China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-25

系列编辑：谢宇　责任编辑：张聪

中国家庭追踪调查
2012 年数据库介绍及数据清理报告

吴琼 戴利红 崔雅红 张文佳

2014.10.8

一． 2012 年各数据库介绍

CFPS 基线调查（2010 年）数据共有五个基本数据库：社区库、家庭经济库、成人库、少儿库以及家庭成员关系库。2012 年的首轮全国追踪调查对基线调查所界定出来的 57,155 名基因成员及其所在家庭进行追踪。具体追踪策略见《中国民生发展报告 2013》第一章。虽然问卷内容部分有所调整，但是除社区问卷在 2012 年省略之外，其它各问卷结构与 2010 年总体类似。2012 年 CFPS 数据库基本情况如表 1 所列。

表 1　CFPS2012 年各库基本状况[1]

数据库	样本量	变量数
成人数据库	35720	1765
少儿数据库	8624	867
家庭关系数据库	55014	318
家庭经济数据库	13315	816
跨年 ID 库	61423	18

1. 成人库：

成人库包括 2010 年界定出来的基因成员中 2012 年追踪调查时年龄处在 16 岁及以上的个人和 2012 年新增家庭成员中年龄处在 16 岁及以上的个人。访问方式为面访(IWmode=1)或电访(IWmode=0)，问卷形式为长问卷(longform=1)和/或短问卷(shortform=1)。成人库中的个人样本包括来自 2010 年基线调查的 34425 个基因成员、2012 年新加入的 23 个新基因成员以及与基因成员有直系亲属关系但本身并不属于基因成员的 1271 个核心成员。

[1] 此篇报告中的统计量根据 CFPS2012 数据版本 5.0 得出。

2. 少儿库：

少儿库包括 2010 年界定出来的基因成员中 2012 年追踪调查时年龄处在 15 岁及以下的个人，以及 2012 年新增家庭成员中年龄处在 15 岁及以下的个人。其中 10 岁及以上的少儿既有家长代答问卷，也有少儿自答问卷；而 10 岁以下的少儿只有家长代答问卷。访问方式依然为面访或电访，问卷形式为长问卷和/或短问卷。少儿库中包括来自 2010 年基线调查的 7257 个基因成员、2012 年新进的 1264 个新基因成员以及核心成员 103 人。

3. 家庭经济库：

家庭经济库以家庭为单位，包括 2010 年基因成员所在原家庭以及由 2010 年家庭因婚姻变化、子女经济独立等原因所派生出来的另组家庭。其中 2010 年基线家庭 12625 户，剩下 690 户为 2012 年另组家庭。访问方式为面访或电访。

4. 家庭成员关系库：

家庭成员关系库以家庭成员为单位，包括 2010 年基因成员及 2012 年新增家庭成员的配偶、父母及子女的基本信息。2012 年家庭成员关系库中包括来自 13453 个家庭的 55014 条个人样本。需要注意的是：这 55014 条观测并不代表着 55014 个独立个人，这是因为 2012 年家庭成员库将另组家庭成员分别放在原家庭列表和另组家庭列表中，并用是否在家（co_a12_p=1 表示在家，0 表示离开原家庭）来表明该个体在 2012 年经济上属于哪个家户。此种安排是为了更好的显示个体在 2010 年到 2012 年间的动态过程。详细情况可参见《中国家庭追踪调查 2012 年家庭成员库的分解与家庭关系库的构建》。家庭成员关系库中包括的个人独立样本有 53759 条，其中有记录的死亡人数为 639 人。住在原家庭的基因成员为 40757 人（75.8%），新进基因成员 1354 人（2.5%），新进的非基因成员 2713（5.0%）经济上归属原家庭但物理外出的基因成员 5488 人（10.2%），另组基因成员 2672 人（5.0%）。另外还有少量基因成员由于出境、服刑等原因不需要追踪（n=136）。

