China Family Panel Studies

CFPS

中国家庭追踪调查

技术报告系列：CFPS-33

系列编辑：谢宇　责任编辑：张聪

中国家庭追踪调查
2012 年家庭关系原始库的分解与重构

戴利红　孙妍　许琪　吴琼

2015.08.10

中国家庭追踪调查（CFPS）是一项全国性、综合性的社会追踪调查项目。2010 年 CFPS 进行了基线调查，每两年进行一次追踪访问，2012 年成功完成了第一次追访。CFPS2012 年家庭成员问卷以家庭为调查单位，但是不同于 CFPS2010 年家庭成员问卷的设计思路。在 2010 年基线调查时，CFPS 家庭成员问卷利用 T1 表采集同住家庭成员的性别、出生年份、婚姻状况、受教育程度等基本社会人口信息，T2 表采集家庭成员直系亲属关系，T3 表采集不同住直系亲属的基本信息。基于这些信息，我们可以建立完整的 CFPS 基线家庭关系结构（详细可见 2010 年的家庭成员问卷和技术报告）。基于 CFPS2010 年已有的信息， 2012 年进行追踪调查时，问卷设计重在采集家庭结构的变化：譬如家庭的分裂、家庭内部的成员流动，以及新进成员的家庭关系，在形式上没有保留 2010 年的 T 表（T1、T2、T3）格式，但是依然存续了 T1 表、T2 表的思想。通过数据清理，可以恢复 T1 表和 T2 表。

一、家庭成员问卷的设计理念

CFPS2010 年的家庭成员问卷采用 T 表来采集家庭关系与家庭成员信息，从而建立了一个完整、精确的家庭结构网络，它为 CFPS 之后的追踪访问奠定了基础。2012 年的家庭成员问卷以 2010 年的家庭结构为基础，通过采集新进个体与原家庭成员的血缘关系、经济联系的相关问题来判断其是否为 CFPS 定义的家庭成员和成员类型；对于离开原家庭较长时间的个体，通过询问离家的原因、与原家庭在家成员的关系、是否养家等相关问题来判断离家人员是否仍为 CFPS 定义的原家庭成员、是否需要继续追踪访问。此外，家庭成员问卷还采集需要追访的外出个人和另组家庭的地址、电话等联系方式，为继续追访提供联系信息。

（一） 基本概念

1. 家庭类型：

原家庭：

当期调查前存在于家庭成员数据库中的完访家庭。

另组家庭：

当期调查从原家庭中分裂出来，与原家庭经济上相互独立的经济体。

2. 家庭成员类型：

基因成员：

2010 年所有家庭成员及其之后的新生血缘子女或是不超过 10 岁的领养子女。

核心成员：

基因成员的非基因直系亲属（父母、配偶、子女）

非核心成员：

家庭中除了基因成员、核心成员之外的家庭成员。

三种成员类型没有包含，并且不可以转化。

针对不同的家庭成员类型，CFPS 制定了不同的追踪策略：如果是基因成员，CFPS 将永久追踪，但可能因为某些原因（如服刑、参军/服役、出家和出境），当期不追踪；如果是核心成员，则进行有条件的后续追踪，当核心成员与基因成员同时在一个家庭中，则实施追踪，如果不在一个家庭，则不实施追踪。

（二）问卷主要内容

原家庭成员确认：首先，基于 2010 年的家庭成员，请受访者确认哪些成员还在家，哪些人离家，离家的原因[1]，和家庭的联系。然后，根据以上信息来判断这些人是否是该家庭的家庭成员，并确定其在 CFPS 中的身份类型（包括在家基因成员、需要追访的离家基因成员、需要追访的另组基因成员、不需要追访的离家基因成员、不需要追访的另组基因成员、去世基因成员）。由于这些人的家庭成员关系已经在 2010 年的 T2 表中存在，所以不需要重新确认家庭关系。

新进人员身份确认：请受访者判断除了在之前已经确认的在家成员外，是否还有其他成员进入家中。新进成员分为三类：a. 2010 年 T2 表中的人（code=2XX，即基因成员的非基因直系亲戚），被认为是新进家庭的核心成员；b. 2010 年 T1_3 表中的人（code=3XX），被认为是新进家庭的非核心成员；c. 第一次进入 CFPS 调查样本的个人，通过回答一些问题后，再判断是否为家庭成员及成员类型。因为 a 成员的家庭关系已在 2010 年采集，所以在后期

[1] 离家原因：1.外出读书 2.外出打工/工作 3.出家 4.探亲访友/离家出走/2010 年经济上是一家，但是分开住 5.服刑 6.参军/服役 7.处境（包含港、澳、台） 8.分家 9.嫁出 10.去世 11.离婚 12.2010 年记录有误

数据清理时可以从家庭成员关系中恢复；但是 c 类人员是第一次出现在 CFPS 调查中，问卷中采集了这部分人员的直系亲属关系。如果亲属为 2010 年的基因成员，则从加载列表中选择，否则由访员重新采集，以此来完善 2012 年的家庭关系，为恢复 2012 年的 T2 表提供补充信息。在 c 类人员中，因新出生和领养关系进入的小孩被定义为 CFPS 的新进基因成员；对于其他新进人员，系统会根据这些人与原家庭基因成员的关系和家里的经济关系，来判断这些新进人员的类型（核心成员或非核心成员）。

