China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-40

系列编辑：谢宇　责任编辑：赵逸文

中国家庭追踪调查
跨年个人核心变量库清理报告

吴琼 戴利红 甄祺 谷丽萍 王祎睿

2021.04

CFPS 从 2010 年基线调查以来，每两年实施一次全国范围的追踪调查，每轮追踪调查都有新加入的家庭成员，也有离开的家庭成员。每个轮次的数据集只会包含当期调查成功接触的家户（也即存在有效家庭成员问卷的家户），而当期没有成功访问的样本或者是全家去世的样本则不在这些数据集中。为了方便 CFPS 数据用户了解所有 CFPS 样本在每一个轮次的访问状态，我们从发布 CFPS2012 相关数据时起开始创建 CFPS 跨年个人状态库，提供曾经进入 CFPS 的所有个人样本的基本人口学信息和每个轮次的访问状态。在发布 CFPS2016 相关数据时，我们又在之前的基础上添加了数据分析中常用的一些核心变量，包括教育、婚姻状态、就业状态、城乡状态等。自 CFPS2018 开始，我们在发布新一轮次的数据时会更新跨年个人核心变量库，包括截至到当期发布数据的所有 CFPS 个人样本的相关信息，以此取代之前发布的所有跨年库。

在整合跨年数据的过程中，我们对一些变量的跨年一致性进行了核查。对于跨年间出现不一致信息的观测，如果我们结合其他信息能够形成较为明确的判断，我们则对相关数据进行调整，改进跨年间的一致性，因此跨年库的样本以及部分变量的取值与已经发布的数据集可能存在一定的差异。但是由于调查数据本身的特点，我们无法将所有发现的不一致现象都强行处理，对于那些根据现有信息无法明确判断的观测我们保留了它们的原值，并在此技术报告中做了简单介绍，用户在使用时可以根据自己的研究需要进行相应处理。

一．跨年个人库的样本

跨年个人核心变量库是个人层面的数据集，也即每一条记录（数据集中的每一行）代表一位在任意调查轮次进入 CFPS 样本的个人。他们不仅包括 CFPS 基因成员，也包括核心成员和非核心成员。CFPS 的基因成员是指经 2010 年基线调查界定出来的与家庭有血缘/婚姻/领养关系的所有家庭成员，这些基因成员今后新生的血缘/领养子女同样被视为基因成员。基因成员是 CFPS 的永久追踪对象。从 2012 年追踪调查开始，基因成员在家中的非基因直系亲属（父母、配偶、子女）被定义为当期调查的核心成员；既不是基因成员也不是核心成员的家庭成员被被定义为非核心成员。CFPS 的家庭以是否存在经济联系来进行界定，并不要求成员们居住在同一地址。

此次发布的跨年库包括 2010 年基线调查开始直至 CFPS2018 完成时所有曾经进入 CFPS 调查的个人样本，共涉及 74130 位个人，其中 64028 位为基因成员，6919 位至少在一次调

查中为核心成员，剩余的 3183 位为非核心成员。

二．跨年库的变量

跨年库中的变量可以分成三类，一类是个人基线基本信息变量（譬如 pid、出生年、性别、民族、基线抽样信息等），在 CFPS 的问卷设计中，它们被认为是恒定不变的，因此它们在跨年库中只会出现一次；第二类是个人或家庭信息的跨年更新变量（譬如婚姻状态、教育水平、户口状态等），它们在每一个调查年份都有可能发生变化，因此在每一轮调查时都存在对应的变量；第三类是访问状态相关变量（譬如个人样本第一次进入 CFPS 调查的年份、每次访问时个体所在的家户号、与家户是否存在经济联系、是否居住在家庭地址、是否存在个人问卷、个人问卷是否是自答等）。