5. 跨年 ID 库：

跨年 ID 库包括在任意调查年被界定为 CFPS 样本家庭的所有家庭成员，而不论其在调查当年的访问状态。跟家庭成员库相比，跨年 ID 库中包含了更多的个人样本（譬如去世及调查当年未接触到的家庭中的家庭成员）。跨年 ID 库的主要用途是记录个人样本的去留情况，比如各调查的受访情况、生存状态以及所在家户等最基本的信息。因为 2012 年调查是跨年 ID 库的第一次发布，我们将在本文的后一部分对这一数据库进行详细介绍。

二． 2012 年问卷数据清理步骤

1. 各库内部样本编码清理

虽然大部分样本异常情况（如在执行过程中的临时置换样本及由于特定原因出现重复样本的数据）已在调查过程中记录并在数据清理的第一阶段处理完成，但有一小部分样本错误的情况需要依靠后期的数据清理发现其逻辑上不合理的地方，再结合当时的访员记录及其它样本管理数据，得出结论。如在跨年 ID 库以及家庭成员关系库清理过程中发现了部分重复样本的情况，大部分是由于原家庭及另组家庭中同一个人重复出现的现象。这种重复样本经各方信息确认后在下一轮数据清理的第一步会先删除。除此之外，还有部分家户号的清理。如在 2012 年的执行过程中，出现了另组家庭又发生分裂的情况，但在后期查看数据时，发现中间过程的家户中并未包括任何 CFPS 基因成员，从本质上来说，这个中间过渡的家庭并非 CFPS 调查所关注的对象，调查中也未产生实质性的信息采集，因此在这一阶段也需要将此类家户删去。在这里值得一提的是，2012 年家庭成员关系清理过程中发现的一些易出错的环节，如上文所提的相关联家户出现重复个人样本，中间家庭并非调查关注对象等情况，2014 年新一轮的调查设计都在相关环节做了调整。这些调整进一步降低了出错的可能性，也为后期的数据清理带来了便利。

2. 面访与电访数据的合并

CFPS 调查主要采用面访的形式进行，但在客观条件受限，面访无法实现时（如家庭中个别成员外出且无法进行面访），调查会采用电访形式。2012 年家庭经济、成人、少儿问卷中电访所占比例分别为 0.89%、2.8%、0.44%。

电访问卷总体上与面访问卷类似，但删除了不适合电话访问的内容（如认知测试），并对一些复杂问题进行了简化。为了方便数据使用，降低用户在分析时遗漏部分样本的可能，我们在数据清理过程中将电访和面访问卷数据进行了合并，并给出了一个标识变量，说明此条样本的访问模式（IWmode=1,0 分别代表面访和电访）。有一点需要数据用户注意的是，面访和电访库虽然在具体问卷问题上采集信息目标一致，但问题本身可能会有差别。在整合电访和面访数据库时，我们对电访和面访问卷进行了逐题的对比，将在两套问卷中完全相同的题取相同的变量名，而有所不同的题采用不同的变量名。

表 2 中列出了电访问卷与面访问卷不完全匹配的变量，我们分别给出其相应的变量名[2]。研究者可根据具体的研究需要对其进行再处理。

表 2　2012 年家庭问卷电访和面访不完全匹配问题信息列表

题号	电访有效变量	面访有效变量	不匹配情况
L 部分　家户农业收入与支出			
L4	FL4	FL401 FL402 FL403 FL404	电访和面访针对种植业生产和林业生产投入金额分别采用总、分不同的提问方式。
L9	FL9	FL901 FL902 FL903 FL904	同 L4 题
P 部分　家户政府转移支付收入			
P301	FP301_T	FP301	电访和面访针对食品，酒水等消费金额的提问，分别提问的内容与时间跨度不同。电访是一个月，面访是一周。
P304	FP304_T	FP304	电访和面访时间跨度不同。电访是一个月，面访是一周。
P401	FP401_T	FP401 FP405	电访和面访针对邮电、通讯、交通支出金额分别采用总、分的提问方式。
P407	FP407_T	FP407 FP508	电访和面访针对教育、娱乐、交通支出金额分别采用总、分的提问方式。
P505	FP505_T	FP504 FP505	电访和面访针对购买、维修各种交通、通讯工具（如汽车、自行车、电动自行车和手机）