已离家的新生血缘子女：采集已经离家的 CFPS 基因成员的新生血缘子女，然后判断他们是否为离家的新进家庭成员或另组基因成员。

外出人员地址采集：采集需要追访的外出家庭成员的外出地址，一方面记录外出成员的流动范围，另一方面为个人问卷的追访提供地址信息。

另组家庭信息表：根据离家人的信息来判断哪些人是从 2010 年接受调查的家庭中分离出来的，并将其分为不同的家庭单元，这就产生了 2012 年的另组家庭。同时采集各自的家庭住址，用于另组家庭成员问卷的调查。

二、家庭成员库的分解工作简要介绍及编码规则

（一）家庭成员数据库的分解

原始的家庭成员数据库是以每个受访家庭的信息为一条观测，包含了近 1.2 万个变量。我们根据问卷设计的各个模块内容，分别将原始的家庭成员数据库进行分解，生成方便使用的数据结构。最终的发布数据库以家庭中每个成员为一条观测，变量包含每个成员的基本信息和关系人信息。清理后的数据库极大地精简了原始库，也更方便用户使用。

（二）家庭样本编码规则

1．2010 年 CFPS 成功访问的家庭，在 2012 年依然是追访的家庭样本，家庭样本编码保持原 6 位码不变；

2．如果受访家庭是 2012 年的另组家庭，那么系统给其分配新的 6 位家庭样本编码。

这样的编码规则保证了在追踪调查中每个受访家庭都保持唯一的样本编码。

（三）个人层面编码规则

1. 户内 3 位码

（1）如果家庭成员是来自 2010 年访问时 T1_1 表中的成员，那么户内 3 位码保持不变（依次为 101、102、……）；

（2）如果原家庭 2010 年 T1_1 表中的成员在 2012 年访问时到了另组家庭，则依然保留原家庭中的户内 3 位码；

（3）如果新进人员是因为新生儿或是领养关系进入家庭，且父母有一方是 2010 年 T1_1 表中的成员，那么此新进成员就被定义为 2012 年 CFPS 的新进基因成员，户内 3 位码依次为 401、402、……；

（4）在另组血缘子女模块采集到的新生儿的编码需继续 3 中的编码往后编；

（5）如果新进人员不是（3）、（4）中的类型时，那么此新进人员的户内 3 位码依次为 431、432、……；

这样的编码规则保证了每个家庭中的家庭成员都有唯一的户内 3 位码，并且可以根据 3 位码的首位知道其进入 CFPS 调查的年份，从第二位获得其是否为新进基因成员的直观信息。

2. 个人编码

（1）如果家庭成员是来自 2010 年访问时 T1_1 表中的成员，无论 2012 年追访时属于原家庭还是另组家庭，依然保持 2010 年的 9 位码（家户号+户内 3 位码）；

（2）如果家庭成员是 2012 年新进成员，那么个人编码是所在家庭样本编码连接户内 3 位码生成。

这样的个人编码保证了 CFPS 样本中每个受访者都有唯一的样本编码，确保了追踪调查时的信息连续性。

三、家庭关系库的构建

（一）家庭关系表（T2 表）的建立

2010 年 CFPS 成功访问的家庭样本中，有一些家庭在 2012 年调查中没有追访成功，2012 年的家庭关系库不再包括这些未追访成功的家庭信息，因此 2012 年家庭关系库的构建思想是：在 2010 年 T2 表中那些 2012 年追访成功家庭（即产生了有效家庭成员问卷）的成员构成及其家庭关系信息的基础上，增加 2012 年新进家庭成员及其家庭关系，同时更新原家庭关系，并加入另组家庭的家庭成员关系信息，就构成了 2012 年的 T2 表（家庭关系表）。

建立 2012 年的 T2 表的具体操作如下：

1．从 2010 年的 T2 表中筛选出 2012 年有有效家庭成员问卷的家庭关系信息，建立数据集 T2_2012_v1；

2．T2_2012_v1 中的家庭是 2012 年追踪成功的 2010 年的家庭信息，其中部分家庭会分裂出新家庭并成为 CFPS 调查家庭样本，我们将其定义为 2012 年 CFPS 的另组家庭。问卷设计中不会重新收集原家庭成员的家庭关系，因此我们需要从 2010 年 T2 表中获取另组家庭的信息。其基本步骤是：先找到另组家庭和原生家庭样本编码的对应关系，再从原家庭关系中复制另组成员的信息到另组家庭中，这样就形成了另组家庭的基础家庭关系，生成了 2012 年所有受访家庭的临时家庭关系库 T2_2012_v2。

