北京大学 中国社会科学调查中心
Institute of Social Science Survey, Peking University

抽样设计和权数应用

吕萍

主要内容

一、抽样设计

二、权数应用

一、中国家庭追踪调查抽样设计

1、调查目的

中国家庭追踪调查（Chinese Family Panel Studies, CFPS）通过跟踪搜集个体、家庭、社区三个层次的数据，反映中国社会、经济、人口、教育和健康的变迁，为学术研究和政策分析提供数据。

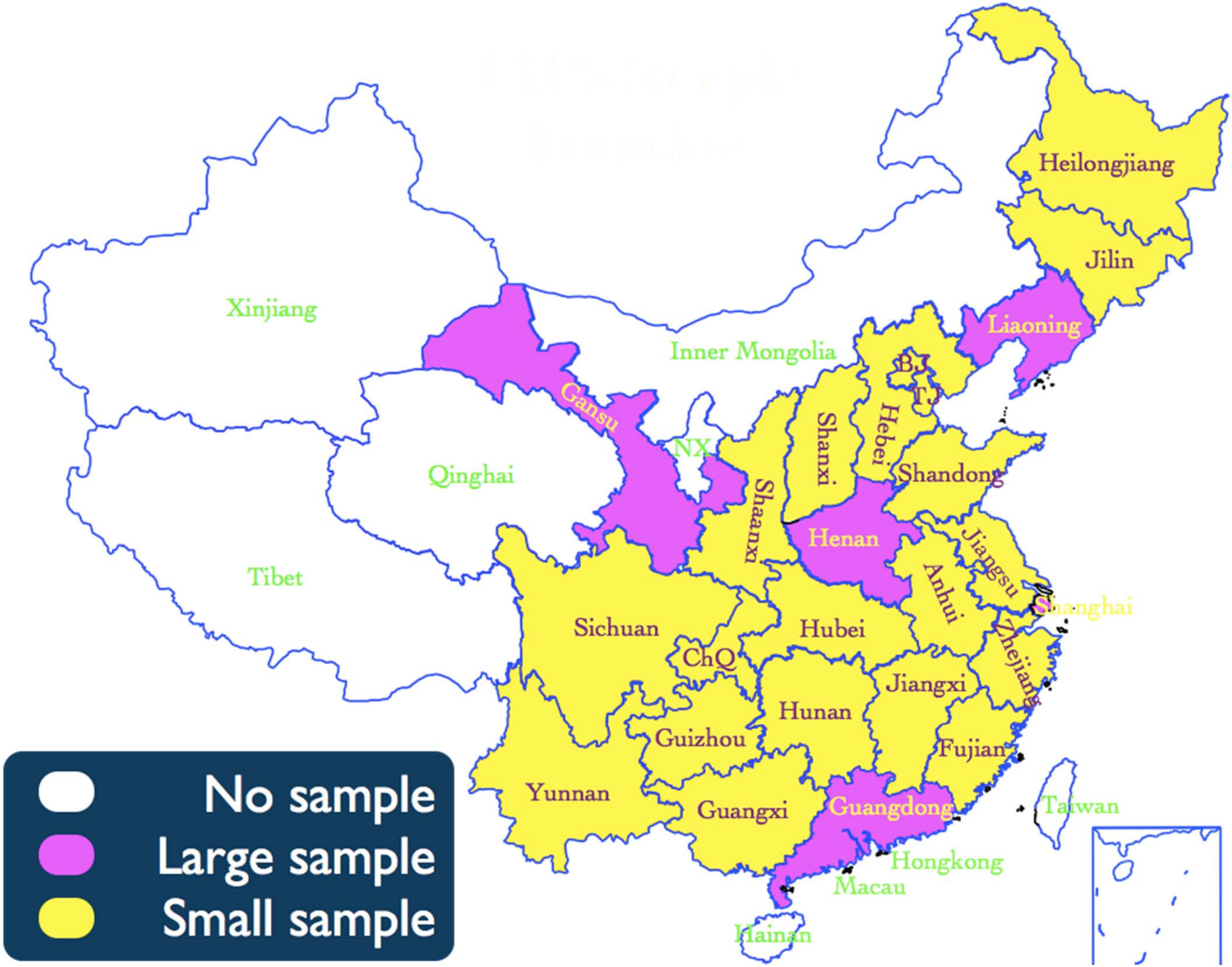
2、调查对象和样本量

调查对象是中国（除香港、澳门、台湾、新疆维吾尔自治区、西藏自治区、青海省、内蒙古自治区、宁夏回族自治区、海南省）25个省市的常住家庭户并对家庭户中所有满足条件的家庭成员进行调查。

子总体

5个省级单位推断单位，为省级层次的自我代表层。目的是获得省级资料，进行省级之间的比较。

- 广东省
- 上海市
- 辽宁省
- 甘肃省
- 河南省


Heilongjiang
Jilin
Liaoning
Xinjiang
Inner Mongolia
Gansu
NX
Qinghai
Tibet
BJ
TJ
Hebei
Shanxi
Shaanxi
Shandong
Henan
Jiangsu
Anhui
Shanghai
Zhejiang
Sichuan
ChQ
Hubei
Jiangxi
Hunan
Guizhou
Fujian
Yunnan
Guangxi
Guangdong
Taiwan
Hongkong
Macau
Hainan
No sample
Large sample
Small sample

➢区分大、小省，一方面通过大省样本获得不同类型地区的省级资料，了解不同类型省/市的发展；另一方面，通过合并大小省资料了解25个省市自治区的总体发展，为研究中国家庭问题提供更加丰富的数据资料。

样本量

省市类型	省的名称	户数（有效户数）
大省（省级推断单位）	上海市	1600
	辽宁省	1600
	河南省	1600
	甘肃省	1600
	广东省	1600
小省（非省级推断单位）	江苏省、浙江省、福建省、江西省、安徽省、山东省、河北省、山西省、吉林省、黑龙江省、广西壮族自治区、湖北省、湖南省、四川省、贵州省、云南省、天津市、北京市、重庆市、陕西省	8000