[2] 2012 年极少量面访的样本（约 10 人次左右）用了电访的问卷，这会导致有些面访的样本在具体变量上可能与电访一致。我们在数据集中用 batch 表示了不同批次的数据。有关 batch 变量的详情请见本报告第 4 部分：特殊变量解释。

			及配件的费用分别采用总、分的提问方式。
P507	FP507_T	FP506 FP507	电访和面访针对各类办公类电器、家具和其他耐用消费品支出金额分别采用总、分不同提问方式。
P509	FP509_T	FP510 FP509	电访和面访针对直接支付的医疗支出及保健费用支出金额分别采用总、分不同提问方式。
P511	FP511_T	FP404 FP511	电访和面访针对美容消费及雇佣保姆、小时工、佣人等各项服务支出金额分别采用总、分不同的提问方式。
P515	FP515_T	FP516 FP515	电访和面访针对购买商业性医疗保险及商业性财产保险支出金额分别采用总、分不同的提问方式。
R 部分 其他房产			
R2	FR2A FR2B FR2C FR2D FR2E	FR2A_A_(1-10) FR2A_B_(1-10) FR2A_C_(1-10) FR2A_D_(1-10) FR2A_E_(1-10)	电访和面访分别针对房产方面的提问。电访是其他家庭成员拥有其他房产，面访是“离您家最近”／“您家的”这套房子。
R2CKP	FR2CKP	FR2CKP_A_(1-10)	同 R2 题
R3	FR3	FR3_A_(1-10)	同 R2 题
R4	FR4	FR4_A_(1-10)	同 R2 题
R5	FR5	FR5_A_(1-10)	同 R2 题
R501	FR501M	FR501M_A_(1-10)	同 R2 题

表 3　2012 年成人问卷电访和面访不完全匹配问题信息列表

题号	电访有效变量	面访有效变量	不匹配类型
P509	SP509A	QP509A QP510	电访和面访针对医疗费用分别采用总、分的提问方式。
	I403	I5 系列变量	电访和面访分别采用总、分的提问方式。

3. 个人问卷中长短问卷的合并

CFPS2012 在成人与少儿主体问卷（长问卷）的基础上，添加了各自的短问卷。短问卷主要有两种用途：1) 非核心家庭成员只需完成短问卷；2）基因成员外出时，由在家的其它

家庭成员先帮其完成代答的短问卷，在异地追访到个人时，再由其自身完成个人长问卷。由短问卷的适用范围可知，在 2012 年问卷库中包括三种长短问卷类型结合，一种是只有长问卷的，这是个人库中大部分观测的情况；另一种是只有短问卷中，其中包括非核心家庭成员问卷（这部分人员因为并不是 CFPS 定义中的家庭成员，故在发布时删除）及异地追访不成功的个人；第三种是既有长问卷又有短问卷的，这些人是外出的基因成员并且在异地追访成功的。成人库中只有长短问卷的比例分别是 87.9%， 7.2%， 剩下 4.9%的人既有长问卷，又有短问卷。而在少儿库中，这三个比例分别是 95.6%，2.8%以及 1.6%。

由于短问卷是代答问卷，因此在问题设置上要避免那些只能由受访者自答的问题（如主观题、认知题）。短问卷中的绝大部分问题是直接从长问卷中摘选的，运用的原始变量名也相同，在合并时只需要考虑当长短问卷均有时，用自答的长问卷数据覆盖代答的短问卷数据即可。需要额外注意的问题有两点。一是当长问卷答案为缺失，而短问卷存在有效值时，用短问卷的数值作为最终值；另外，部分问题存在长短问卷不完全匹配的情况，需要用不同的变量分别记录。表 4 为成人问卷中长短问卷不完全匹配问题列表，其余经过确认长短问卷内容一致的问题已经在数据整理过程中将其变量名进行了统一。