3．在 2012 年家庭成员问卷中将“新进家庭人员确认”中的新进成员分为 3 类（如一、3．中所述），从家庭成员问卷数据中分别选出该模块的数据：

对于 2010 年 T3 表（在 2010 年与该家庭成员不同住）中的成员，到 2012 年接受访问时已经成为了该家庭的成员的情况，2012 年调查时没有单独采集这些人的家庭关系，构建关系库时我们通过 2010 年原有的家庭关系和其它类型的新进成员的信息，利用程序恢复这类成员的家庭关系，将这些人加入到 T2_2012_v2，独立成行，并生成新的临时家庭关系库 T2_2012_v3。

问卷采集了第一次进入 CFPS 家庭样本的家庭成员（因出生、婚姻、非直系的血缘亲属等原因进入家庭）的家庭关系：要求受访者从原来的家庭成员和新进的人员列表中选取他们

的家庭关系人，包括父母、配偶、子女。将这些人加入到 T2_2012_v3，独立成行，生成临时家庭关系表 T2_2012_v4。

4．检查新进成员新采集的家庭关系数据的质量，发现错误或是可疑数据后，根据家庭全部的家庭关系和家庭成员名字、年龄、性别、婚姻等信息修改证据充分的家庭关系。对于证据不充分的错误，则删除局部的家庭关系，避免错误信息的误导。更新后生成新的临时家庭关系库 T2_2012_v5。

5．根据 T2_2012_v5 中 2012 年新进成员提供的家庭关系信息，来更新 2010 年原有家庭成员的家庭关系，添加、删除或是替换。实现关系人的互认，生成新的临时家庭关系库 T2_2012_v6。

6．加入“已离家的新生血缘子女”中的新进基因成员，根据上面一、6 的判断结果，这些小孩被分为原家庭中的基因成员和另组家庭中的基因成员。标记这些成员，将他们独立成行地放入到各自的家庭中，并双向地恢复他们与父母的家庭关系。生成新的临时家庭关系库 T2_2012_v7。

7. 由于 2012 年家庭成员问卷库中的新进成员的户内 3 位码的编码规则和个人问卷数据库中的规则不同，所以必须将两者统一，以方便用户使用。以个人库中的编码规则为依据，更改家庭成员库中的编码。生成新的临时家庭关系库 T2_2012_v8。

8．在上面第 2 步中，有些原生家庭中的人会作为另组家庭成员的关系人出现，为了保证他们 2012 年的最新信息能够出现在该家中，需要区分他们 2012 年是否有个人问卷：如果有，就将他们的户内 3 位码末尾添 0 变成 4 位的，并将其单独记录；如果没有，就保持不变。生成新的临时家庭关系库 T2_2012_v9。

9. 检查 T2_2012_v10 中每个家庭是否存在家庭内部 3 位码和成员姓名有重复的情况（排除 2010 年核查正常的 11 个家庭重名样本），按个人问卷库中的名字和 3 位码进行取舍，另一个 3 位码再重新赋值，从而保证编码的唯一性。生成新的临时家庭关系库 T2_2012_v10。

经过上面一系列操作，2012 年家庭关系库的 T2 表（家庭关系表）的雏形创建完成，但是还需要后期根据 T1 表中的个人基本信息来核实 T2 表的正确性，并将错误的家庭关系尽可能地更正。

从上面叙述的操作流程上可以看到，新创建的家庭成员库在 2010 年的家庭中加入了新进家庭成员，但是并没有删除那些 2012 年不与该家庭同灶吃饭的人（也即 2012 年已经分裂出去，经济上不属于原家庭的成员）。我们在后面整理的 T1 信息表的某些变量会标记是否物理离开原家庭、离家原因、离家的大致区域、是否同灶吃饭等信息，到 2014 年的家庭关系库清理时再将这些依然不在家的人删除。这样有利于数据使用者了解家庭成员组成的动态变化。基于这样的理念，2012 年的 T2 表又不同于 2010 年的 T2 表，前者是后者家庭关系的延续。

（二）家庭人员基本信息表（T1 表）的建立

上面讲了 2012 年的家庭关系库中的 T2 表不同于 2010 年的 T2 表，那么当然 2012 年家庭人员基本信息表（T1 表）也就不同于 2010 年的 T1 表的设计理念。2012 年的 T1 表更像是一个材料仓库，里面不仅包含了 2010 年的所有家庭成员、不同住个人的简单信息，还包含了 2012 年新进家庭成员的基本信息。也就是说只要在 CFPS 调查中采集过的人员都是一条观测。

建立 2012 年的 T1 表的具体操作如下：

1．从 2010 年家庭关系库中分解出 2010 年的 T1 表，作为 2012 年 T1 表的基础数据来源。生成 2012 年临时的 T1 表：T1_2012_V1。

2．整理 2012 年原生家庭样本编码和另组家庭的对应关系，将另组家庭中的原家庭成员及其关系人的信息从原家庭中复制到另组家庭中，并增加一个变量来标记另组家庭的原生家庭的样本编码。生成 2012 年临时的 T1 表：T1_2012_V2。