2.1 个人基线基本信息变量

这些变量包括出生年、性别、民族等。它们一般在 CFPS 调查中只会被采集一次，大多是在该样本初次进入 CFPS 调查或首次接受个人问卷访问的轮次采集。从理论上来说，这些变量只有一个真实值，不会随着时间变化，我们将其称为基线变量(baseline variables)。出于数据质量的考虑，部分变量也许会采集多次，但原则上来说，各次采集的数据应该一致。对于存在多个来源的此类变量，数据清理人员会依据一定的原则在多个来源中选择较为合理的一个数值。以下为此系列中每一个变量的生成过程。

pid：每个个体样本第一次进入 CFPS 调查时由访问系统自动生成，pid 一旦确立后便恒定保持不变，不管在后续调查中个体的状态如何变化，pid 是唯一的，因此用户可以使用 pid 在不同数据集或者不同调查轮次中链接个体层面数据集。pid 一般为进入 CFPS 调查时所在的家户号+个人户内三位代码。在少量情况下，样本进入 CFPS 调查时被原家庭成员认为是存在经济联系但居住在外的状态（也即物理离家），但离家单元自己又认定为是经济上与原家庭独立时，系统自动生成的 pid 前六位的家户号部分可能与该样本当年的家户号不一致。

出生年(birthy)、性别(gender)：CFPS 调查对于个体出生时间与性别的采集方式，在历年的访问中存在一定差异。2010 年基线调查时在家庭成员问卷中对所有成员采集出生时间与性别，如果出生时间不详还会询问受访者年龄和属相，家庭成员问卷由家庭回答人统一回答。为了提高数据的准确性，在 2010 年的个人问卷中，访员对受访者再次直接询问出生时间。

从 CFPS2012 至 CFPS2018，访员会在确认问卷或个人问卷中对满足特定条件的非初访个体信息进行更正。此外，CFPS2014 至 CFPS2018 的家庭成员问卷也会对一部分个体的信息进行补充，补充信息主要针对新进成员、离家成员以及过往轮次相关信息缺失的样本。我们在整合不同来源的数据集时，只考虑那些新采集的数据点，去除来自以往加载的数据，避免部分信息由于加载的原因重复出现。

综合过往历年信息，我们以如下的方式进行出生年和性别的整理：首先通过 pid 链接各年成员问卷及确认问卷、个人问卷新采集的信息，之后根据信息的不同来源进行数值的取舍。当多来源的信息无任何不一致时，直接取对应值。在出生年份方面，当多来源信息出现矛盾且最大和最小值之间相差超过 10 年，我们结合姓名、家庭关系、教育、婚姻状态等逐一查看，并最终确定合理值；若最大最小值不超过 10 年，优先考虑个体内部出现次数更多的数值，此外自答数据优于代答，较早年份优于较晚年份。

在整合不同来源的出生年份原始数据集中，15849 个样本只被采集过一次出生年信息，占总样本数量的 21.4%；49031 个样本存在两条记录，占 66.1%；6080 个样本存在三条记录，1738 个样本存在四条记录，33 个样本存在五条记录。在存在多个数据源的样本中，不一致的情况如下。有两条记录的观测中，1328 条存在不一致信息，占所有存在两条记录的样本的 2.7%；有三条记录的观测中,有 1092 条存在不一致信息，占比 18.0%；在存在四条记录的观测中，有 558 条存在不一致信息，占比 32.1%；有五条记录的观测中有 17 条存在不一致信息，占比 51.5%。另外，有 216 条样本的取值源自逐一查看历年数据后的综合判断，有 391 条使用了往年调查中以询问年龄或属相的方式采集的信息。此次发布的数据集中未包含出生月份数据，这是因为加上出生月份数据后，跨年一致性核查的复杂度大大增加，为了保证年份数据的准确性，我们在此版数据中集中核对了其跨年一致性。数据用户可以在 CFPS2010 年关系库或者具体年份的发布数据中查看出生月份数据。

在性别变量的清理过程中，我们查看一致性信息时仅考虑问卷回答人直接提供的性别信息，暂不考虑访员直接记录的性别，后者只在问卷回答人提供的性别信息缺失时，作为补充信息直接填补。当多来源的性别信息出现不一致现象时，若姓名未出现不同则采取多数优先于少数、较早年份优先于较晚年份的原则；若姓名出现过不同记载则进一步具体查看并结合年龄、家庭关系等信息进行赋值。在此种取值方式下，有 183 条样本出现了过往年份矛盾记载，结合对应样本的信息进行了处理。

