China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-34

系列编辑：谢宇　责任编辑：张聪

中国家庭追踪调查
2014 年数据库介绍及数据清理报告

吴琼 戴利红 张聪 王玉磊 张文佳

2016.6.1

一． 2014 年调查问卷更新情况

中国家庭追踪调查（CFPS） 基线调查在 2010 年实施，2012 年该项目实施全国范围的首轮全样本追踪。CFPS2014 为第三轮全样本调查，集中的面访时间为 2014 年 7 月至 11 月，但由于后期有追访工作的补充以及电话调查，实际调查执行期持续到 2015 年 5 月。在此报告中，我们以 CFPS2014 来指代 CFPS 的第三轮全国调查，并不特指在 2014 年完成的 CFPS 调查。CFPS 三轮调查的问卷结构基本保持一致，但在具体模块以及具体问题的问法上可能会有调整。关于问卷设计细节上的更新，用户可以通过比较不同年的问卷得到，以下是以 CFPS2012 的问卷为基础，CFPS2014 问卷设计上的主要更新。

1. 家庭成员经济联系的双重界定

在 CFPS2012 追踪调查时，外出单元与原家庭是否存在经济联系（也即是否还属于 CFPS 定义上的同一家庭的家庭成员）由原家庭的家庭成员问卷中几个问题的结果单纯的进行系统判定：当外出单元与原家庭有经济联系时，该外出单元被判断为物理上离家但经济上仍然归属原家庭的个体，他们的家户号 fid 不会发生改变（也即 fid14 与 fid12 保持一致）；自答个人问卷在原家庭中产生，个人 pid 也不会发生改变；当外出单元与原家庭没有经济联系时，该外出单元被判断为经济上独立的个体，他们将会被分配新的家户号 fid（也即 fid12 与 fid10 不同），个人自答问卷在新家庭中产生，但其个人 pid 不会发生改变。

而在 CFPS2014 调查时，首先，由原家庭回答人判断外出单元与原家庭是否经济独立，同时系统给其分配各自不同且不同于 fid12 的单元编码，在此阶段不产生个人自答问卷；其次，当外出单元被追踪成功时，他们自我界定与原家庭的经济联系，同时在该单元中产生个人自答问卷，pid 维持不变。当外出单元认为其经济独立时，该单元被定义成 CFPS2014 的新家庭，它在各数据集中 fid14 将会与 fid12 不同，同时新家户产生相应的家庭经济问卷。当外出单元认为其经济不独立时，不产生经济问卷，且在后期清理时，将个人问卷所归属的家庭编码调整成原家庭编码。

这种经济联系界定上的更新会给数据结构带来如下变化：一是对家庭关系库的影响。CFPS2014 调查时和清理版本家庭成员的界定可能有所不同。调查时家庭成员的界定是按照家庭成员问卷回答人的结果来判定自家的成员构成，该家庭其它类型的问卷以此作为家庭成员的加载列表。然而，数据清理时家庭成员的界定需要考虑原家庭和其外出单元的双重回答，如果双方回答不一致，就会影响家庭成员的构成。当外出单元认为自己经济独立时，数据清

理才会最终定义这个外出单元是 CFPS2014 的新家庭（也称另组家庭）；否则，外出单元不是 CFPS 的独立家庭样本，而是外家庭的一部分物理外出样本。我们将其总结为下表 1，其中情况 1 和 4 原家庭与外出单元判断一致，情况 2 和 3 原家庭与外出单元判断不一致。当二者判断不一致时，还会带来第二个方面对数据的影响，即对家庭关系库的影响。

表 1　CFPS2014 中外出单元经济独立性的判断

原家庭的判定	外出单元的判定	
	独立	不独立
独立	1. 外出单元为经济独立的另组家庭，调查版本与清理版本一致	2. 外出单元为原家庭经济不独立的物理外出样本，调查版本的家庭成员少于清理版本
不独立	3. 外出单元为经济独立的另组家庭，调查版本的家庭成员多于清理版本	4. 外出单元为经济不独立的物理外出样本，调查版本与清理版本一致

由上表可知，当情况 2 发生时，外出单元的家庭成员没有被包含在调查时的原家庭成员列表中，而外出单元也不会生成独立的家庭经济问卷，因此这部分人员在家庭经济问卷的信息采集时会被遗漏。 同时，上表中的情况 3 会造成家庭经济问卷中原家庭所包含的家庭成员列表与由其生成的新家庭之间可能存在着不同程度的重叠，直接导致同个家庭成员的收入支出项目出现在多个家庭中。为了方便用户使用，我们在经济问卷中对于每一个家庭，都列出了该家庭与其相关联的家庭之间在家庭成员列表上的关系。具体的生成方法和涵义在以下对家庭经济库的介绍中将详细列出。