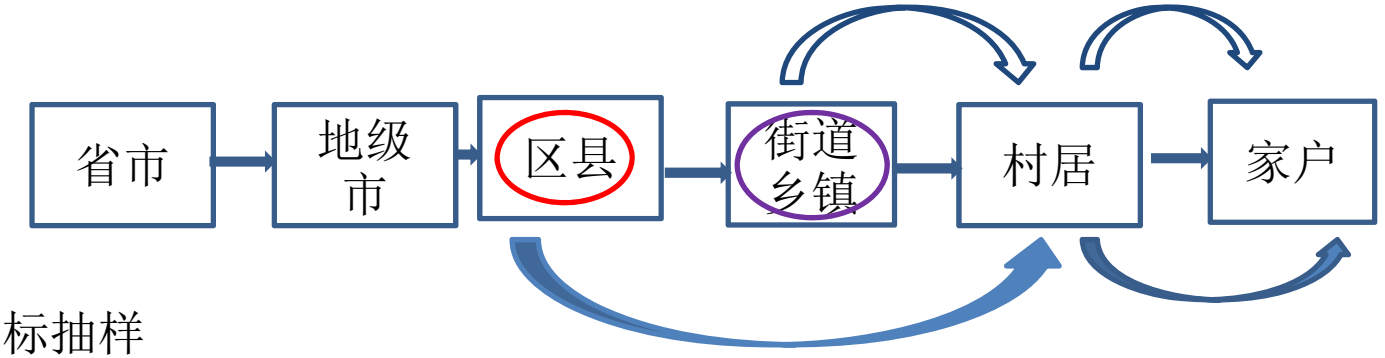
表1 全国25个省市的分类

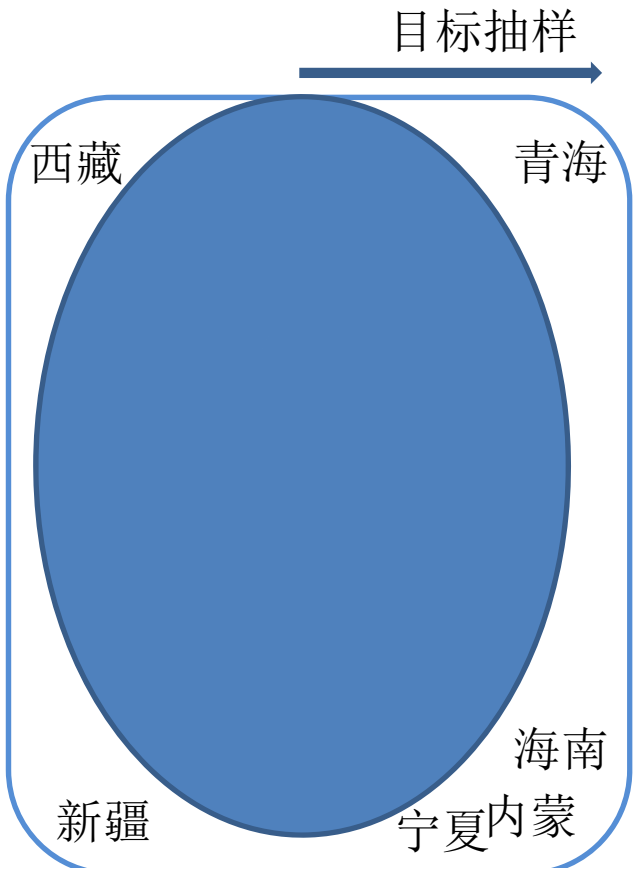
3、抽样设计

抽样设计是在充分利用辅助信息对抽样框进行有效排序的基础上的三阶段不等概率的系统PPS整群抽样设计。其中，上海市由于自身区县数比较少，抽样设计稍有不同。

特点：

- 区分大、小省，一方面通过大省样本获得不同类型地区的省级资料，了解不同类型省/市的发展；另一方面，通过合并大小省资料了解25个省市自治区的总体发展，为研究中国家庭问题提供更加丰富的数据资料。
- 不区分城乡，但是在设计过程中尽可能多的利用辅助信息对抽样框进行排序。在中国，家庭户的各个指标更多的受地理环境、行政区划设置等定性因素以及省市、区县的人均GDP、非农业人口比例、人口密度等定量指标密切相关。在抽样设计中，选取这两部分定性和定量指标对抽样框中的区县、村居进行排序，最大程度的提高样本的代表性。





