CHFS Sampling Design

The sampling design for the China Household Finance Survey (CHFS) consists of two major components, an overall sampling scheme and an onsite sampling scheme based on mapping. This design meets two goals. On one hand, it is to draw a random sample that is representative of all Chinese households. On the other hand, it aims to provide sufficient data for answering important research questions such as household assets allocation, consumption, and saving, to name a few. To achieve these goals, the sampling design has the following four features. First, we oversample observations from relatively wealthy regions. Second, we oversample observations from urban areas. Third, the sample is representative of the diverse geographic regions of China. Fourth, other things being equal, we choose the procedures that cost least.

1. The overall sampling scheme

This project employs a stratified three-stage probability proportion to size (PPS) random sample design. The primary sampling units (PSU) include 2,585 counties (including county level cities and districts) from all provinces (including provincial cities) in China except Tibet, Xinjiang, Inner Mongolia, Hong Kong, Macau, and Taiwan. The second stage of sampling involves selecting residential committees/villages from the counties/cities selected at the earlier stage. The last stage is to select households from the residential committees/villages chosen at the previous stage. Every stage of samplings is carried with PPS method and weighted by its population size. In result, we set the sample size at somewhere between 8,000 and 8,500 households.

In practice, we selected about 80 counties from the PSU, and then 4 residential communities from each of the 80 counties, and then 20-50 households from each of the selected residential communities depending on the level of urbanization and economic development. The average number of households from each residential community is 25. This produces a sample size of 8,000 (4x25x80=8,000).

(1) The first-stage sampling

The first-stage sampling is to select 80 counties out of the 2,585 PSUs. Ideally, the 80 counties should not only cover diverse geographic regions but also contain enough observations from relatively wealthy areas in China. To achieve this outcome, we sort the 2,585 counties into 10 strata based on their GDP per capita. In each stratum, 8 counties are randomly drawn with PPS where each county is weighted by its population size. By this way we got 80 counties covering 25 provinces in China. Table 1 compares some descriptive statistics of GDP per capita of the selected 80 counties with that from the national statistics. They are very close to each other according to the illustration of table 1.

Table 1 The GDP Per Capita of Overall and Sampled 80 Counties (Unit: Yuan)

GDP Per capita	Mean	Standard deviation	Median	Q25	Q75	Kurtosis	Skewness
overall	17334.8	17736.9	11370	7173	20263	3.2	17.64
sample	17809.2	19336.3	11349	7232	21143	3.5	20.41

Note: Q25 and Q75 is quartile of 25% and 75% respectively.

To examine the geographic distribution of the selected counties based on the abovementioned

sampling scheme, we repeated the PPS sampling procedure by random simulation for 1,000 times and compared the average with the national statistics. The small standard deviations shown in the Table 2 suggests that current sampling scheme has produced consistent geographic distributions of the selected counties across trials. On average, the ratio for selected counties in the Eastern, Central, and Western China is about 37: 30: 33. Comparing it to the national statistics, the proportion of counties from the Eastern China is a little bit higher. However, this does not pose any serious problem in that our priority is to have a geographically balanced distribution of counties/cities from all over China. In the final sample of 80 counties/cities from 25 provinces, the ratio for selected counties in the Eastern, Central, and Western China is 32: 27: 21.

Table 2 The Geographic Distribution of Overall and Sample

	Overall			Sample（Simulation:1000times）		
	East	Central	West	East	Central	West
Mean	0.343	0.272	0.384	0.367	0.306	0.327
Standard deviation	—	—	—	0.023	0.023	0.023

(2) The second-stage sampling

At this stage, we select residential communities from the counties. The key is to decide the ratio of urban residential committees over rural villages. If the sample is drawn based on the household registration, it would produce a sample with fewer observations from the urban areas. Given one of the key purposes of the survey is to study the household assets where urban residence are likely to have more assets, we oversample the urban population by the following procedures.

First, we sort the counties according to the proportion of non-agricultural population and divide them into five groups, i.e., quintiles.

Second, for counties in the top quintile with the highest level of non-agricultural population, the ratio of sampled residential communities from the urban areas over sampled villages from the rural areas is 4: 0.

Third, for counties in the quintile below the top one, the ratio of sampled residential communities from the urban areas over sampled villages from the rural areas is 3: 1.

Fourth, accordingly, for counties in the bottom quintile with the lowest level of non-agricultural population, the ratio of sampled residential communities from the urban areas over sampled villages from the rural areas is 0: 4.

Following the above scheme, we got two separate sampling frames, an urban one and a rural one. Given the numbers of residential communities or villages we are supposed to draw from each sampling frame, we then conducted PPS sampling according to the number of households in each residential community. Table 3 illustrates the distribution of urban residential communities in the 80 counties.

Table 3 The Sampling Distribution of Urban Residential Communities

Urban communities	Frequency	Percent (%)
0	15	18.75
1	10	12.50

2	15	18.75
3	15	18.75
4	25	31.25

Table 3 shows that there are 15 counties out of the 80 ones where none urban residential community is drawn. They account for 18.75% of the county level sample; whereas there are 25 counties where none rural village is drawn, accounting for 31.25% of the sample. This outcome meets our goal of oversampling the urban population. Accordingly, among the 320 selected residential communities at the second-stage sampling, the ratio of the urban one to the rural one is 181: 139.