民族(ethnicity)：CFPS 在问卷设计时对于民族信息采集基本以一次性采集为主，也即只要该

样本曾经在任一轮次提供过相关的民族信息，后期便不会对民族信息再次提问。鉴于这样的数据源，我们整理民族信息的方法如下：以 2010 年基线采集的民族信息为基础，缺失则以调查轮次的先后顺序，用采集到信息进行填补，其中少量个人样本的民族信息来源于并未发布的冗余个人代答问卷。绝大部分样本只有最多一条有效的民族信息，但其中 3,628 条存在跨轮次重复采集的情况，我们对这些样本不同来源的民族信息进行了一致性核查，发现其中有 71 条样本存在跨年信息不一致的现象。对于跨年信息不一致的样本，我们进行取值的顺序如下：优先使用 2010 基线调查采集的数据；基线信息不存在时，我们优先采用来自自答样本的信息；如果上述两种取值方式依然无法解决问题，我们则采用最新轮次采集的信息。

基线源头家户号(fidbaseline): 由于 CFPS 调查尚未进行样本补充，因此无论是基线调查界定的样本还是后续加入的样本，他们都存在一个与其关联的基线源头家庭。如果个人样本在基线调查时就存在，基线源头家庭就是该样本在 2010 年所处的家庭，如果个人样本在基线调查时并不存在，我们需要根据该样本进入 CFPS 调查时的家户号来确定其基线源头家庭。如果该家户号在基线调查时就存在，基线源头家户号就等于加入时的家户号；如果并不存在，我们则需要依次查找其上级家户，以最终确定其基线源头家户号。

基线源头家户号对应的抽样子总体(subpopulation)：CFPS 在基线抽样时有六个子样本框，分别是五个自代表性的“大省”所在的子样本框（1：上海市子总体；2：辽宁省子总体；3：河南省子总体；4：甘肃省子总体；5：广东省子总体）和一个其它地区样本所在的子样本框（6：其他省市子总体）。追踪调查时新加入的样本会根据其基线源头家户号所在的子总体来确定。

基线源头家户号是否在全国再抽样样本中(subsample)：基线抽样时，CFPS 对五个“大省”子总体实施了过度抽样。针对这五个“大省”样本框进行二次抽样后，所得的样本与“小省”样本框共同构成具有全国代表性的样本。当个人样本处在“二次抽样”的家户或“小省”样本框时，他们属于全国再抽样样本(subsample=1)，否则不属于(subsample=0)。追踪调查时新加入的样本会根据其基线源头家户号是否属于全国再抽样样本来确定。

基线源头家户号对应的 PSU（PSU）：在基线调查时，CFPS 的每个子总体通过三个阶段抽样，其中第一阶段（PSU）为行政区/县（上海地区为街道）。追踪调查时新加入的样本会根据其所在家户或上级家户在基线时基线源头家户号所对应的 PSU 来确定。

2.2 跨轮更新变量：个人基本信息

这些个人信息变量在每个轮次都有可能会被重新采集，在跨年库中每个轮次有对应的数值，变量名中用后缀体现采集的轮次。这一系列主要包含如下变量：

教育系列：这里面包括四组变量，分别是已完成的最高学历（cfps20XXedu），离校/上学阶段（cfps20XXsch），已完成的教育年限（cfps20XXeduy），已完成的教育年限-插补值（cfps20XXeduy_im）。对于年龄在 45 岁以下的受访者，我们每轮都会在个人问卷中询问他们的上学状态，这些受访者的教育系列变量的数值在不同轮次可能会有所不同。我们对跨年间的受教育状态会进行一定程度的互相校验，但并未完全要求保持一致。对于那些跨年间存在矛盾（也即后面的教育阶段低于早期）的样本，我们筛选了最高学历前后差异过大的样本（前面比后面大了两个以上等级），共涉及 100 条样本。这 100 条样本中，有 77 条样本在至少一个轮次存在个人的自答问卷，如果我们只考虑存在自答问卷的年份，53 条样本在最高学历上并无矛盾，对于这 53 条样本，我们将其矛盾的代答最高学历设置为缺失。剩余的 47 条样本亦或全部为代答数据，亦或自答样本本身在跨年间存在矛盾，我们暂未处理，保留了原值，由用户自己根据研究需要来进行数据处理。