2. 事件日历调查(Event History Calendar, EHC)方法的引入

EHC 是指通过勾勒出时间轴线的模式，帮助受访者更加准确地回忆在一定时间段内某些事件的发生结点。它的主要目的在于增加回忆的准确性，特别是对于那些在回忆时间段内可能发生多次的事件来说（如工作状态的变化）。有关 EHC 更多的背景信息，请参考后续的 EHC 设计和清理报告。

在 CFPS2014 中，我们对成人自答问卷中的迁移、婚姻以及工作模块引入了 EHC 的设计。与往期的相应模块相比，基于 EHC 的这些模块数据最大的区别在于以下几点。一是系

统会自动检查时间跨度上的完整性，譬如从 CFPS2012 访问月份到 CFPS2014 访问月份之间是否有任何月份的状态信息缺失，并提醒访员进行追问。二是 EHC 设计会构造出一系列数组变量，对不同的地址、婚姻关系、工作的具体信息分别进行存储。三是 EHC 的题逻辑跳转更复杂，系统采集的受访者信息更加丰富。

二． 各数据库介绍

CFPS2014 全国追踪调查以 CFPS2010 和 CFPS2012 全国调查所界定出来的家庭为基础，发放的样本包括 CFPS2010 的基线家庭以及 CFPS2012 产生的新家庭样本[1]。CFPS2014 访问问卷包括家庭成员问卷、家庭经济问卷、成人问卷、少儿问卷以及村居问卷。CFPS2014 的数据库基本情况如表 2 所列。

表 2　CFPS2014 年各库基本状况

数据库	样本量	变量数
成人数据库	37147	959
少儿数据库	8617	719
家庭关系数据库	57739	278
家庭经济数据库	13946	398
村居库	621	230

1. 成人库：

成人库包括往期调查界定出来的家庭成员中 CFPS2014 调查时年龄处在 16 岁及以上的基因及核心成员， 以及 2014 年新增家庭成员中年龄处在 16 岁及以上的基因和核心成员。访问方式为面访或电访，回答人可能是受访者自己(selfrpt=1)或家人代答(proxyrpt=1)。需要注意的是，同一个成人样本，可能只有自答或代答问卷，也可能自答和代答问卷都有。自答

[1] 去除那些在往期调查中已经确认的所有家庭成员已经死亡的家庭。

问卷可以用面访(self_IWmode=1)或电访(self_IWmode=2)进行，代答问卷也可以用面访(proxy_IWmode=1)或电访进行（proxy_IWmode=2）。成人库中的个人样本包括来自 2010 年和 2012 年调查的 34751 个基因成员、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 2392 个核心成员。

2. 少儿库：

少儿库包括往期调查所界定出来的家庭成员中 CFPS2014 追踪调查时年龄处在 15 岁及以下的基因和核心成员，以及 2014 年新增家庭成员中年龄处在 15 岁及以下的基因和核心成员。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 年和 2012 年调查的 7095 个基因成员、2014 年新进的 1334 个新基因成员以及核心成员 188 人。

在 CFPS2014 中，有可能存在一个少儿出现多份代答问卷的情况。这是因为对于物理上离家的少儿来说，原家庭将会提供一份代答问卷，当异地追踪成功时，如果在异地有与该少儿同住的家长，将会提供另一份家长代答问卷。也即对于少儿来说，有可能存在两份家长代答问卷共存的情况。当多于一份代答问卷同时存在时，我们保留了与少儿物理上同住的家长的代答问卷。少儿库中只有自答、只有代答和自答及代答共存的比例分别是 0.09%，70.02%以及 29.88%。

3. 家庭成员关系库：

家庭成员关系库以家庭成员为单位，CFPS2012 家庭关系库为基础，通过 CFPS2014 家庭成员问卷和个人问卷的最新信息，重新构造 CFPS2012 到 CFPS2014 之间家庭成员构成及其流动状态，补充、变更家庭关系和个人基本信息。CFPS2014 家庭成员关系库包括来自 14219 个家庭的 57934 条观测量。需要注意的是，这 57934 条观测并不代表着 57739 个独立个人：为了体现人员跨家庭的流动性，我们将另组家庭成员分别放在原家庭列表和另组家庭列表中，并用是否同灶吃饭(也即是否与相应家庭有经济联系，其中 co_a14_p=1 表示经济有联系，0 表示经济独立)来表明该个体在 CFPS2014 调查时在经济上属于哪个家户。详细情况