3．对于 2010 年的 T3 表中的新进家庭成员，他们的 code 由 2010 年的 2XX 变成 2012 年的 4XX。因此需要变更这些成员的 code，并将 2012 年的在家状态变量设置为“在家”且“同灶吃饭”。2012 年调查时没有对未进入家庭的 T3 表人员的信息进行更新，因此将他们“是否健在”、“年龄（周岁）”设置为“不适用”。生成 2012 年临时的 T1 表：T1_2012_V3。

4．根据问卷“原家庭成员确认”模块数据整理出 2010 年成员仍在家的成员的列表、离家人员列表、离家原因和离家时间；从“外出人员地址采集”、“另组家庭信息表”模块整理

出离家人员大致的现住区域，用这些家庭原成员的信息来更新 T1_2012_V2 中的在家状态、离家时间、访问时所在区域等信息。生成 2012 年临时的 T1 表：T1_2012_V4。

5．2012 年的个人确认问卷对所有 2012 年追访到的个人样本的性别、出生年份进行再次确认，从这个问卷数据库中整理出最新的性别和出生年份的变量。一方面用这些数据更新 T1_2012_V4，另一方面用 2012 年的新样本对 T1_2012_V4 中的样本量进行扩充。生成 2012 年临时的 T1 表：T1_2012_V5。

6．从 2012 年的成人、少儿问卷库中整理出个人的最高学历、户口类型，添加到 T1_2012_V5 中。生成 2012 年临时的 T1 表：T1_2012_V6。

7．从家庭成员问卷数据中分别整理出离家原因是离婚、嫁出、去世的人，再从个人问卷中整理出 2012 年调查时的婚姻状态。前者可以弥补没有生成个人问卷的个人调查时的婚姻状态，其中去世人的婚姻状态是缺失的，其配偶如果没有产生个人问卷，那么就是“丧偶”的状态。将整理的信息添加到 T1_2012_V6，生成 2012 年临时的 T1 表：T1_2012_V7。

8．根据（一）中建立 T2 表的第 8 步的操作，从 T2_2012_v9 中整理出另组家庭中有 4 位码的个人，保留他们的原生家户号、新家户号和户内码，并将这些人在原家庭中的信息复制到另组家庭中，并且将这些人是否在家的状态改为不在家、不同灶吃饭。生成 2012 年临时的 T1 表：T1_2012_V8。

9. 检查 T1_2012_V7 中变量缺失的情况并查明原因，将确认没有信息的变量赋值为-8，有信息补充的变量再用对应信息填充。生成 2012 年临时的 T1 表：T1_2012_V9。

（三）生成初始版 2012 年家庭关系库

将家庭人员基本信息表（T1 表）与 2012 年的家庭关系表（T2 表）进行合并，将个人信息放到对应家庭成员及其关系人的后面。比如家庭某个成员，他可能与多个人存在血缘或是婚姻关系，那么他的个人信息就会在他出现的每个位置上。这样就生成初始版本的 2012 年家庭关系库（T_2012 数据库）。它的一个观测是 2010 年调查时的一个家庭成员或是 2012 年的新进成员，变量是每个成员的相关信息。在该数据库中同时存在个人及其家庭关系人的个人基本信息。

四、检查并更新数据库

上面的工作主要是对分散信息的重组，还需要通过数据检查工作来验证数据是否符合逻辑。如果发现问题，再通过 2012 年、2010 年的原始数据和执行反馈信息进行更正。

（一）个人层面的核查

在家庭关系库中，结合成员的年龄、性别、婚姻和家庭关系人及其信息之间的逻辑关系，可以发现存疑数据。

1. 年龄：

（1）筛选较可能存疑的观测：夫妻的年龄差大于等于 10 岁的，父母至少有一方的年龄和孩子的年龄差小于 15 岁的。

（2）检查（1）中出现上述现象的原因：人工查看这些家庭的家庭关系，如果是家庭关系不清晰，就改正 T2 表中的家庭关系；如果家庭关系正确，有必要就查看 2010 年、2012 年的原始数据是否一致。若两年的数据一致，就认为数据本身是正确的；若不一致，再根据具体情况修正 T1 表，当没有合理的理由时，则以 2012 年的数据为准。

（3）根据修正后的 T1、T2 表，重新生成家庭关系库 T_2012_v2。

2. 性别：

（1）筛选存疑观测：母亲的性别是男或父亲的性别是女；夫妻的性别一样。

（2）检查（1）中出现上述现象的原因：人工查看这些家庭的家庭关系，如果是家庭关系不清晰，就改正 T2 表中的家庭关系；如果家庭关系正确，再结合 2010 年、2012 年的原始数据加以判断，并修正 T1 表。

（3）根据修正后的 T1、T2 表，重新生成家庭关系库 T_2012_v3。

3. 婚姻：

（1）检查配偶彼此是否互认，将不互认的家庭逐一提出。

（2）查找配偶不互认的原因：通过查看夫妻双方的在家状态、离家原因、婚姻、和孩子的关系来确认是关系匹配错误还是婚姻重组造成的。根据错误原因合理地更正 T1、T2 表。

（3）根据修正后的 T1、T2 表，重新生成家庭关系库 T_2012_v4。

（二）家庭层面的核查

1．检查同一个人是否同时存在于原家庭和另组家庭之中（co_a12_=1），根据原始问卷数据进行判断，最终将其归属于更符合问卷设计的那个家庭，重新生成家庭关系库 T_2012_v5。