广东

甘肃

河南

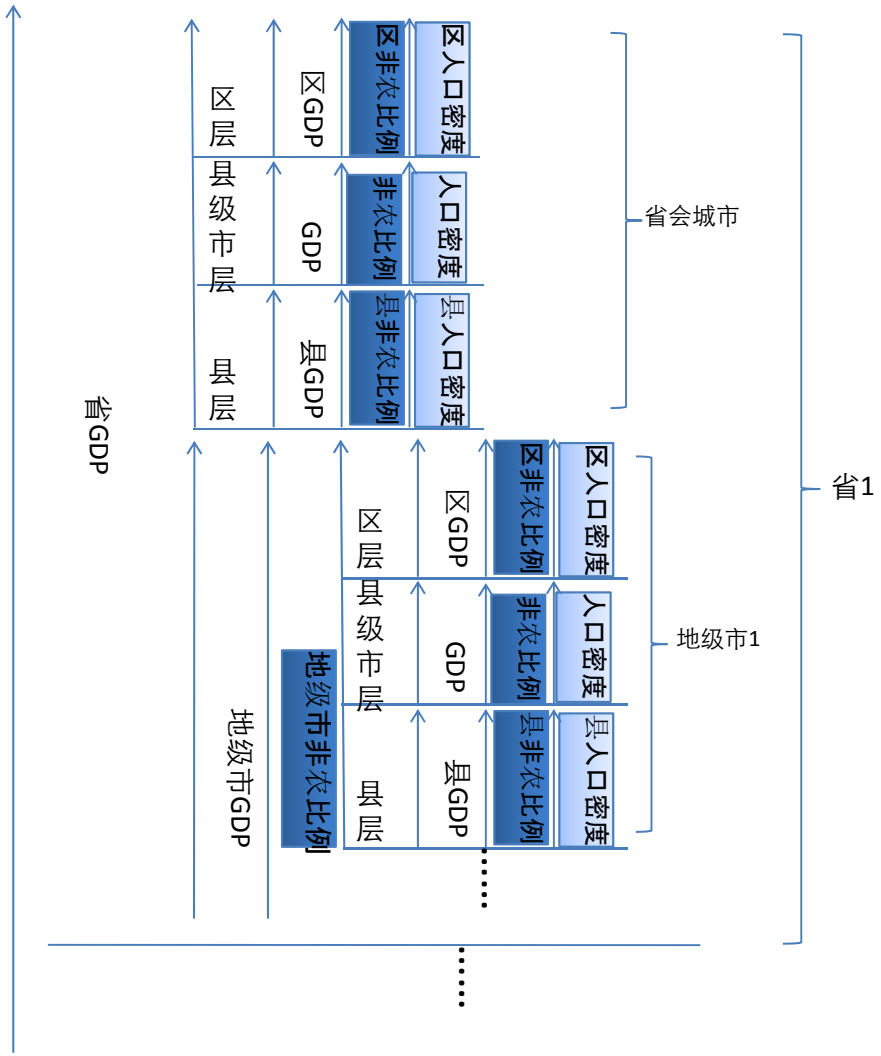
辽宁

上海

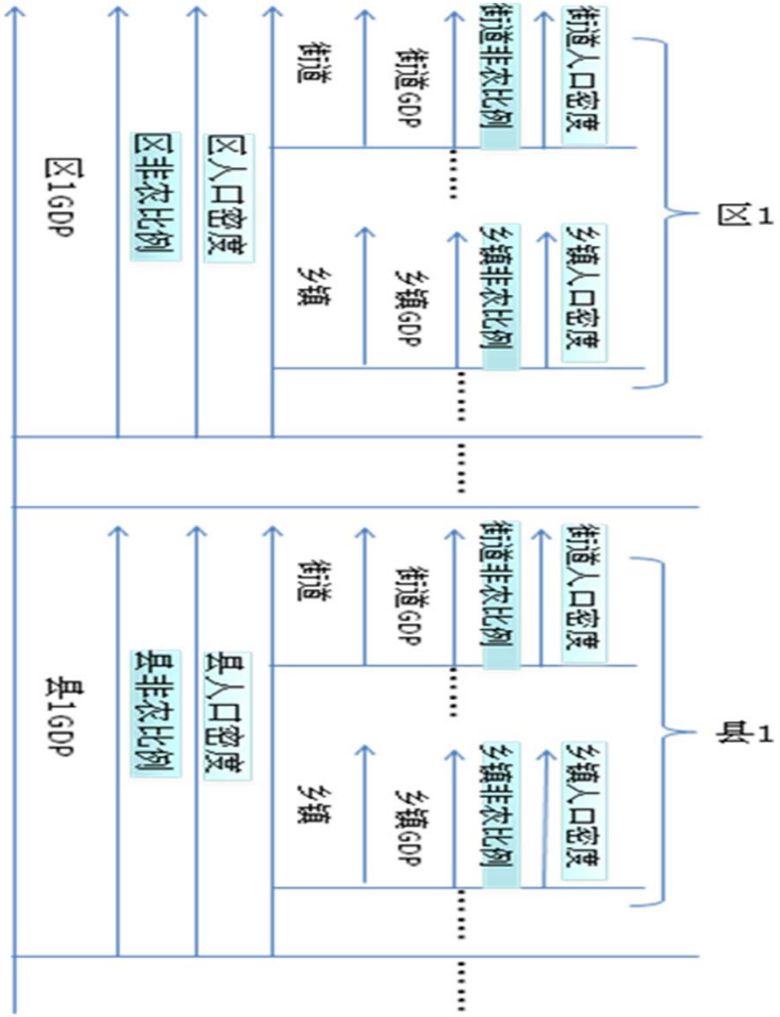
6个子总体

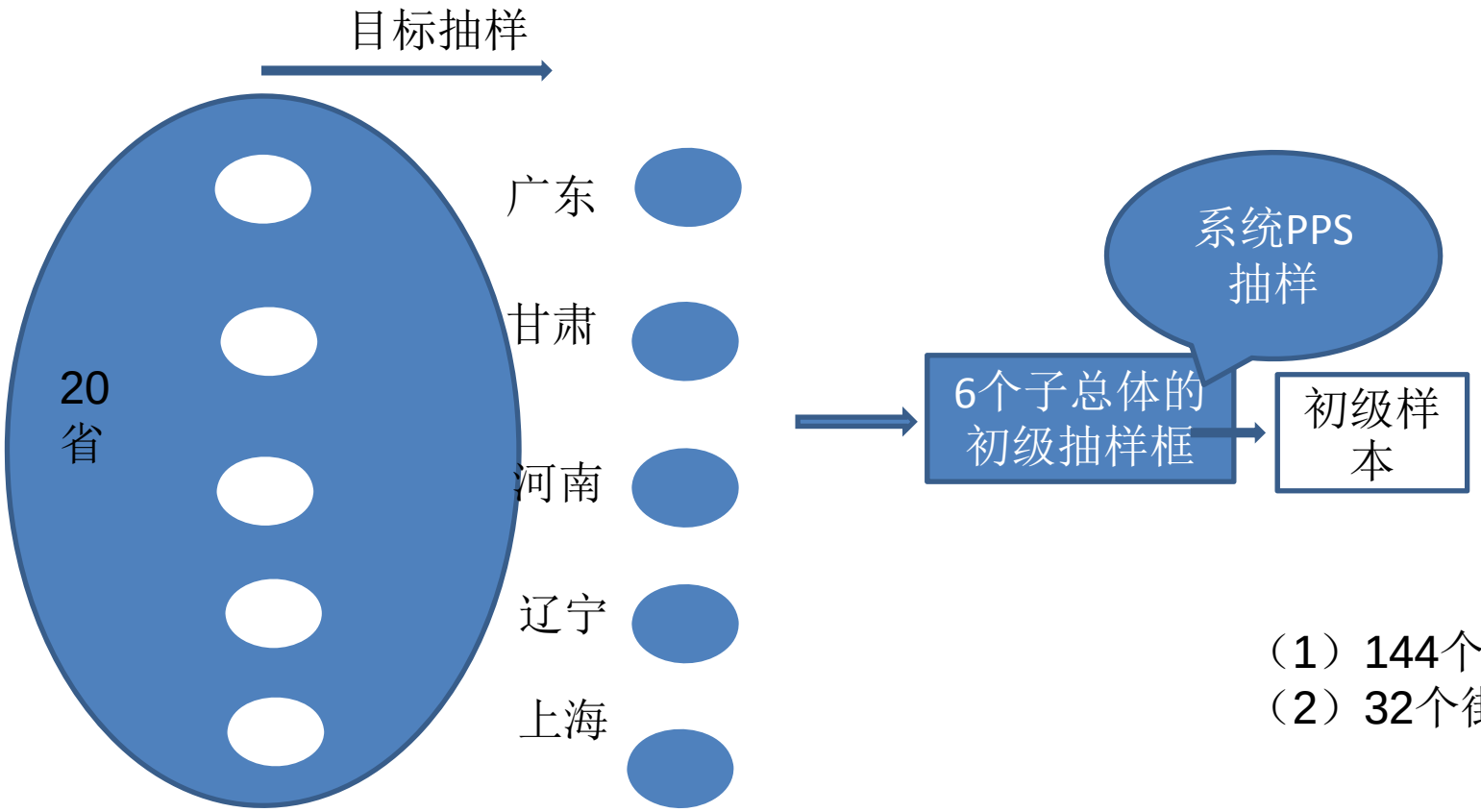
（1）第一阶段抽样

小省，大省（除上海）的抽样框（区县抽样框）



上海市的抽样框（街道、乡镇抽样框）





（1）144个区县样本
（2）32个街道乡镇样本

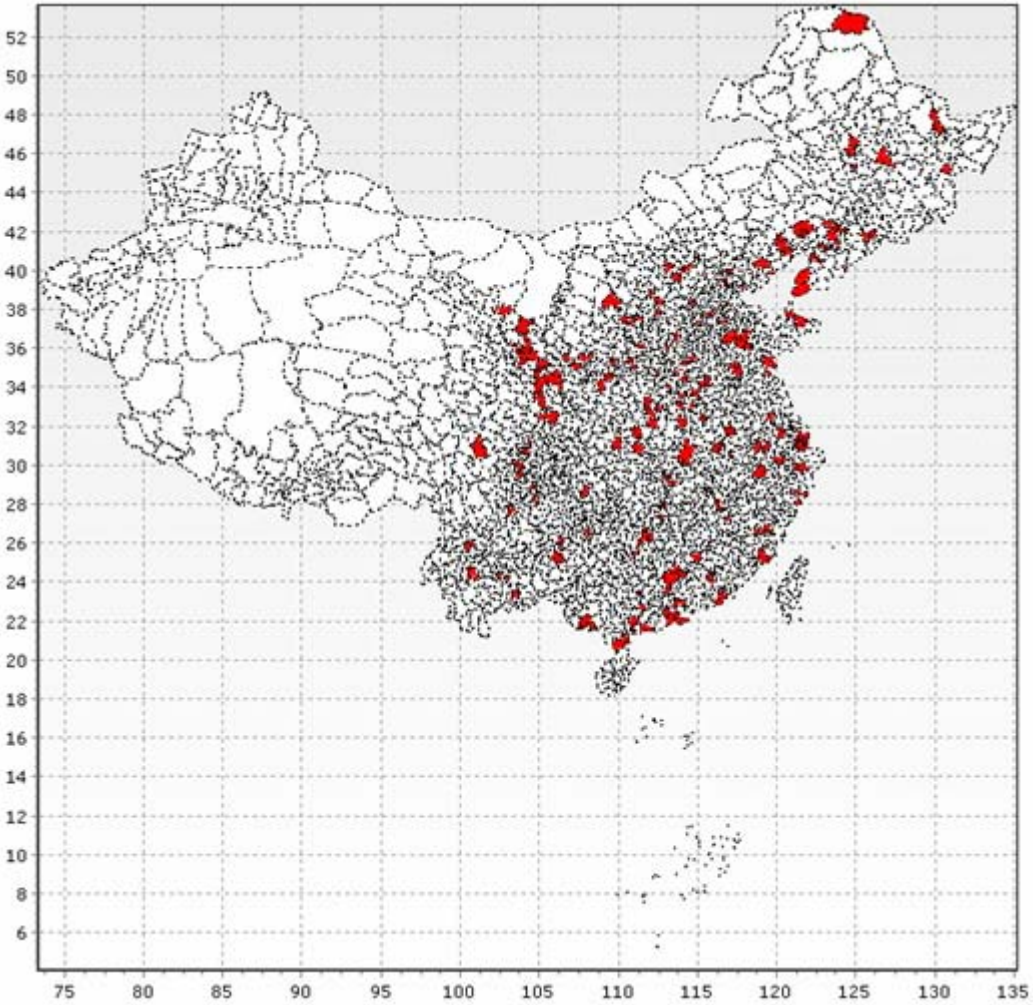
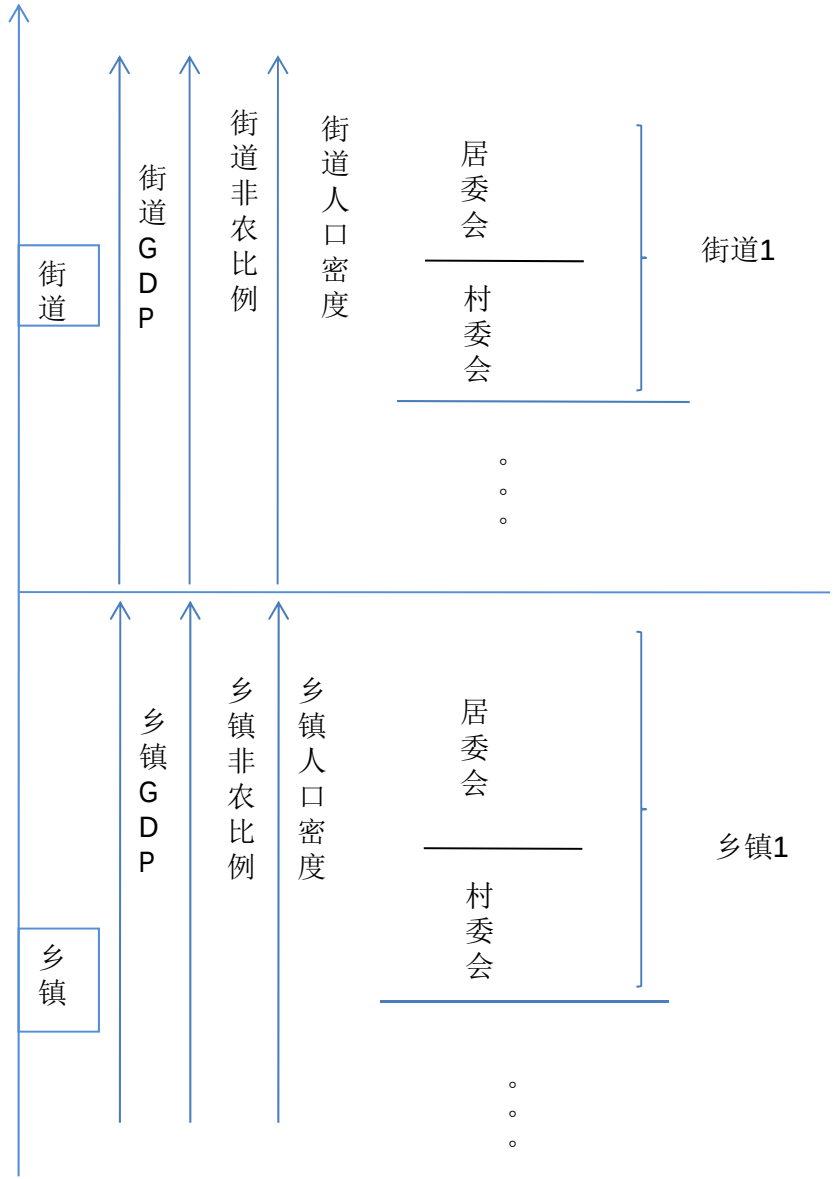

52
50
48
46
44
42
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
75
80
85
90
95
100
105
110
115
120
125
130
135

表2 小省的样本区县在各省市的分配

省市	样本区县数	省市	样本区/县数
北京市	1	重庆市	2
福建省	2	江西省	3
黑龙江省	5	陕西省	3
山西省	7	广西省	3
安徽省	3	湖北省	3
浙江省	3	云南省	4
天津市	1	贵州省	5
江苏省	3	湖南省	6
吉林省	3	山东省	7
河北省	8	四川省	8

（2）第二阶段抽样

2.1 村居抽样框



小村居的合并
大村居的拆分

（2） 超大社区（人口数大于10000）

超大社区（主要是居委会）将居委会人口数大于10000的村居委会，结合当地的地理位置、建筑物特征等进行拆分，拆分后每个单元的人口数不低于4000人，进行二次抽样选取一个单元做为调查对象。