可参见后续的技术报告《CFPS2014 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 55600 条，其中有记录的死亡人数为 559 人。住在原家庭的成员为 41869 人（75.2%），新进基因成员 1617（2.9%），新进的非基因成员 2598（4.7%），经济上归属原家庭但物理外出的成员 4193 人（7.5%），另组成员 3904 人(7.0%)，跨家庭流动的成员 458 人（0.8%）。另外还有少量成员由于存疑、出家、参军、服刑等原因不需要追踪（n=961）。

在 CFPS2014 家庭关系库中，按照个人样本在 CFPS 中第一次出现的时间可分为三种情况：CFPS2010 的基线个人样本 pid 的 fid14、fid12、fid10 都不缺失，表示该个人当期所属的家庭以及在 CFPS2012 及 CFPS2010 时所属家庭；CFPS2012 新进个人样本 pid 的 fid10 是缺失的（-8），fid14、fid12 表示该个人当期所属的家庭及在 CFPS2012 调查时所属家庭；CFPS2014 新进个人样本 pid 的 fid10、fid12 是缺失的（-8），fid14 表示该个人当期所属的家庭。

4. 家庭经济库：

家庭经济库以家庭为单位，包括往期调查所界定出来的原生家庭以及在 2014 年调查时发现由家庭因婚姻变化、子女经济独立等原因所派生出来的新组家庭。在 2014 年家庭经济库的 13946 户中，有 1250 户为当年调查时所界定的另组家庭。访问方式为面访（IWmode=1）或电访(IWmode=2)。

在上一节中我们提到，在 CFPS2014 调查中，对于外出单元与原家庭的经济联系，存在原家庭和外出单元自己的双重界定。当二者的界定不一致时，对家庭经济问卷数据有一定的影响。在家庭经济问卷的开始，访员会根据之前家庭成员问卷的回答情况给受访者列出目前与该家庭有经济联系的每个个人，然后提醒受访者经济问卷的问题应考虑到该列表中的所有家庭成员。如前所述，当双重界定之间不一致时，有可能产生家庭成员在多个关联家庭中重复出现的状况。为了便于用户在后期对这些家庭的数据进行调整，我们在经济问卷中生成了一系列变量来指征与一个家庭有关联的家户号，以及与每一个关联家户之间是否有经济问卷指待家庭成员的重叠，该系列变量如下。

表 3. 家庭经济库中经济问卷关联家庭相关变量信息

变量名	变量标签	值标签
overlapfid1	经济问卷关联家户 1	
overlapfid2	经济问卷关联家户 2	
overlapfid3	经济问卷关联家户 3	
overlapfid4	经济问卷关联家户 4	
overlapfid1type	与关联家户 1 相关类型	1=fid14 与关联家户完全相等 2=fid14 与关联家户交叉；即 fid14 与关联家户有部分共同成员，但是不等 3=fid14 完全包含于关联家户 4='fid14 完全包含关联家户'
overlapfid2type	与关联家户 2 相关类型	
overlapfid3type	与关联家户 3 相关类型	
overlapfid4type	与关联家户 4 相关类型	

5. 村居库：

村居问卷在 CFPS2012 调查时暂停一轮，到 CFPS2014 时再次执行时，CFPS 样本所分布的村居已由基线时的 643 个扩充到 1976 个。由于大部分新增的村居中只有一个或几个 CFPS 样本，CFPS2014 的村居问卷只对基线村居以及 CFPS2014 调查时聚集达到 5 个或以上样本的村居进行村居问卷数据的采集，最终成功获得村居问卷样本数 621 个，其中包括 594 个基线村居。

三． 2014 年问卷数据清理步骤

1. 中断样本的确认

在访问已经开始后，由于各种原因（如受访者中途退出，访问系统问题以及其它原因）

而需要中止访问的样本都属于调查中的中断样本。中断样本中的大部分在后期成功补充完成了并收录在数据库中，但有少量并未完成这部分工作。我们针对这种没有后续完成问卷的中断样本，检查其问卷进展的完整度，如果超过 80%的完成率，则将其纳入发布数据集中。根据该筛选标准，CFPS2014 共纳入 XX 个中断样本观测，其中家庭成员观测 89 个，家庭经济库观测 12 个（10 个面访，2 个电访），成人库观测 20 个（10 个自答面访，8 个自答电访，1 个代答面访，1 个代答电访），少儿库观测 41 个（39 个面访，2 个电访）。在各问卷中中断样本均以 interrupt 变量来指征，中断样本的 interrupt 变量值取 1。