2．检查同一个人是否存在多个个人样本编码。比如：原家庭和其另组家庭都说这个人是各自家庭的新进成员；或者原家庭成员又被当作另组家庭的新进成员，根据个人样本生成的规则，此时同一成员将错误地生成两个不同的编码。更改时，我们根据执行反馈或是实际产生问卷信息来确认新进成员所属家庭和个人编码，若 2010 年已经产生个人编码，则保留原有的，确保编码的唯一性和一致性。重新生成家庭关系库 T_2012_v6。

（三）跨库的核查

1．将 2012 年家庭关系库与家庭经济数据库中的家庭样本编码 fid12 进行匹配，主要目的是确保另组家庭编码在两个数据库中的一致性。

2. 将家庭关系库中的家庭编码 fid12、个人编码 pid 与个人问卷数据库进行匹配。对于匹配失败的情况，是由成员归属家庭错误、个人编码不一致造成的。查找依据进行更正，保证个人库全部与家庭关系库对应。重新生成家庭关系库的 T_2012_v7。

五、添加个人和家庭属性的变量

（一）添加个人属性的变量

根据 2012 年家庭成员问卷设计的理念，2012 年家庭关系库会体现 2010 年家庭成员在 2012 年调查时哪些依然在该家庭中、哪些人已经流出、新成员流入情况。流出的成员要判断 2012 调查时是否需要追访，追访的成员是物理离家还是已经产生另组家庭；新进成员是基因成员、核心成员，还是非核心成员。这就使得同一个另组到新家庭的成员在不同家庭中的身份属性变量不同。

1. 基因成员类型

根据问卷中“原家庭成员确认”和“新进人员身份确认”模块的变量，来判断家庭中 CFPS 基因成员类型[2]，将其分为 8 种，生成变量 genetype，它的取值和值标签如下：

genetype=1: 在家的 CFPS 基因成员

genetype=2: 新进 CFPS 基因成员：基因成员的新生子女和不超过 10 岁的领养子女

genetype=3: 需要追访的外出 CFPS 基因成员

genetype=4: 需要追访的另组家庭的 CFPS 基因成员

genetype=5: 不需要追访的外出 CFPS 基因成员

genetype=6: 不需要追访的另组家庭的 CFPS 基因成员

genetype=7: 死亡 CFPS 基因成员

genetype=0: 不是 CFPS 基因成员

2. 是否是核心成员

根据问卷中“新进人员身份确认”模块的变量来判断与基因成员的亲密性，如果新进成员是基因成员的非基因直系亲属（父母、配偶、子女）那么他是核心成员。那新进成员中既

[2] CFPS 基因成员类型的判断条件请参见中国家庭追踪调查 2012 年面访问卷《家庭成员问卷》中的“CFPS2012 个人问卷生成规则”表。

不是基因成员又不是核心成员的家庭成员则是非核心成员。生成是否核心成员的变量coremember:

coremember=1: 核心成员

coremember=0: 非核心成员

coremember=-8: 不适用

（二）添加家庭属性的变量

1. 同灶吃饭总人数

根据上面 T_2012 表生成的规则，另组基因成员、不需要追访的基因成员（如：去世、入狱、出国、出家人员等）依然会显示在 2012 年家庭关系库中，但是他们已经不是家庭成员，所以 2012 年家庭成员的人数不是将家庭中的每个观测累加，而是将家庭中同灶吃饭的人（co_a12_p=1）的成员累加，并生成变量 fam_sum（同灶吃饭的总人数）。

2. 同灶吃饭中基因成员人数

根据基因成员的七种类型，同灶吃饭中基因成员人数包含：在家的原基因成员（genetype=1）+新进基因成员（genetype=2）+需要追访的外出基因成员（genetype=3），生成变量 gene_n（同灶吃饭中基因成员人数）。

3. 同灶吃饭中核心成员人数

根据同灶吃饭的变量（co_a12_p=1）、核心成员的变量（coremember=1），将每个家庭中符合这两个条件的成员累加，生成变量 core_n（同灶吃饭中核心成员人数）。

4. 父母背景的综合变量

为了补充家庭关系库中父母年龄、最高学历信息缺失，在 2012 年成人问卷中以子女代答的方式，采集了父母的这些信息。存在的主要缺点是，如果存在婚姻重组，父母信息可能有不统一的情况。我们将新采集的数据与原有数据相比较、整合，生成父母的出生年(fbirthy12，mbirthy12)、父母最高学历(feduc12，meduc12)的综合变量。制定的生成原则：父母自答信息优先，其次是子女代答的，然后是其他成员代答的。