(3) The third-stage sampling

The last stage of sampling in CHFS is to select households from the chosen residential communities. In each rural village, we randomly draw 20 households; whereas in the urban areas, the number of households that we select varies according to the housing price of the residential communities. Based on the average housing price of each neighborhood, we sort the residential communities and divide them into quartiles. For the top quartile where the average housing price is the highest, we draw 50 households from each residential community; for the bottom quartile where the housing price is the lowest, we select only 25 households. Thus we are able to have a greater number of wealthy households in the sample. See Table 4 for the distribution of the number of households across urban residential communities.

Table 4 The Distribution of Number of Households across Urban Communities

Number of households	Number of communities	Percent of communities (%)	Cumulative percent of communities (%)
25	53	32.72	32.72
30	52	32.10	64.82
35	28	17.28	82.10
50	29	17.90	100.00

2. The onsite sampling scheme

(1) Mapping residential areas

The onsite sampling is based on the mapping of the residential areas and the collection of household lists in the area. The extent to which the map is drawn precisely directly affects the quality of this last stage of sampling.

The CHFS develops a geographic information sampling system using the technologies of remote sensing, GPS, and GIS to collect geographic information of the targeted areas. The fine-grained digital imagery and vector maps used in the mapping come from the Institute of Geographic Information of the Chinese Academy of Sciences. When on the field, our trained mapping technicians use an electrical measuring instrument and a GPS system to collect accurate electronic data, which are automatically transferred to computers to create high-quality vector maps. We also take into account the potential changes of the geographic data after we collect the data in the first place and manually check and record any change at later stages. In this way, we make sure the geographic information in the virtual world in our system matches that in the real world.

The system we developed not only allows our mapping technicians to draw residential household location directly on the electronic map but also stores relevant household location information used for the last stage sampling. This innovation improves efficiency, decreases potential errors in the mapping and sampling process, and helps to protect the privacy of household information. Our working procedure is illustrated in Figure 1 below.

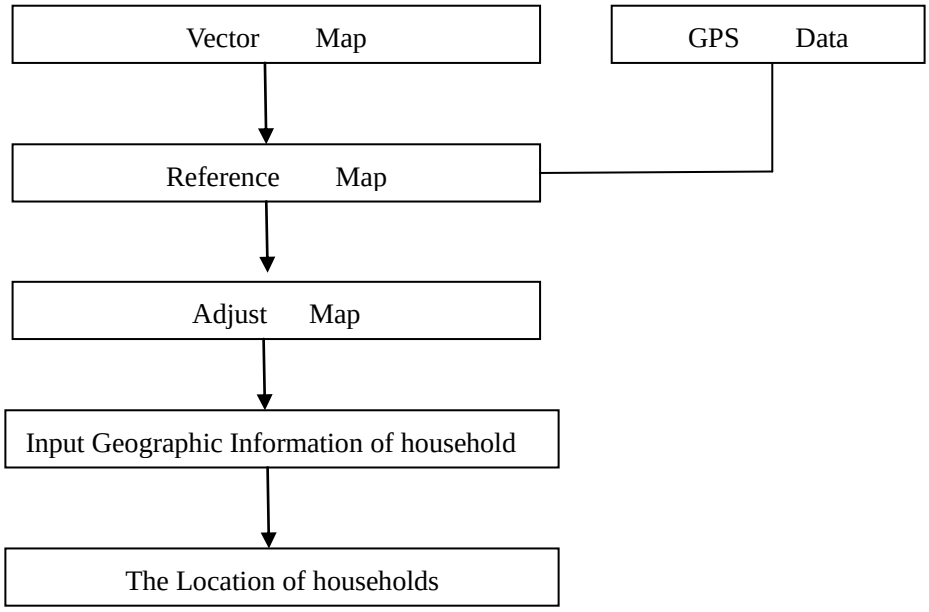

Figure 1 The Critical Process of Mapping

(2) Selecting households

We use the equal-space sampling procedure to draw households from the household list collected from the previous mapping stage.

First, we calculate the sampling interval, i.e., out of how many households one is chosen, using the following formula:

Sampling interval = total number of households in the community / number of households to be selected (round up to the closest integer)

E.g., If we plan to draw 30 households from the 100 households of the residential committee/village, we get 100/30=3.33. Then the sampling interval should be 4.

Second, the random starting point is decided by the unit digit of the clock time when the procedure is carried out. For example, if the clock time is 15:34, then 4 is the starting point; if it is 12:03, then 3 is the one.

Third, we draw the households. The first selected household is the one that the random starting point corresponds to on the household list. Using above example again, if 3 is the starting point and 4 is the sampling interval, the 3rd household on the list is then the first chosen one in the sample, so are the 7^{th}, 11^{th}, 15^{th}, 19^{th} …until all 30 households are drawn from the list.