2. 各库样本编码清理

CFPS2014 的各库样本编码清理工作包含以下环节。1）清理各库内部 id 的重复情况，其中家庭经济库为 CFPS2014 家户号 fid14，成人和少儿库为 pid。各库内部 id 的重复主要是由于在执行过程中启用了备用问卷的原因造成的，这些重复样本的确定绝大部分可以通过结果代码来进行判定。2）清理跨库间样本编码的逻辑关系，确保所有数据集的观测以家庭成员库为出发点。3）以家庭关系库为基础，确认和清理所有关联家户中相同名字但不同 pid 的个体。关联家庭中同名不同 pid 出现的主要原因有三个方面：a) 个人是原家庭成员，但是又被其另组家庭认为是第一次新进的成员，b) 个人同时第一次进入多个关联家庭中，c) 成员的名字是谐音，但是从家庭关系判断是同一个人。处理的原则是，往期调查中已经进入 CFPS 的个体在关联家庭中以新人的身份再次进入 CFPS，我们对此类样本将保留其当前所在家庭的观测，并将其 pid 号在各库中统一为个人初次进入 CFPS 调查时的 pid。CFPS2014 中共清理此类样本约 1000 条。4） 以家庭成员库为基础，清理各问卷中所涉及到的家庭成员列表中的样本编码。

3. 部分数值变量的录音核查

CFPS2014 在后期清理中，对各库中部分数值变量的存疑观测进行了录音核查。对原库中确认访员记录错误的观测进行了更新。表 4 中列出了录音核查中确认修正数值的观测条目在 10 条以上的变量信息。

表 4　CFPS2014 存疑观测的部分录音核查变量

变量名	变量标签	所修改的观测数

家庭经济库		
FQ5	房屋购建成本（万元）	193
FM401	全部经营总资产（万元）	80
FT1	您家现金及存款总额（元）	66
FT302	所有住房房贷总支出（元）	61
FT101	您家定期存款总额（元）	53
FT401	购房建房装修借款额度（元）	28
FQ6	房子当前市价（万）	18
FP505	物业费（元）	16
FR2	其他房产市价（万元）	16
FN301	领取离退休或养老金总额（元）	15
FP507	汽车购置费（元）还有保养、维修	12
FINC	过去 12 个月总收入（元）	10
FP506	住房维修费（元/年）	10
FO7	工资收入总额（元）	10
成人库		
qq402	午休时长（分）	375
qq403ab	晚上睡觉时间（点）	59
qp702	一周锻炼时长（小时）	41
qq9010	干家务时长（小时）	40
qi202	税后领退休金数额（元）	31
qm9	嫁妆价值（元）	17
qp510a	其他伤病费用（元）	17
qq1001	每周看电视、电影时长（小时）	14
qq2011	开始吸烟年龄（岁）	13
qp512a	自家直接支付（元）	11
qc303_a	非学历教育时长（小时）	10
少儿库		
KS1011	非周末学习时间（小时）	11

通过录音核查所最终修正的变量值只包括那些核查员能较为肯定的判定是由于访问员在记录时有误的情况，但由于受访者或访员理解原因，采集的数据并不符合常识，我们无法通过录音核查进行有效地改动。同样，部分存疑变量及观测由于录音缺失、录音质量较差等问题不能核查，用户可根据具体的研究需要进行相应地处理。

4. 面访与电访数据的合并

CFPS2014 继续沿用 CFPS2012 的面访为主，电访为辅的原则。电访比例总体上相比 2012 年有一定程度的上升。 CFPS2014 中家庭经济、成人自答、成人代答、少儿代答问卷中电访所占比例分别为 3.88%、5.55%、19.65%和 6.32%。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 5 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，研究者可根据具体的研究需要对其进行再处理。 注意此表中不包括那些只在面访中出现但并未在电访中出现的变量。