5. 家庭代际数

根据家庭成员（同灶吃饭 co_a12_p=1）的家庭关系，生成家庭代际数变量 generation，生成规则如下：

Generation=1：一代户，单人家庭、兄弟姐妹、夫妇为同一代人

Generation=2：二代户，父母（岳父母、公婆）与子女（儿媳、女婿）为两代人

Generation=3：三代户，爷爷、奶奶、外公外婆与孙子女为三代人

Generation=4：四代户，如果出现隔代，按最高代与最低代之差来计算（如爷爷奶奶与孙子女同住，中间父母这一代不住在家中，算三代）

Generation=5：五代户

Generation=79：其他，如果家庭关系库中有某个家庭成员与其他所有家庭成员都不存在婚姻和直系血缘关系，则该户家庭标记为其他。

六、数据库变量及标签

变量模块	变量名	变量顺序	变量标签
	NAME	VARNUM	LABEL
成员编码	pid	1	个人样本编号
家庭编码	fid12	2	2012 年家户号
	fid10	3	2010 年家户号
家庭人员的信息变量名	code_a_p	4	2012 年个人家庭内部编码
	cfps_interv_p	5	个人产生有效问卷的情况
	TB1A_A_p	6	个人属相
	TB1B_A_p	7	2012 年个人年龄
	TB1Y_A_p	8	个人出生（年）
	TB1M_A_p	9	个人出生（月）
	TB2_A_p	10	个人性别
	ALIVE_A_p	11	2012 年个人是否健在
	deathreason_p	12	2012 年个人去世的原因

	TB3_A12_p	13	2012 年个人婚姻状况
	TB4_A12_p	14	2012 年个人最高学历
	qa301_a12_p	15	2012 年个人户口类型
	qa302_a12_p	16	2012 年个人户口所在地
	TB6_A12_p	17	2012 年个人是否在家住
	TB601_A12_p	18	个人离家的原因
	outpers_where12_p	19	离家人（个人）的居住区域
	TB602ACODE_A12_p	20	离家人（个人）的省国标码
	leavingtime_y_p	21	个人离家的年份
	leavingtime_m_p	22	个人离家的月份
	co_a12_p	23	个人是否与该家庭同灶吃饭
离家人与家庭经济联系	e7_a_p	24	过去一个月外出成员是否给家里寄钱
	e701_a_p	25	外出成员给家里寄钱金额
	E8_a_p	26	过去一个月家里是否给外出成员寄钱
	E801_a_p	27	家里是否给外出成员寄钱金额
父亲的基本信息变量名	code_a_f	28	父亲家庭内部编码
	fid10_f	29	2010 年父亲的家户号
	id_f	30	父亲在调查中的样本编码
	cfps_interv_f	31	父亲产生有效问卷的情况
	TB1A_A_f	32	父亲属相
	TB1B_A_f	33	2012 年父亲年龄
	TB1Y_A_f	34	父亲出生（年）
	TB1M_A_f	35	父亲出生（月）
	TB2_A_f	36	父亲性别
	ALIVE_A_f	37	2012 年父亲是否健在
	deathreason_f	38	2012 年父亲去世的原因
	TB3_A12_f	39	2012 年父亲婚姻状况
	TB4_A12_f	40	2012 年父亲最高学历
	qa301_a12_f	41	2012 年父亲户口类型

	qa302_a12_f	42	2012 年父亲户口所在地
	TB6_A12_f	43	2012 年父亲是否在家住
	TB601_A12_f	44	2012 年父亲离家的原因
	outpers_where12_f	45	离家人（父亲）的居住区域
	TB602ACODE_A12_f	46	离家人（父亲）的省国标码
	leavingtime_y_f	47	父亲离家的年份
	leavingtime_m_f	48	父亲离家的月份
	co_a12_f	49	父亲是否与该家庭同灶吃饭
母亲的基本信息变量名	将上面父亲信息的变量名的后缀换成“m”	50--71	将上面父亲信息的变量标签中的“父亲”换成“母亲”
配偶的基本信息变量名	将上面父亲信息的变量名的后缀换成“s”	72--93	将上面父亲信息的变量标签中的“父亲”换成“配偶”
10 个孩子的基本信息变量名	将上面父亲信息的变量名的后缀分别换成“c1”、“c2”、… …、“c10”	90--313	将上面父亲信息的变量标签中的“父亲”换成“孩子 1”、……“孩子 10”
统计变量	fam_sum	314	同灶吃饭成员的总人数
	gene_n	315	同灶吃饭中基因成员人数
	coremember	316	是否是核心成员
身份类别	genetype	317	所属基因成员类型
	core_n	318	同灶吃饭中核心成员人数
决策人	tf10pid	319	家庭重大事件决策人样本编码
抽样信息	subpopulation	320	抽样子总体
	subsample	321	是否在全国再抽样样本中
父母的综合变量	fbirth12	322	父亲出生年份（综合变量）
	mbirth12	323	母亲出生年份（综合变量）
	feduc12	324	父亲最高学历（综合变量）
	meduc12	325	母亲最高学历（综合变量）
家庭结构	Generation	326	家庭代际数
发布版本	ReleaseVersion	317	发布版本