表 4　2012 年成人长短问卷不完全匹配问题信息列表

题号	长问卷有效变量	短问卷有效变量	不匹配类型
N401	QN401S_S_1-QN401S_S_14	SN401	长问卷针对多种组织成员类型进行多项选择，短问卷是单选题形式，只关注是否是中国共产党党员。
P509	QP509A	SP509A	长问卷针对住院花费中自家支付部分提问；短问卷针对伤病花费中自家支付部分提问。
G401	QG401	SG401	长问卷针对挣工资的非农工作提问；短问卷针对所有非农工作提问。

4. 综合变量的添加

除了从问卷中直接生成的变量之外，CFPS 发布数据中还包括了部分项目团队人员基于问卷变量后期生成的综合变量。这些综合变量的基本情况如下：

1） 家庭收入（家庭经济库）

家庭收入包括总的家庭收入、人均家庭收入和具体分项收入（如家庭工资性收入、经营性收入、转移性收入、财产性收入等）。由于 CFPS2012 在采集家庭收入方面的设计与 CFPS2010 有所不同，为方便用户使用，我们另生成了可用 2010 年比较的估计值。

2） 家庭支出（家庭经济库）

家庭支出包括家庭总支出以及分类别的支出，其中总支出考虑到居民消费性支出、转移性支出、福利性支出、以及建房购房贷款支出。

3） 城乡状态（家庭经济库、成人库、少儿库）

在 2012 年各库中，我们不仅提供了各村居根据国家统计局的城乡类型划分（urban12），还提供由访员在村居现场记录的村居类型（“城市”和“乡村”两种类别由变量 urbancomm 标识），并进一步通过访员对村居的具体情况的选择将其细分类为“城市”、“城镇”和“农村/城郊村居”（由 typecomm 变量标识）。

以上三类变量如表 5 所示。

表 5 家庭经济库中综合变量

变量名	变量标签
家庭收入部分	
WAGE_1	工资性收入
WAGE_2	工资性收入（与 2010 年可比）
wage_1_adj	工资性收入-调整
wage_2_adj	工资性收入-调整（与 2010 年可比）
foperate_1	经营性收入
foperate_2	经营性收入（与 2010 年可比）
ftransfer_1	转移性收入
ftransfer_2	转移性收入（与 2010 年可比）
fproperty_1	财产性收入
fproperty_2	财产性收入（与 2010 年可比）
FELSE_1	其他收入
FELSE_2	其他收入（与 2010 年可比）
FINCOME1	2011-2012 全部家庭纯收入
FINCOME2	2011-2012 家庭纯收入(2010 可比)
fincome1_adj	2011-2012 全部家庭纯收入-调整
fincome2_adj	2011-2012 家庭纯收入-调整(2010 可比)

fincome1_per	2011-2012 人均家庭纯收入
fincome2_per	2011-2012 人均家庭纯收入(2010 可比)
fincome1_per_adj	2011-2012 人均家庭纯收入-调整
fincome2_per_adj	2011-2012 人均家庭纯收入-调整(2010 可比)
fincper_p	家庭人均收入分位数
fincperadj_p	家庭人均收入分位数-调整
家庭支出部分	
PCE	居民消费性支出-加总
FOOD	食品支出-调整
DRESS	衣着支出
HOUSE	居住支出-调整
DAILY	家庭设备及日用品支出-调整
MED	医疗保健支出
TRCO	交通通讯支出-调整
EEC	文教娱乐支出
OTHER	其他消费性支出
EPTRAN	转移性支出
EPWELF	福利性支出-含插补
MORTAGE	建房购房贷款支出-估计
EXPENSE	家庭总支出
其它	
typecomm	居住社区类型
urbancomm	城乡(访员观察)
urban12	国家统计局划分的城乡类型