◆抽样后进行（CFPS）

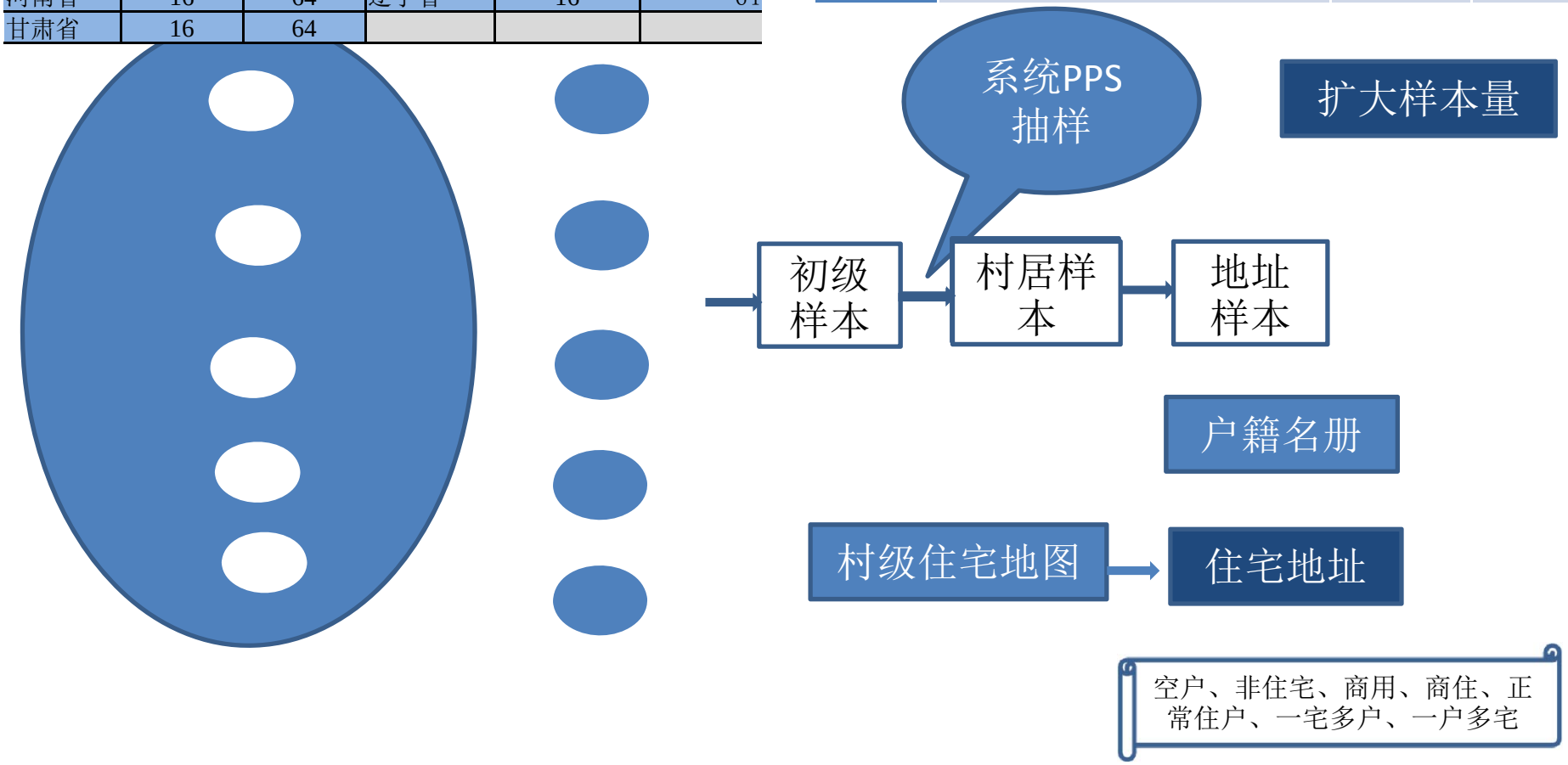
◆抽样前进行

（3）末端抽样框

省市	样本区/县数	村居数	省市	样本区/县数	村居数
北京市	1	4	重庆市	2	8
福建省	2	8	江西省	3	12
黑龙江省	5	20	陕西省	3	12
山西省	7	28	广西省	3	12
安徽省	3	12	湖北省	3	12
浙江省	3	12	云南省	4	16
天津市	1	4	贵州省	5	20
江苏省	3	12	湖南省	6	24
吉林省	3	12	山东省	7	28
河北省	8	32	四川省	8	32
广东省	16	64	上海市	16	64
河南省	16	64	辽宁省	16	64
甘肃省	16	64			

表 2010年中国家庭动态跟踪调查初访末端样本量

地区	类型	应答率	接触样本数量
一类地区	居委会（主城区和城乡结合区的村委会）	60%	42
	其他村委会	70%	36
二类地区	居委会（主城区和城乡结合部的村委会）	70%	36
	其他村委会	80%	32
三类地区	居委会	80%	32
	村委会	90%	28



一宅多户

- 一宅多户，即一个住宅地址对应着两个或两个以上的住户，例如兄弟分家后仍同住、某间房子出租给多人、老人与成家后的儿女同住但经济独立等，此时会遗漏目标单元，产生低覆盖的抽样误差。需要对地址中的多个住户进行拆分编号，并在备注中标明一宅的第几户，以避免实际调查中利用CAPI系统住户过滤系统再次进行住户过滤。
- 一宅多户是实质意义上的一宅多户，一个住户拥有两个户口本的情况不是一宅两户。

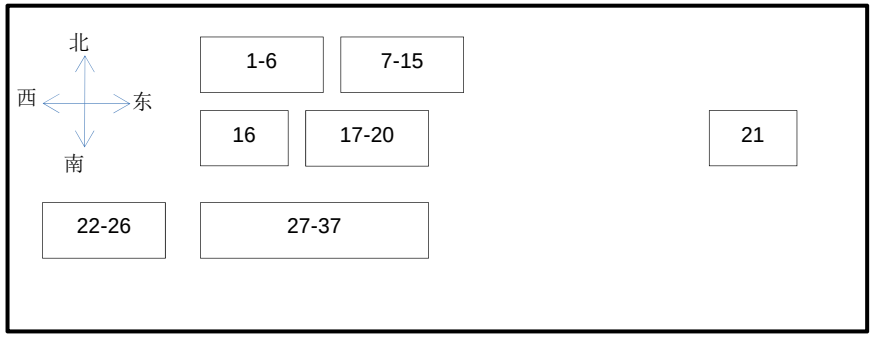
一户多宅

- 一户多宅，即一个住户拥有两套或两套以上的房子，若直接按照地址列表清单进行抽样会产生过覆盖的抽样误差。此时，需要按照户主姓名进行筛选，若确实属于一户多宅的情况，需要对其进行合并。
- 一户多宅是实质意义上的一户多宅，若是一家人有两套房子，但是一套房子闲置，不是实质意义上的一户多宅，空置的房子需要标注为空户。同样的，若是另一套房子出租，也不是实质意义上的一户多宅，出租的房子仍是正常住户。