表 5　2014 年家庭问卷电访和面访不完全匹配问题信息列表

电访有效变量	面访有效变量	不匹配类型
家庭经济库		
FN6	FN3	电访和面访针对过去 12 个月中领取离/退休金或养老金情况分别采取询问人数和是否有人领取两种方式提问
F08	F01 F02	电访和面访针对作农活或外出打工人员数量分别采用总、分不同的提问方式。
F09_1	F05_1	电访和面访针对除了帮别人家做农活和外出打工外，过去 12 个月的工资收入情况分别采取询问人数和是否有人领取两种方式提问
F09_2	F05_2 F06	电访和面访针对在过去 12 个月中家里几人有工资收入分别采用总、分不同的提问方式。
FL5	FL501 FL502 FL503 FL504 FL505	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。
FL8	FL801 FL802 FL803 FL804 FL805	电访和面访针对饲养家禽、家畜及水产品投入金额分别采用总、分不同的提问方式。
FP408	FP401 FP405	面访针对邮电通讯和交通支出金额采取分别提问的方式。
FP520	FP402 FP403 FP404 FP504 FP505 FP506	电访和面访针对水电费、燃料费、取暖费、物业费（包括车位费）、住房装修维修费分别采用总、分的提问方式。

FP519	FP502 FP503	电访和面访针对文娱旅游支出金额分别采用总、分的提问方式。
FP521	FP507 FP508	电访和面访针对购买、维修、保养汽车及其他各种交通工具（如自行车、电动自行车）、通讯工具（如手机）及配件的费用分别采用总、分的提问方式。
FT1001	FT601 FT602	电访和面访针对亲友/民间借款待偿总额(元)分别采用总、分的提问方式。
成人和少儿库		
DSA1TOTAL_M	KS10 KS11	面访将亲朋好友支付与学校政府支付教育费用区分，电访将两者合并不进行区分。
DR413B_T	KR413B	面访关注参加哪一种辅导班的每周平均时间，电访仅关注辅导班每周平均的时间
DF104M	WF104M	电访和面访选项不一致
DD6TOTAL_M	WD6 WD7	电访和面访针对家庭外教育费用分别采用总、分的提问方式。

5. 成人问卷中自答和代答数据的整合

CFPS2014 在成人自答问卷的基础上，另外设计了代答问卷。代答问卷在结构上与自答问卷类似，但去除了不适合代答的主观题和认知题等。

对于成人来说，代答问卷主要有两种用途：1)原家庭的基因成员和核心成员物理外出时，由在家的其它家庭成员先帮其完成代答问卷，在异地追访到个人时，再由其自身完成个人自答问卷。2）本应自答的人员由于身体原因而无法完成自答时，由对其情况熟悉的家庭成员完成代答问卷。由代答问卷的适用范围可知，在 CFPS2014 的个人库中将包括三种自答和代答问卷类型结合的情况，一种是只有自答问卷的观测，这在个人库中占最大的比例；第二种是只有代答问卷，这包括由于身体原因无法自答的个人以及异地追访不成功的个人；第三种是既有自答问卷又有代答问卷的，这些人是外出的基因成员和核心成员并且在异地追访成功的。成人库中只有自答或代答问卷的比例分别是 85.69%、9.88%，剩下 4.42%的人既有自答问卷，又有代答问卷。

代答问卷中的大部分问题是直接从自答问卷中摘选的，对于这些题需要统一其变量名，并在合并时考虑当长短问卷均有时，用自答问卷数据覆盖代答问卷数据。需要额外注意的问

题有两点。一是当自答问卷答案为缺失，而代答问卷存在有效值时，用代答问卷的数值作为最终值；另外，部分问题存在自答代答问卷不完全匹配的情况，需要用不同的变量分别记录。表 6 为成人问卷中自答和代答问卷不完全匹配问题列表。

表 6　2014 年成人自答和代答问卷不完全匹配问题信息列表

自答问卷变量	代答问卷变量	不匹配类型
GA1	PG01	成人自答问卷关注 2012 调查以来是否有过全职工作，代答问卷关注现在是否有工作
QGA101	PG02	成人自答问卷关注过去 12 个月是否有实习兼职经历，代答问卷关注过去 12 个月是否有工作
GC104	PG501	成人自答问卷关注工作结束时间，代答问卷针对的是工作月数
EGC103	PGC103A	成人自答问卷加载的是 2012 年更新后的工作确认，代答问卷针对的仅是 2012 年的工作
QN401S_S_1-QN401S_S_14	PN401a	成人自答问卷针对多种组织成员类型进行多项选择，代答问卷是单选题形式，只关注是否是中国共产党党员。
KW1	PW1R	成人自答问卷关注的是在哪个阶段离开正规教育额，代答问卷关注的是最高学历
QZ201	PZ201	成人自答问卷是对受访者直接提问，代答问卷是对代答者提问
QZ207	PZ207	成人自答问卷是对受访者直接提问，代答问卷是对代答者提问
QZ209	PZ209	成人自答问卷是对受访者直接提问，代答问卷是对代答者提问
QZ211	PZ211	成人自答问卷是对受访者直接提问，代答问卷是对代答者提问
QZ5	PZ5	成人自答问卷是对受访者直接提问，代答问卷是对代答者提问