城乡状态：这里面包括两组变量，分别是受访者居住地的城乡状态（urbanXX），以及受访者的户口状态（hkXX）。urbanXX 来自个人问卷库，因此只对存在个人问卷的受访者存在，它根据调查时受访者所在的村居在当年国家统计局中的城乡分类决定。而 hkXX 主要来自个人问卷中的受访者自报（16 岁及以上）或家人代答（16 岁以下或是只有代答问题），但 2018 家庭成员问卷对于之前年份部分个人信息进行了补充，因此此库中 2018 年的户口信息用家庭成员库信息进行了填补。

婚姻状态 (marriage_XX)：婚姻状态的信息来自于个人问卷和家庭成员问卷。针对 16 岁及以上年龄的个人问卷受访者，我们每一轮会询问其婚姻状态，来自受访者自答或家人代答。我们以 CFPS2010-CFPS2018 各轮次个人库发布的信息为基础，如果缺失则以相对应年份家庭成员问卷新采集的婚姻信息进行填补。由于各年份间成员问卷个人信息采集的筛选条件有所不同，各个年份能补充的样本规模并不一致。2010 年基线家庭成员问卷采集了所有个人（包括少儿样本）的婚姻信息，这一年份的婚姻状态相对完整；2012 年成员库没有相关设计；CFPS2014-CFPS2016 成员问卷采集了新进人员的婚姻信息；2018 不仅采集了新进人员的信息，也补充了已有家庭成员中缺失的婚姻状态。我们将各年份数据整合在一起后，对跨年间婚姻状态的逻辑关系进行了检验。跨年间婚姻状态在个人问卷设计中做出了一些逻辑限

制，譬如之前处于在婚、离异、丧偶状态的后续不可出现“未婚“状态。但是，实际数据中存在一部分不符合此逻辑的情况。如果这种不一致的现象来自于不同轮次间自答和代答信息的矛盾，我们保留了自答值，删除了矛盾的代答值。经此处理后的婚姻数据中仍有 209 条样本存在跨年间信息矛盾的情况，这其中大部分（n=167）是因为受访者在当轮访问确认上一轮次的婚姻状态时，直接予以否认并更新了上一轮次的婚姻状态。

就业状态（EMPLOYXX）：就业状态的信息来自于个人问卷。 从 2012 年开始，CFPS 基于国际劳工组织的定义，对工作状态的判断进行了较为标准化的提问方法。问卷通过一系列问题将受访者工作状态进行如下三类区分：在业、失业、退出劳动力市场。在 CFPS2012 和 CFPS2014 数据中，我们只能对自答样本的就业状态进行判断；从 CFPS2016 开始，我们不仅能对自答样本的就业状态进行判断，还能对代答样本的就业状态基于如下问题进行一定程度的判断：“QGB1 过去一周是否至少工作了 1 个小时”，如果受访者该题回答“是”，其就业状态就是在业，但是对于代答样本，我们无法对这道题回答“否”的样本的就业状态进一步确认，只能设置为缺失。

2.3 跨轮更新变量：个人访问状态

这一系列的变量主要与 CFPS 访问过程中个体或者个体所在家庭的访问状态有关，它们也是针对每轮次更新的。

个人样本进入 CFPS 年份（entrayear）：此变量记录了个人样本第一次出现在 CFPS 家庭成员关系库中的轮次。

是否在 CFPS20XX 家庭关系库中（inrosterXX）：这个变量表示个人所在的家户在相应的轮次是否完访，如果个人所在的家户完访，则个体会出现在当年的家庭关系库中。CFPS 的调查从家庭成员问卷开始，然后由完访的家庭成员问卷衍生出其它各类问卷 。因此如果家庭成员问卷未完访，其它各类问卷就没有完成的可能性。