七、生成家庭关系库的案例

（一）2012 年家庭关系表的例子

表一　2012 年家庭关系 T2 表结构

变量标签	2012 年家庭样本编码	2010 年家庭样本编码	个人姓名	个人家庭内部编码	父亲姓名	父亲家庭内部编码	母亲姓名	母亲家庭内部编码	配偶姓名	配偶家庭内部编码	孩子 1 姓名	孩子 1 家庭内部编码	……	孩子 10 姓名	孩子 10 家庭内部编码
变量名	Fid12	Fid10	Name_a_p	Code_a_p	Name_a_f	Code_a_f	Name_a_m	Code_a_m	Name_a_s	Code_a_s	Name_a_c1	Code_a_c1	…….	Name_a_c10	Code_a_c10
原家庭	100000	100000	张三	101	张三爸	201	张三妈	102	-8	-8	-8	-8	……	-8	-8
	100000	100000	张三妈	102	爸爸 1	202	妈妈 1	203	张三爸	201	-8	-8	……	-8	-8
	100000	100000	张四	103	张三爸	201	张三妈	102	-8	-8	-8	-8	……	-8	-8
	100000	100000	张三爸	431	爸爸 2	-8	妈妈 2	-8	张三妈	102	张三	101		-8	-8
另组家庭	100001	100000	张三	101	张三爸	431	张三妈	102	张三妻	432	张小三	401	……	-8	-8
	100001	-8	张三妻	431	张三妻爸	-8	张三妻妈	-8	张三	101	张小三	401	……	-8	-8
	100001	-8	张小三	401	张三	101	张三妻	431	-8	-8	-8	-8	……	-8	-8

注：构建过程中有名字方便信息核查，但名字属于隐私信息，所以在发布版的数据中将其删除。

由上面的数据可以看出：样本编号为 100000 的家庭在 2010、2012 年都接受了访问，张三（101）是 2010 年 100000 的家庭成员，但是 2012 年因为娶妻分家，离开了原家庭，到了另组家庭(样本编码记为 100001)，且新进了张三的配偶：张三妻（431），期间有了儿子：张小三（401）。例子 1 中家庭 100000 中，新进了张三的父亲：张三爸（431），2010 年时张三爸（201）不同住，但是 2012 年回家了，则重新给了户内 3 位码 431，并通过其他成员的家庭关系恢复了张三爸的家庭关系人信息；另组家庭 100001 中，张三的父母关系人是从 100000 中复制过来的，配偶和孩子的关系人是根据 2012 年采集的信息添加的，张小三和张三妻的关系人是 2012 年新采集的。因为张三妻的父母与他们不同住，所以是没有编码的。这个较复杂的案例中，张三从原家庭到了另组家庭中，为了保存这种流动性，他在两个家庭中都有一条观测。

（二）2012 年家庭人员基本信息表的例子

表二　2012 年家庭人员基本信息 T1 表结构

个人 id	2012 年家户号	2010 年家户号	户内编码	年龄变量	性别	是否健在	去世原因	婚姻状况	最高学历	户口类型	户口所在地	是否在家住	离家信息	是否同灶吃饭	最近一次访问时间	样本代码（原始）
pid	Fid12	Fid10	Code_a_	……	TB2_a_	Alive_A_	Deathreason_	TB3_A12_	TB4_A12_	Qa301_a12_	Qa302_a12_	TB6_A12_	……	Co_a12_	Interviewyear_latest_	Id12_
……	……	……	……	……	……	……	……	……	……	……	……	……	……	……	……	……
100000101	100000	100000	101	……	1	1	-8	1	6	1	-8	0	……	0	2010	-8
100000102	100000	100000	102	……	0	1	-8	2	3	3	-8	1	……	1	2012	100000102
100000103	100000	100000	103	……	1	1	-8	1	4	1	-8	1	……	1	2010	-8

100000431	100000	-8	431	… …	1	1	-8	2	4	1	-8	1	… …	1	2012	100000431
-8	100000	100000	201	… …	1	1	-8	1	5	1	-8	1	… …	1	2010	-8
-8	100000	100000	202	… …	1	1	-8	5	3	1	-8	1	… …	1	2010	-8
-8	100000	100000	203	… …	0	0	1.1.1	-8	2	1	-8	0	… …	0	2010	-8
100000101	100001	100000	101	… …	1	1	-8	2	6	1	-8	1	… …	1	2012	100000101
100001431	100001	-8	431	… …	0	1	-8	2	6	1	-8	1	… …	1	2012	100001431
100001401	100001	-8	401	… …	1	1	-8	1	1	1	-8	1	… …	1	2012	100001401
100000102	100001	100000	102	… …	0	1	-8	2	3	3	-8	0	… …	0	2012	100000102
… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …

表中的信息是家庭 100000 和 100001 中所有人员的信息，包括家庭成员和不同住的直系亲属，特殊说明如下：

1．张三（100000101）会在两个家庭中都出现（fid12=100001，fid10=100000），2012 年他是 100001 中的家庭成员，在家同灶吃饭（TB6_A12_=1，Co_a12_=1），因此在 100001 家中产生个人问卷，在 100000 中的状态是不在家且不同灶吃饭（TB6_A12_=0，Co_a12_=0），“最近一次访问时间”为 2010，其他信息保留；在 100001 中的“最近一次访问时间”为 2012，其他信息为 2012 年更新后的信息，明显看出“婚姻状态”不同。