6. 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，包括家庭工资性收入、经营性收入、转移性收入、财产性收入和其他收入。其中工资性收入是指家庭成员从事农业或非家受雇工作争取的税后工资、奖金和实物形式的福利。经营性收入是指家庭从事农林牧副渔业生产经营扣除成本后的净收入（包括自产自销部分），以及从事个体经营和开办私营企业获得的净利润。转移性收入是指家庭通过政府的转移支付（如养老金、补助、救济）和社会捐助获取的收入。财产性收入是指家庭通过投资、出租土地、房屋、生产资料等获得的收入。其他收入是指通过亲友的经济支持和赠予获取的收入。CFPS2014 在经营性收入、转移性收入、财产性收入和其他收入方面的设计与 CFPS2012 相同，但在工资性收入上，CFPS2012 在个人问卷中对逐个家庭成员询问他们自己的工资收入，然后通过加总的方式得到家庭总的工资收入。但这种由个人问卷中逐个成员加总方式得到总工资收入的缺点是一旦部分家庭成员的个人问卷缺失，该家庭的总工资性收入无法直接得到，在 CFPS2012 中我们通过个体特征对缺失个人问卷的家庭成员工资进行了插补。与 CFPS2012 不同，CFPS2014 的家庭问卷包含了家庭工资性收入部分，由家庭问卷回答人汇报家庭总体的工资收入。但对于家庭内有物理外出的个人可能存在着低估，因此在家庭工资收入生成时，如果遇到存疑情况（包括缺失值、0 值或农村家户个人问卷加总工资收入大于家庭问卷报告的收入时），则用个人问卷加总的工资性收入来替代家庭问卷数值。关于 CFPS2014 家庭收入计算的更多信息，请参考《中国民生发展报告 2016》的“收入分配”章节。

为了方便用户与 CFPS2010 基线数据的比较，我们还同时生成一版与基线可比的系列变量。

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四大类支出，它们分别是居民消费性支出（包含食品、衣着、居住、家庭设备及日用品、交通通讯、文教娱乐、医疗保健、其他消费性支出），转移性支出（包括家庭对非同住亲友的经济支持、社会捐助以及重大事件中人情礼），保障

性支出（包括家庭购买各类商业保险），建房购房贷款支出。有关 CFPS2014 家庭支出更详细的信息，请参考《中国民生发展报告 2016》中的“家庭消费”章节。

以上两类变量在家庭经济库中的具体变量列表如下。

表 7　家庭经济库中综合变量（已更新）

变量名	变量标签
家庭收入部分	
fWAGE_1	工资性收入（调整后）
fWAGE_2	工资性收入（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	全部家庭纯收入
FINCOME2	家庭纯收入(与 2010 可比)
fincome1_per	人均家庭纯收入
fincome2_per	人均家庭纯收入(与 2010 可比)
fincome1_per_p	人均家庭纯收入分位数
fincome1_per_p	人均家庭纯收入分位数（与 2010 可比）
家庭支出部分	
PCE	居民消费性支出-加总
FOOD	食品支出-调整
DRESS	衣着鞋帽支出
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	房贷支出
EXPENSE	家庭总支出
其它	
urban14	国家统计局划分的城乡类型

3) **个人收入（成人和少儿库）**

个人收入（p_income）计算中包括了实习收入、正式受雇工资、养老金、奖学金，将实习收入和正式受雇工资合并成为个人工资性收入(p_wage)，并将电访、面访、代答的收入进行调整，形成了个人累加工资性收入（f_wage）。

4) **认知水平（成人和少儿库）**

CFPS2014 的认知测试从设计上来说基本沿袭了 CFPS2010 的问卷，包括识字测试和数学测试两部分。但在 CFPS2010 的基础上，也做出了调整，其中最重要的一点是 CFPS2010 中依据教育水平将受访者分成由低到高三阶段，每个阶段认知测试的起点不同，且较高起点的受访者不会退到较低起点测试。这种方法的缺点是如果起点设置过高，则无法较好估计受访者的真实水平。因此在 CFPS2014 中，我们允许在较高起点的首道题答错后降到低一级起点再尝试，以更准确地估计受访者的认知水平，由变量 wordtest14_sc2 和 mathtest14_sc2 表示。同时，我们另外生成了一系列假设起点固定时，受访者有可能得到的分数，以确保其与 CFPS2010 认知分数的可比性，由变量 wordtest14 和 mathtest14 表示。