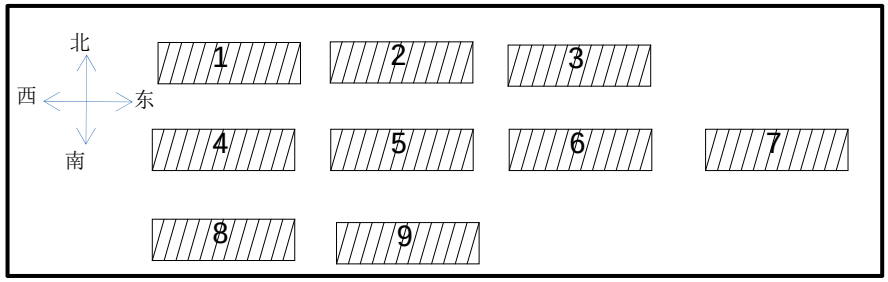
末端抽样框绘制：

- 确定行走路线

1、如果建筑物有自然编号，按照原有编号从小到大的顺序行走。
2、在村（居）委会，原则上以西北角为起点，按照从西到东、从北到南的路线行走。
3、原则上保证不重不漏。按照行走路线规则或原有自然编号清楚标出各建筑物编号。
务必做到在给定住宅楼院编号、住宅楼院中的住宅编号后，无论任何人能唯一确定一个住宅。