2．户内编码为 2XX 的人是不同住的直系亲属（Co_a12_=0）），不需要产生个人问卷，因此样本编码 pid=-8。

3．根据表一可知：在家庭 100001 中，张三妈（100000102）不是 100001 中的成员，但是因为是张三的母亲，所以她的信息被复制到了该家中，是不同住的直系亲属（TB6_A12_=0，Co_a12_=0）。

4．因为 100000431 是 2012 年新进家庭成员，所以 2010 年家庭样本编码 Fid10=-8，其他新进成员类似。

（三）将 T1 表与 T2 表合并成 T_2012 数据库的例子

表三　2012 年家庭关系库的变量结构

2012 年家户号	2010 年家户号	个人户内 3 位码	来自表二中个人基本信息	父亲户内 3 位码	来自表二中父亲基本信息	母亲户内 3 位码	来自表二中母亲基本信息	配偶户内 3 位码	来自表二中配偶基本信息	孩子 1 户内 3 位码	来自表二中孩 2 基本信息	孩 2—孩 9 的相应变量	孩子 10 户内 3 位码	来自表二中孩 10 基本信息
Fid12	Fid10	Code_a_p	Person_information（一系列变量）	Code_a_f	father_information（一系列变量）	Code_a_m	mother_information（一系列变量）	Code_a_s	spouse_information（一系列变量）	Code_a_c1	Child1_information（一系列变量）	……	Code_a_c10	Child10_information（一系列变量）
……	……	……	……	……	……	……	……	……	……	……	……	……	……	……
100001	100000	101	张三的信息	201	张三爸的信息	102	张三妈的信息	431	张三妻信息	401	张小三信息	……	-8	全为-8
100001	-8	401	张小三信息	101	张三的信息	431	张三妻信息	-8	全为-8	-8	全为-8	……	-8	全为-8
100001	-8	431	张三妻信息	-8	全为-8	-8	全为-8	101	张三的信息	-8	全为-8	……	-8	全为-8
……	……	……	……	……	……	……	……	……	……	……	……	……	……	……

表三是根据表一和表二生成的，此处显示家庭 100001 的信息。由于家庭样本和个人户内 3 位码的编码规则确保了每个人有唯一的代码，而且这两个变量都在表一和表二中，因此可以作为链接变量，将表二中的个人基本信息添加到表一对应家庭关系人的位置上。

八、数据库变量使用说明

（一）家庭关系库内部使用

上面的几个部分详细介绍了2012年家庭关系库的结构。在家庭层面上，可以根据变量fid12识别出每个家庭，再根据pid和co_a12_p识别出同灶吃饭家庭成员或是经济独立的外出人员，pid和b6_a12_p识别出在家居住和物理外出的家庭成员；同时可根据变量genetype，可以判断家庭成员在两次调查期间的流动性；2012年家庭编码fid12不同于2010年家庭编码fid10的家庭，表示前者是从后者家庭中分裂成的另组家庭，这样家庭中会有流动成员的重复观测，但是他们的家庭状态变量co_a12_p的取值不同的。在个人层面上，研究个人离不开他所处家庭关系人，通过关系位置上的编码可以识别其父母、配偶、子女，分别根据变量tb6_a12_（后缀）、co_a12_（后缀）可以知道该成员是否和父母、配偶、子女在物理上同住，是否为经济上的家庭单元；根据变量tb601_a12_（后缀），知道是离家的原因；根据变量outpers_where12_(后缀)，可以知道离家外出地点与原家庭住址的远近；若在离开本省，根据变量 b602acode_a12_（后缀），可以知道所在的省份。

家庭关系库会有原家庭和另组家庭的区别与联系。如果2012年家庭编码fid12与2010年家庭编码 fid10（非“-8”）相同，则表示是原家庭（基线家庭）；如果不同，那么这类家庭是由原家庭分裂出来的另组家庭。用另组家庭的fid10与全库中的fid12进行匹配，就找到了两个家庭完整信息。

（二）与其他数据库的跨库使用

家庭关系库确定了每个家庭中成员的构成、家庭关系、流动人员的离家信息。CFPS2012还包含家庭经济库、成人库、少儿库。因此要全面了解个人和家庭之间的紧密联系，跨库使用数据是必不可少的。

1．与成人库和少儿库：由于在家庭关系库中完访的另组成员在原家庭和另组家庭中各有一条观测，因此和个人数据库进行匹配时，应该同时使用2012年家庭编码fid12和个人编码pid。这样就知道了个人的家庭关系及关系人的基本信息。如果还需要了解关系人的详细信息，

就需要用关系人的编码再次匹配了；这时用家庭和个人编码匹配时，缺点是只能匹配到同一个家庭的关系人的个人问卷信息，而匹配不到不同住关系人的。

2．与家庭经济库：这两个数据库分别包含了家庭层面的人口和经济方面的详细信息，直接用 2012 年的家庭编码 fid12 匹配，就建立了两个库的联系。