4) 个人收入（成人和少儿库）

个人收入包括调整前(income)及调整后(income_adj)的个人收入，其中调整后的个人收入对其中就业状态为受雇但工资性收入为 0 或缺失的个人进行了插补。

5) 受教育水平（成人和少儿库）

受教育水平包括已经完成的最高学历水平（edu2012）以及当前的教育年限(sch2012)。

6) 就业状态（成人库）

成人的就业状态分为“在业”、“失业”和“退出劳动力市场”三种。

7) 认知水平（成人和少儿库）

2012 年认知测试包括测量记忆的单词回顾题和数列逻辑题，其中数列逻辑题采用分阶段适应性设计，也即第一阶段的得分决定了第二阶段试题组的难度。成人和少儿库中分别提供了单词和数列题的综合得分。由于数列题设计的复杂性，我们提供了基于原始问题绝对排序的 Guttman 计分法的得分，以及基于项目反应理论模型(Item Response Theory)中 Rasch 模型的 W 分数。

以上四类变量如表 6 所示。

表 6　成人与少儿库中综合变量

变量名	变量标签
个人收入部分	
INCOME	个人收入
INCOME_adj	个人收入（调整后）
教育水平部分	
edu2012	已完成的最高学历
sch2012	当前教育年限
认知水平	
NS_G	数列题：Guttman 计分法(0-15)
NS_W	数列题：Rasch 模型计分法 W 分数
NS_WSE	数列题：W 分数的标准误差
IWR1	短期记忆得分：第一次尝试
IWR2	短期记忆得分：第二次尝试
IWR	短期记忆得分：两次加总
DWR	长期记忆得分
其它	
employ	就业状态

注：这些综合变量的具体生成规则将在随后发布的其它相关技术报告中有详细介绍。

5. 最佳变量

CFPS2012 有多个以_best 结尾的变量名，这是最佳变量的标识。这些变量主要是针对以下两种可能出现的状况生成：1）当同一种变量可能出现不同来源时：如性别、年龄等信息可能从家庭成员库中获取，也可能从个人库中的自答信息获取；2）在数据清理的过程中发

现了一些极端不合理的数值。对于第一种不同信息来源的情况，我们的基本原则是自答优于代答，新数据优于旧数据；但在处理过程中如果发现有逻辑不合理的地方，则会综合考虑其它信息。对于第二种情况，我们会提取与该变量相关的其它信息进行综合判断，给出一个最合理的值，但同时保留原始值。譬如 CFPS2012 出现了部分房产值疑似过大或过小的观测，我们通过住房类型、住房面积、室内设施、家庭收入与支出等辅助信息来判断是否有可能存在单位错误的问题。有关财产的后期数据整理详情可参见 CFPS2012 技术报告《中国家庭追踪调查 2012 年和 2010 年财产数据技术报告》。

6. 职业编码和行业编码

CFPS2012 采集了受访者的详细工作信息，涵盖了自家农业生产活动、农业打工、受雇、非农自雇以及家庭帮工。工作信息相关变量很多是文字信息，出于以下两点考虑，这些原始变量不在数据库中发布：1）涉及到与隐私相关的具体工作单位信息；2）文字所含信息对于多数分析者来说难以直接运用。因此我们组织工作人员对这些原始的信息进行职业和行业的编码，生成不包含隐私信息且较方便分析的数据。具体的生成规则见《中国家庭追踪调查 2012 年职业编码报告》。

由于 2012 年的工作模块设计与 2010 年有所不同，为了方便大家进行跨年数据分析，我们另生成了一个综合变量：2012 年当前主要职业（job2012mn_occu）。 这个变量的生成规则可参见 CFPS2012 技术报告：《中国家庭追踪调查 2012 年综合变量：当前主要工作》。