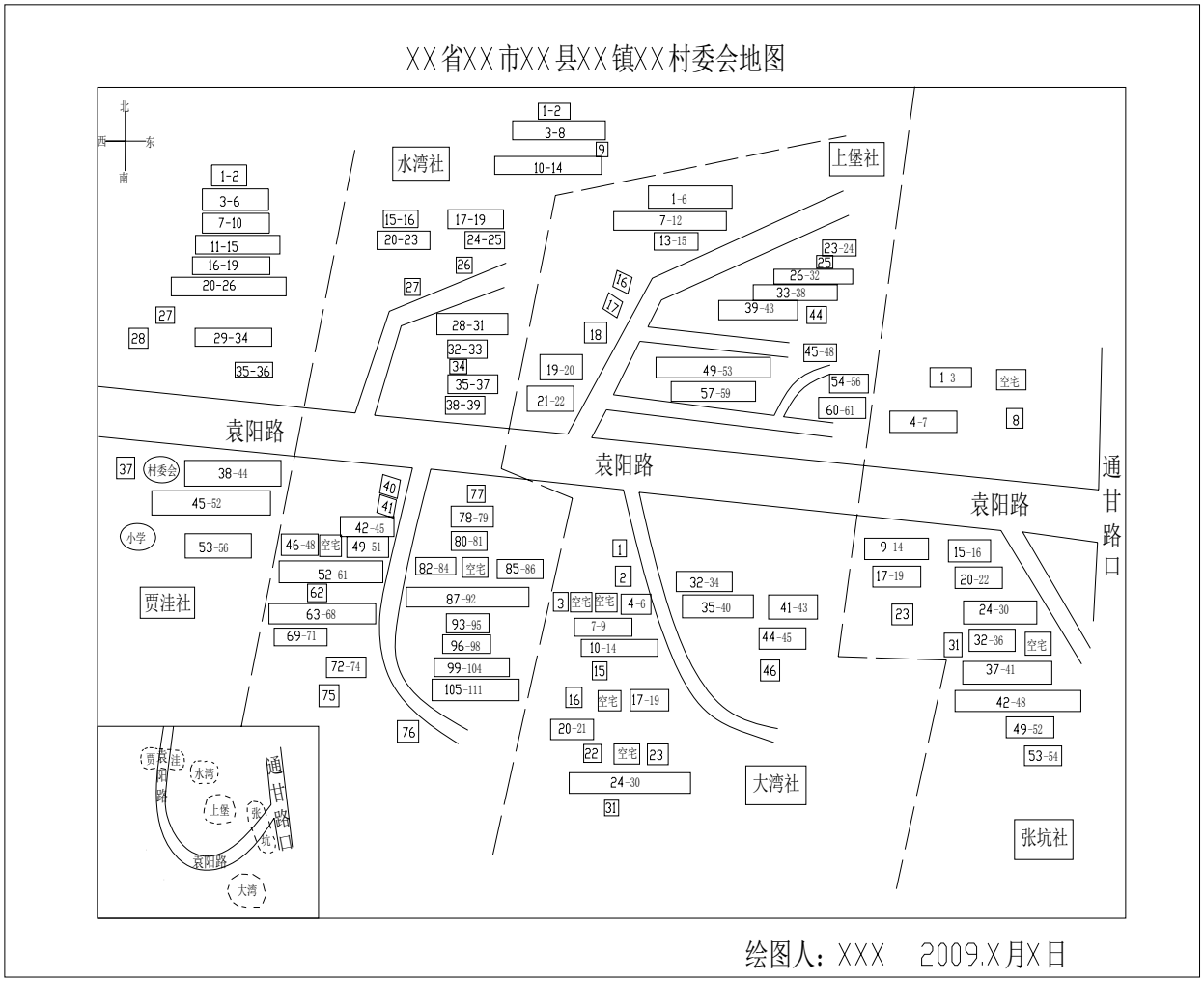


以住宅编码体现的平房区从西到东、从北到南的行走路线



以住宅编码体现的楼房区从西到东、从北到南的行走路线

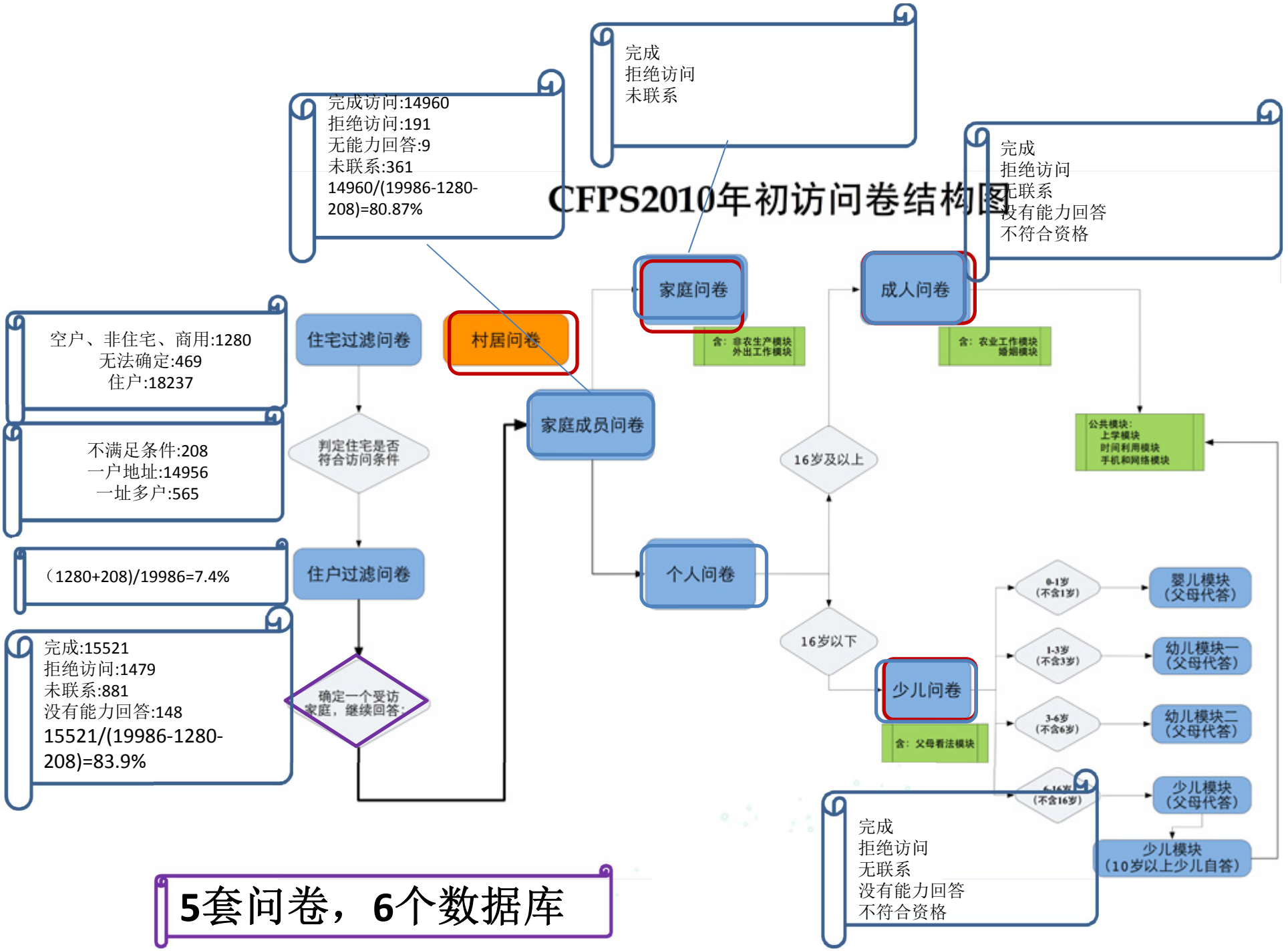
- 如果一个村（居）由多个边界分明的自然村/社、村民小组、大队、小居委会等组成，且没有住宅建筑物自然编号，则在路线上可以考虑以自然村/社/居为单位行走和绘图，但在自然村/社/居内仍需按照从西到东、从北到南的路线行走和绘图，且需补充各自然村/社/居分布概要图。具体图例如下：



住户列表清单

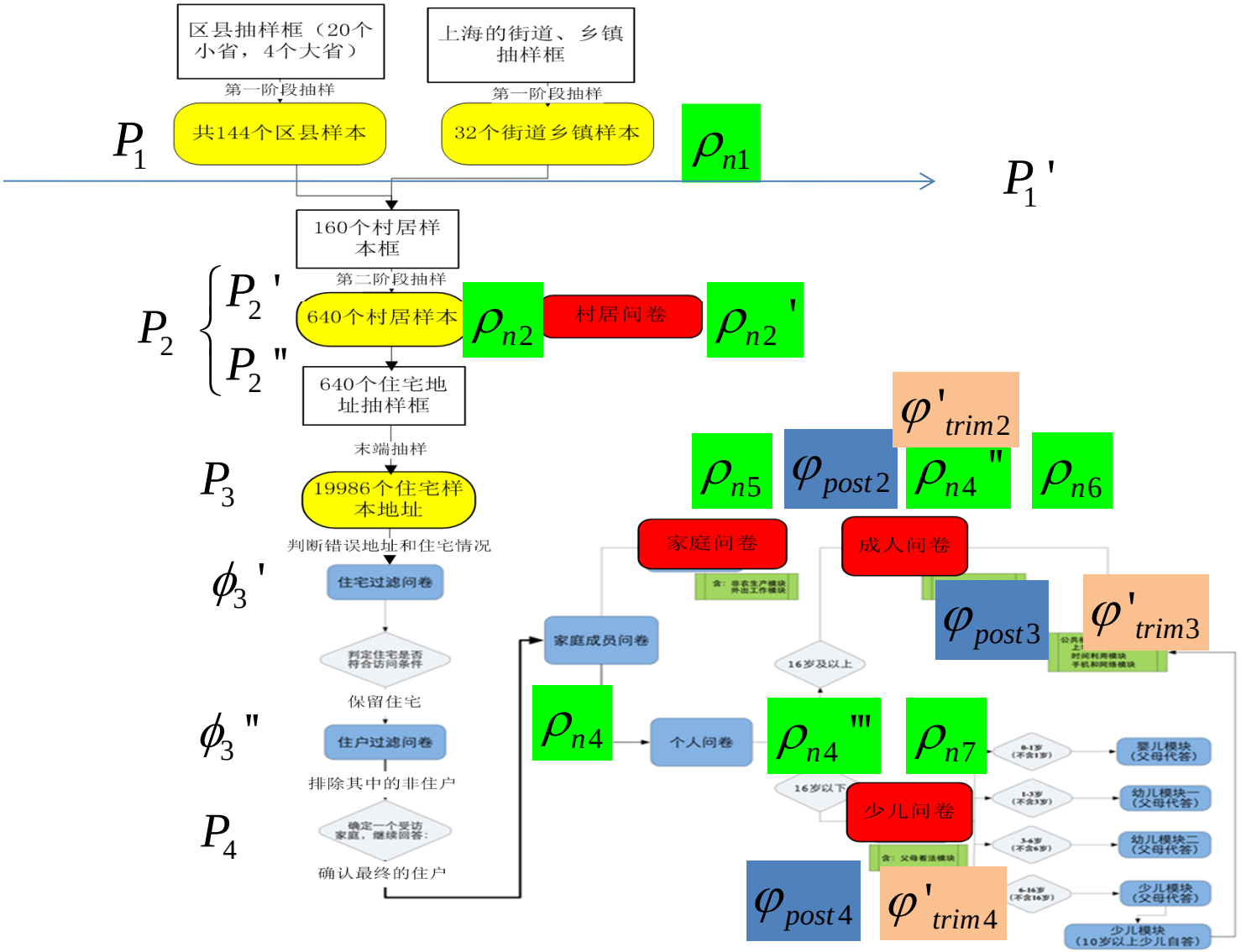
建筑物编号	住户地址	住户编号	累 计 住 户 数	户主姓名	备注
001	向阳小区1号楼东门101	0001	1	张×	原有编码
001	向阳小区1号楼东门102	0002	2	李×	
	…….	0018	18	郑××	
002	向阳小区2号楼东门101	0001	19	王××	
002	向阳小区2号楼东门102	0002	20	赵××	
	.....		199		
011	11号院4-1	0001	200	王××	大杂院（4宅2户）
011	11号院4-2	0002	202	赵××	
011	11号院4-3	/	/	/	不住人商店
011	11号院4-4	/	/	/	空宅
012	向阳小区12号楼101	0001	203	陈××	