以上两个系列综合变量在个人库中的列表如下。

表 6　成人与少儿库中综合变量

变量名	变量标签
个人收入	
p_income	个人收入
p_wage	工资性收入
认知水平	
wordtest14_sc2	14 年词组测试题得分：14 年问卷算法
wordtest14	14 年词组测试题得分：与 10 年可比算法
mathtest14_sc2	14 年数学测试题得分：14 年问卷算法
mathtest14	14 年数学测试题得分：与 10 年可比算法

7. 各类编码工作

1） 职业编码和行业编码

CFPS2014 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。职业和行业编码体系请参见 CFPS2010 技术报告《中国家庭追踪调查 2010 年职业行业编码》。

2） 疾病和死亡编码

在 CFPS2014 调查中，发现在与上次调查间，共有 559 人由家庭成员汇报为去世状态，对于这些个人，我们询问了其死亡原因，然后由访员在现场对死亡原因进行编码。

CFPS2014 年成人和少儿问卷中，分别在健康模块询问了以下关于疾病的信息：成人库采集“您被医生诊断的最主要的慢性疾病名称”，少儿库是“孩子患过最严重的疾病秉承”。后期由编码员按照 CFPS 疾病体系进行编码。按照患病部位和严重程度，CFPS 疾病编码体系将疾病划分为 21 大类，131 小类，详见《CFPS 疾病编码表》。

3） 地址编码和城乡状态

与 CFPS2010 和 CFPS2012 相同，CFPS2014 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在 2010 年的一系列宏观变量，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》。与往年有所不同，CFPS2014 在给出了地址编码以外，还另生成一个指示变量来代表该区域的行政区划是否发生了变化（第二版发布数据更新中将包括）。

我们同时提供了按 2014 年国家统计局网站上定义的各样本所在村居的城乡性质。

4） 方言编码

CFPS2014 记录了访问使用的语言，我们以《中国语言地图集》[2]为依据，结合受访者所在区县、出生地等其他辅助信息，对该文本信息进行编码。方言编码体系的介绍见 CFPS2012 技术报告《中国家庭追踪调查方言编码》。

5） 行政管理职务编码

在工作模块，CFPS2014 以开放式问题询问了受访者的行政管理职务。为便于用户使用，结合采集到的单位性质规模、下属数量等信息，按照表 7 所示体系对采集的信息进行行政管理职务编码。该编码体系的介绍，详见 CFPS2010 技术报告《2010 行政/管理职务综合变量

[2] 中国语言地图集[M]. 朗文出版（远东）有限公司, 1987.

的建构》。

表 7 行政管理职务编码表

代码	标签
0	无职务
1	公共部门基层行政/管理职务
2	市场部门基层行政/管理职务
3	公共部门中层行政/管理职务
4	市场部门中层行政/管理职务
5	公共部门高层行政/管理职务
6	市场部门高层行政/管理职务
7	公共部门顶层行政/管理职务
8	市场部门顶层行政/管理职务

6） *在读本科院校编码*

CFPS2014 的上学模块采集了在读大学生样本的在读本科院校信息共计 358 条。为保护用户隐私，对该文本信息进行了编码处理。编码的分类原则参照入学年份和教育部公布的高等教育学校信息。具体而言，本科院校主要包括以下几类：全国重点院校（985 高校，第一批次录取）；全国重点院校（非 985 的 211 院校，第一批次录取）；普通重点院校（第一批次录取）；普通本科院校（第二批次录取）；三本院校（第三批次录取）；部队院校（提前批录取）；艺术、体育类院校；海外大学。

以上六类编码变量在个人库中的列表如下。

表 5 个人库编码变量列表

变量名	**数据库**	**变量标签**
职业、行业编码		
QG302	成人库	QG302 行业编码
QG303	成人库	QG303 职业编码
QGA4	成人库	QGA4 行业编码
QGA401	成人库	QGA401 职业编码
WGA4	少儿库	WGA4 行业编码
WGA401	少儿库	WGA401 职业编码
疾病编码		
QP402A	成人库	慢性疾病编码 1
QP402B	成人库	慢性疾病编码 2
WC5	少儿库	最严重疾病编码
WC5_2010	少儿库	出生后最严重疾病编码
Deathreason14_p	家庭关系库	个人去世的原因