7. 地址编码

与 2010 年基线相同，2012 年各库的地址信息给出了三级编码：省码、区县码和村居码，其中省码是国标码，用户可以知道是哪个具体省，但区县码和村居码均为顺序码。CFPS 另为用户准备了区县层级的一系列宏观变量，详情请参考 CFPS2010 技术报告《中国家庭追踪调查区县数据库模糊化方法》。

三． 2012 年清理难点

1. 跨年个人 id 库的建立

2012 年 CFPS 第一次实施全国追踪调查。跨年个人 id 库的目标是记录每一个曾经作为 CFPS 个人样本的去留基本信息，其中包括其在每一年的受访状态、其所在家户以及成员类型（基因成员、核心成员）。

构建个人 id 库的第一个难点在于个人 id 在不同年份的匹配。对于追踪调查来说，同一个人有固定且唯一的 id 是基础，但达成这一目标需要后期的清理。由于 CFPS 2012 年调查时所用的原始个人 id 是由 2012 年家户号加个人户内编码组成的，对于那些在两次调查中家户号有所变动的个人来说，其 id 在两次调查时是不同的。对于这些人我们需要首先匹配上 2010 年其所在的原家户号。按照设计，2010 年已属于 CFPS 家庭成员的个人无论其是否进入新家户，其在 2012 年个人户内码不会改变。因此，对于改变家户的个人来说，其匹配到 2010 年的个人 id 可以由其 2010 年原家户号加上其个人户内码得到。对于新进入 CFPS 的个人来说，其 id 维持 2012 年家户号加上其个人户内码。

构建个人 id 库的第二个难点是区分家庭成员在不同调查年份的访问状态。个人 id 库以 2010 年界定出的 57,155 个基因成员为基础，将这个基础库与 2012 年家庭成员库中的成员列表部分进行合并，在合并库中的个人大体可分为三类：第一类个人在 2010 年基因成员库中，且在 2012 年有其家庭层面信息，数据库有其在 2010 年与 2012 年的基本信息；第二类个人在 2010 年基因成员库中，但在 2012 年未追访到其任何相关家户，数据库有其 2010 年的基本信息，但 2012 年的状态缺失；第三类个人不在 2010 年基因成员库中，2012 年由于婚姻、出生等原因进入 CFPS 样本家庭，数据库有其 2012 年的基本信息，但 2010 年状态缺失。在这里，尤其要注意的是在 2012 年有家庭层面信息并不意味着存在个人层面的有效的成人问卷或少儿问卷。

2. 家庭成员关系库的构建

2012 年家庭成员关系库相比 2010 年多了另组家庭和原家庭的联系，一个成员可能与多个样本家庭之间有联系。2012 年家庭成员关系库的创建以 2010 年家庭成员关系库为基础，将离家的个人进行标识，并在新家庭中更新其家庭关系。特别值得指出的是对于由原家庭进

入另组家庭的个人，在家庭关系库中会出现多个记录，即其在原家庭中的一个记录（为离家状态，由 co_a12_p=0 来标识）和在新的家庭中的一个记录（为在家状态，由 co_a12_p=1 来标识）。关于家庭成员关系库构建的详细内容参见技术报告《中国家庭追踪调查 2012 年家庭成员库的分解与家庭关系库的构建》。

四． 特殊变量解释

1. 数据批次（Batch 变量）

CFPS2012 年调查共分四个批次的数据。第一批次为集中的面访数据，这部分数据占问题的绝大多数。第二批数据为针对外出打工人员春节集中返回家乡所设计的春节期间的追访。第三批数据的访问形式为面访，但由于各种原因实际采用的问卷却是电访的问卷，因此采集到的变量也是与电访问卷一致的变量；对于只在面访问卷中出现的变量，第三批面访数据均设为-8；这批数据只占很小的一部分比例，涉及到 10 个家庭经济问卷和 28 份成人问卷。第四批数据全部为电访的问卷。