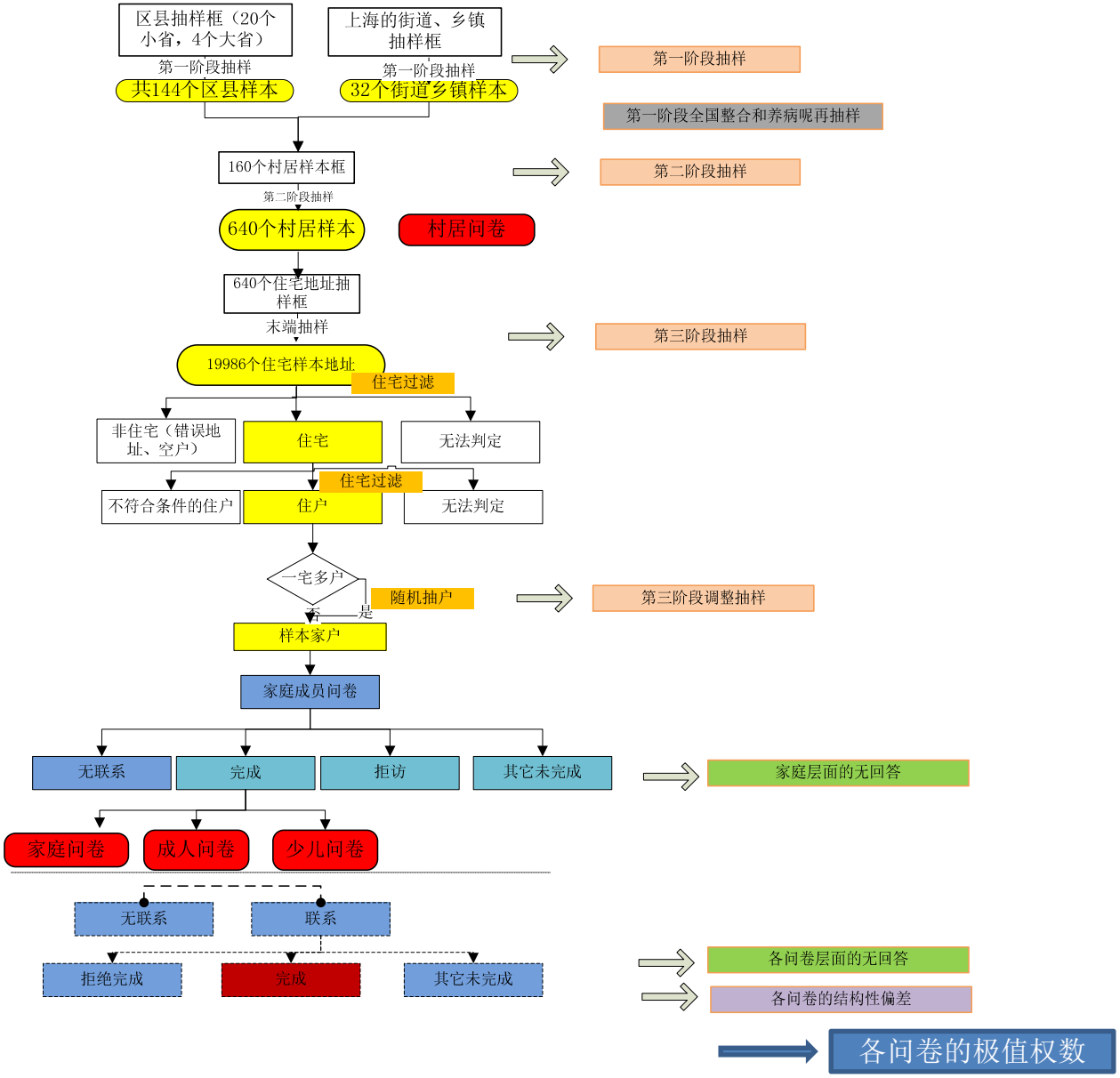
CFPS2010年初访问卷结构图



5套问卷，6个数据库

2010年CFPS初访调查流程图




区县抽样框（20个小省，4个大省）
第一阶段抽样
共144个区县样本
上海的街道、乡镇抽样框
第一阶段抽样
32个街道乡镇样本
第一阶段抽样
第一阶段全国整合和养病呢再抽样
160个村居样本框
第二阶段抽样
第二阶段抽样
640个村居样本
村居问卷
640个住宅地址抽样框
末端抽样
第三阶段抽样
19986个住宅样本地址
住宅过滤
非住宅（错误地址、空户）
住宅
无法判定
住宅过滤
不符合条件的住户
住户
无法判定
一宅多户
随机抽户
第三阶段调整抽样
否
是
样本家户
家庭成员问卷
无联系
完成
拒访
其它未完成
家庭层面的无回答
家庭问卷
成人问卷
少儿问卷
无联系
联系
拒绝完成
完成
其它未完成
各问卷层面的无回答
各问卷的结构性偏差
各问卷的极值权数

- libname shuju "C:\Users\520\Desktop\CFPS2010shuju";
- options nofmterr;
- option compress=yes;
-
- **proc contents** data=shuju.Cfps_2010_family out=famvar(keep=NAME VARNUM LABEL) ;**run**;
-
-
-
- /*NAME VARNUM LABEL*/
- /*PROVCD 3 省代码*/
- /*Fswt_Nat 622 家庭权重-全国样本*/
- /*Fswt_Res 623 家庭权重-全国再抽样样本*/
-
- /*31 上海市 */
- /*41 河南省*/
- /*44 广东省*/
- /*62 甘肃省*/
- /*21 辽宁省*/

- **proc means** data=shuju.Cfps_2010_family n nmiss mean stderr clm ;
- var faminc_net_old ;
- **run** ;

- **proc means** data=shuju.Cfps_2010_family n nmiss mean stderr clm ;
- var faminc_net_old ;
- weight Fswt_Nat ;
- **run** ;

proc surveymeans data=Cfps_2010_family nobs nmiss mean stderr min max ;

- var faminc_net_old ;
- stratum stra ;
- cluster COUNTYID ;
- weight Fswt_Nat ;
- **run** ;

- **proc freq** data=Cfps_2010_family1;
- tables FD1 ;
- **run**;
-
- **proc freq** data=Cfps_2010_family1;
- tables FD1;
- weight Fswt_Nat ;
- **run**;

- **proc surveyfreq** data=Cfps_2010_family1;
- stratum stra ;
- cluster COUNTYID ;
- weight Fswt_Nat ;
- tables FD1 ;
- **run** ;

- svyset psu [pweight=Fswt_Nat], strata(stra)
- svy: mean a
- svy: prop a

谢谢！