方言编码		
QZ104	成人库	访问使用什么方言
行政管理职务编码		
QG1401	成人库	行政管理职务描述
在读本科院校编码		
KRA603	成人库	目前在哪所学校读本科

四． 2014 年清理难点

1. 家庭关系库的构建

CFPS2014 家庭关系库以 CFPS2012 完访家庭关系库、2012 年未完访家庭的部分 CFPS2010 基线家庭关系库为基础，根据 CFPS2014 家庭成员问卷、个人确认、成人、少儿多套问卷采集的信息，经过了信息的分解、重组、核查、修正等多个环节、多轮循环操作生成的。它的内容一方面反映了 CFPS2012 至 CFPS2014 两轮调查间每个成员在家庭中的状态、家庭成员构成、家庭网络关系、个人基本信息的变化；另一方面，也在确认 CFPS2016 调查时发放的家庭和个人追访样本的构成和种类，是下次调查的依据。

CFPS2014 较 CFPS2012 家庭关系库的结构相同，除了要克服 CFPS2012 经历过的难度外，CFPS2014 家庭关系库在生成过程中增加以下新的难点：

1．家庭成员问卷有新增加模块、修改模块、逻辑校验，原始数据库的变量总体增加到近 2.6 万，不仅给原数据库的分解造成难度，同时判断变量之间的关联性、逻辑性的难度也变大。

2．原家庭成员确认模块中采集的存疑人员与实际情况可能不一致，需要特殊处理。

3．考虑原家庭和另组家庭的成员之间存在跨家庭流动，因此家庭成员问卷中新增关联家庭之间成员流动的内容。在数据清理过程中，需要额外整理这类流动人员的个人信息。

4．不同于 CFPS2012 问卷生成规则的设计，CFPS2014 的所有外出单元都会有家庭成员问卷，如果不被定义成家庭，这些人需要合并到原家庭，会涉及个人所属家庭编码的调整，但是新进非基因成员不放回，因此数据上要做相应的调整。

5． CFPS2014 新家庭样本的界定不同于 CFPS2012，见（一、1）中介绍的家庭成员经济联系的双重界定的设计，从家庭成员构成、家庭网络关系的补充、个人信息的匹配上都增加了难度。

6． CFPS2014 产生的外出单元家庭成员问卷量增加，随之同一个人在多个家庭出现的

样本也增加，出现了近 1 千条观测 pid 不一致的问题，需要人工判断该成员应该在哪个家庭中，以及保留哪个 pid。

7. CFPS2014 是第三次调查，清理、核查中需要综合考虑前两次的信息，以保证多次调查间一致。

2. EHC 中冗余和中断的清理

CFPS2014 在成人自答问卷及成人自答电访问卷的迁移、婚姻和工作三个模块采用了历史事件日历记录法（EHC）设计，以便更加详尽的采集受访者在两次调查之间的时段内的状态变化。成人自答问卷及成人自答电访问卷使用相同的 EHC 设计。

由于技术实现上的难点，CFPS2014 的 EHC 模块的问卷系统不能清除冗余值，所以会记录下受访者回答的所有痕迹。如果受访者在回答 EHC 过程中意识到某个题回答错误，回去修改那道题目的答案，然后逻辑跳转至另一条路径，并继续答题。这时候，新路径上的应答与旧路径上本不应该答的题目答案都被系统保存下来了。此外，由于实地调查过程中的各种原因导致在回答 EHC 模块过程中的中断，这时候中断点之后的数据也会相应缺失。

EHC 清理工作主要是针对冗余和中断数据的清理。首先识别不需要储存答案的旧路径，将冗余值改为了-8；同时根据问卷逻辑跳转判断应该回答但却没有回答的题目，将原来的-8 改为了-9，标识此题答案缺失；如果恰好是逻辑跳转点题目的答案缺失，则由此题引出的几条逻辑跳转路径中的题目都被改为了-10，标识无法判断是否应该回答这些题目。例如，在成人自答面访数据集中，EEB5 变量有 97 个观测从-8、1、5 修改为了-10，有 14 个观测从-8 修改为了-9，有 111 个观测从 1、 5 修改为了-8。

除了各个题目对应的变量之外，EHC 模块内部还会生成一系列综合变量。在对 EHC 模块各个题目的值进行清理之后，再按照 EHC 模块中的规则，用清理后的数据生成新的 EHC 模块的综合变量（如婚姻模块的 cmstart 以及工作模块的 IncomeA, jobstartN, firstjob 等）。