家户号（fidXX）：虽然个人的 pid 在不同轮次间保持不变，但个人所在的家户是有可能发生变化的，譬如说分家、子女由于婚姻组建自己的小家庭、离婚等原因都有可能产生新的家户而使个体的家户号在跨年间发生变化。我们将每个个体在每一轮次所在的家户号分别列出。如果个体所在的家户成员问卷未完访（inrosterXX=0），则其相应的 fidXX 被设置为-8。

CFPS20XX 的个人数据集类型（indsurveyXX）：这个变量表示个人所在的个人数据集类型，也即个人问卷的完成问卷情况。在 CFPS2010-CFPS2016 年间这个变量有如下四类状态：“0

未完成；1 成人库；2 少儿库；-8 不适用”。不适用的意思是个人所在的家户当年没有完成家庭成员问卷；未完成表示个人所在家户完成了家庭成员问卷，但受访者未完成个人问卷。2018 年的个人问卷设计上做了调整，不再区分成人和少儿问卷，而是个人问卷和少儿家长代答问卷，因此这个变量在 0 和-8 的意义保持不变的基础上新添了如下 3 个数值：“3 同时存在于个人库和少儿家长代答库； 4 只在个人库； 5 只在少儿家长代答库“。

CFPS20XX 是否存在个人自答问卷（selfrtpXX）：从 2012 年追踪调查开始，CFPS 为了最大限度采集个人信息，对于那些无法完成个人问卷或外出的家庭成员提供了代答问卷，由熟悉其情况的家庭成员进行代答，而后再继续尝试对外出人员发放自答问卷，以提高数据的完整度。此变量表示对于 10 岁以上的样本，完成的个人问卷是否为自答问卷。

个人与 fidXX 在该轮次是否经济独立(co_aXX_p)：在跨年核心变量库中，每个个体用一行表示，他们在每个轮次对应一个家户号，如果该个体所在家户当年完成了访问，则相应年份的 fidXX 为其所在的家户 。如果个体与该家户的关系是个人在经济上属于该家户，则 co_aXX_p=1；但在一些情况下，如果个体与该家户的关系是个人在经济上不属于该家户，同时个体所在离家单元又未完成家庭成员问卷，则 co_aXX_p=0，在当期数据中该个体的家户号依然为原家庭。co_aXX_p=-8 表示该个体不在当期的家庭关系库中。

个人是否物理上居住在 fidXX（tb6_aXX_p）：对于每一位经济上属于 fidXX 的家庭成员（也即 CFPS 定义的家庭成员：co_aXX_p=1），tb6_aXX_p 变量表示该个体是否物理上在家居住（1：是；0：否）。所有 co_aXX_p=0 的成员其 tb6_aXX_p 自动赋值为 0, -8 表示个人并不在当年的家庭关系库中。

成员类型(genetypeXX，corememberXX)：2010 年进入 CFPS 的样本均为基线基因成员，也即存在于 2010 年家庭关系库中的成员全部为基因成员。从 2012 年开始，存在于家庭关系库中的成员不仅有基因成员，还有核心成员和非核心成员；而基因成员又根据其与相应家户号 fidXX 的经济联系、居住情况、离家原因等分为在家基因成员、新进基因成员、外出基因成员、另组基因成员等。2012 年家庭成员的设计与后续年份有所不同，因此基因成员的区分类型也与后续的追踪年份有差异，因此 2012 年变量名(genetype12r)及其值标签与后续年份的格式不完全一样。

个人是否健在(aliveXX)：对于每一轮次调查中 CFPS 接触过的样本，我们对于家庭成员的健在状态有一个记录，一般由家庭成员问卷的回答人提供（1：健在；0：去世）。还有少量在调查过程中由他人汇报的全家去世的情况，这些个人样本显示去世状态，但并不在当年的家

庭关系库中，因此跨年库中 aliveXX 变量显示的死亡人数比相应的家庭成员库中的死亡人数要多。与个人健在相关的还有四个变量，分别是个体是否去世”deceased”（1：是；0：否），去世年份(death_year)，去世月份(death_month)，去世原因（deathcause_code）。后面三个变量只针对那些由家人汇报去世的个体才存在，也即在当轮次调查中存在家庭关系库记录的；而对于那些由他人汇报全家去世的样本，我们没有采集去世时间以及原因信息。