
China Family Panel Studies



中国家庭追踪调查

技术报告系列：CFPS-42

系列编辑：谢宇 责任编辑：赵逸文

中国家庭追踪调查 2020 年婚姻模块数据清理报告

张厚义¹ 左舒艺² 孙妍²

1. 安徽师范大学 2. 北京大学中国社会科学调查中心

2022. 09

目录

目录	1
一. 清理背景	2
二. 问题描述	2
三. 2020 年婚姻数据库重构	3
3.1 重构数据结构	3
3.2 数据清理	4
四. 补全往期婚姻数据	4
五. 合并构建用户友好婚姻数据库	5
5.1 构建婚姻状态最佳变量	5
5.2 不同测量方式的矛盾数据处理	5
5.3 构建用户友好婚姻数据库	6

一. 清理背景

婚姻是个人生命历程中的重要事件。婚姻的历程与变动,关联着父母身份的转变和生育子女的状态,对个人的工作、健康和生活模式有着长期性和持续性的影响。因此,婚姻信息是实证研究中普遍关注的重要指标。

CFPS 中的婚姻信息有三种来源:个人自答问卷、家庭成员问卷和个人代答短卷。针对 16 岁及以上的受访对象,我们每一轮会在其个人自答问卷中通过婚姻模块采集两轮调查间经历的婚姻史。CFPS 问卷中将婚姻状态分为五类:未婚、在婚/有配偶、同居、离婚和丧偶。婚姻模块主要采集回溯期内每次婚姻状态变化的经历,采集每段婚姻或同居经历开始或结束的时间、结束的原因,如果存在配偶/同伴也会采集配偶/同伴的个人信息。针对因各种原因未能完成个人自答的受访者,则会通过家庭成员问卷或个人代答短卷采集其简要的婚姻状态信息。

由于追踪调查项目后续追踪侧重于了解受访者两期调查间的变化,数据采集策略是以受访者在往期调查中提供的信息为出发点,采集并更新变化的信息。对于婚姻状态没有改变的受访者,其往期调查已采集的信息通常不会在当期的发布数据库中体现。结合数据用户的使用反馈,我们了解到大量用户希望在已公布的 CFPS 个人数据库中更直观、全面地展现出受访者完整的婚姻信息。为方便用户使用,项目组在对 2020 年 CFPS 个人库及家庭关系数据进行综合清理的基础上,尝试构建出用户友好的婚姻数据库,希望降低广大数据用户使用该部分数据的难度。以下是用户友好婚姻数据库的清理过程与方案说明。

二. 问题描述

结合用户的反馈意见,我们认为在当前婚姻数据库中主要存在以下两方面的问题可以继续优化:

第一,数据结构不够简洁,存在同类变量。CFPS 婚姻信息的采集方式与婚姻状态相关联,五种婚姻状态有各自的问卷分支,在数据库中分散于不同位置存储;但不同婚姻类型的问卷分支中又采集了相同的信息,如开始时间。目前发布版的数据基本保留了实地数据采集的结构框架,虽然能较好地满足用户数据与问卷相匹配的需求,但在一定程度上也复杂化了婚姻数据结构。考虑到一位受访者在某一时点上只存在一种婚姻状态,研究者往往更希望用一个变量标识受访者现有的婚姻状态,并提取与该段经历相关的各类信息。用户友好数据库

将对所有婚姻状态类型的问题进行整理，整合成用婚姻状态作为标识变量的简化数据版本，以使用户能够快速识别和提取。例如：将婚姻及同居部分都出现的开始年份变量替代为开始年份单一变量。

第二，无法直接获取当前婚姻信息。CFPS 访问是以受访者在往期调查中回答的婚姻状态为出发点，请受访者回忆两轮调查间经历的所有婚姻状态变化。若受访者的婚姻或同居经历（无论是否属于当前婚姻状态）起始于本期调查之前，且往期调查时已采集过婚姻及配偶的相关数据，而本轮调查回溯期内婚姻状态并未发生变化，当期将不再重复提问，往期采集的数据也不会直接在发布库中体现。这种数据采集方式虽然提高了婚姻数据的完整性和精细度，但也意味着除了婚姻状态，用户无法从当期数据中完整地获取所有受访者当前婚姻的详细信息。

考虑到研究者在使用婚姻数据时，不仅需要使用婚姻状态数据，还需要诸如配偶的个人属性、婚姻开始时间等详细信息。为方便用户使用，我们在清理构建简化版 2020 年婚姻数据库的基础上，尝试纳入往期调查（侧重 2018 年）所获取的婚姻信息，构建完整的婚姻数据库。

三. 2020 年婚姻数据库重构

从以下两方面对 CFPS2020 年个人库中的婚姻模块数据进行基础清理，主要包含数据结构简化及清理两部分工作。

3.1 重构数据结构

首先，保留 2020 年个人库中婚姻模块采集的每一段婚姻或同居的开始时间、结束时间、是否持续到现在、结束原因¹、与配偶/同居对象的相识方式、配偶/同居对象的出生年月、配偶的受教育程度、是否存在婚前同居、婚前同居月数等信息。

其次，将“婚姻开始年月”与“同居开始年月”合并为“婚姻/同居开始年月”，将“婚姻是否持续到现在”与“同居是否持续到现在”合并为“婚姻/同居是否持续到现在”，利用婚姻状态识别该段经历的类型。

最后，将“宽数据”转变为“长数据”，使每位受访者的每段婚姻或同居经历成为一条

¹ CFPS2020 问卷中，同居部分未问及结束原因。

单独的观测，再按照个人 pid 及婚姻/同居的开始时间重新排序和编号，便于研究者根据自己的研究需求，识别特定的经历或存在特定经历的受访对象。

3.2 数据清理

通过校验重构后的婚姻数据，我们发现 22 人汇报了不止 1 段持续到现在的婚姻或同居，共涉及 45 条数据，主要分为以下两种情况：

第一，有 17 人汇报了开始年月一致的 1 段婚姻和 1 段同居（共计 34 条数据），其中一个解释是：受访者错误理解了同居的含义，在婚姻及同居模块汇报了同一段婚姻经历，导致重复采集。结合受访者汇报的当前婚姻状态，采用当前婚姻状态为在婚，但仍保留同居记录。

第二，受访者汇报了开始年月不一致但时间存在部分重合的婚姻和同居，考虑到 CFPS 数据采集过程中允许某一时段内同时存在婚姻与同居。因此，对该部分数据不做处理。

四. 补全往期婚姻数据

CFPS 个人婚姻模块数据采集以最近一次调查中受访者的回答为出发点进行提问，若婚姻状态没有变化则不再询问其配偶/同居伙伴的相关信息。因此，CFPS2020 年个人库中 2018 年婚姻数据存在缺失。在重构 2020 年婚姻数据库后，通过 2018 年个人库和家庭关系库加载补充 2020 个人库中缺失的 2018 年婚姻数据。根据访问的不同类型，数据补全的方案有所不同，具体操作流程如下：

首先，针对 2018 年有配偶或同居对象且在 2020 年访问时确认加载婚姻状态的人员²，从 2018 年家庭关系数据库中获取婚姻或同居开始时间，配偶/同居对象的 pid、出生年月、受教育程度等基本信息，合并至重构数据库中，补全 2018 年婚姻的相关信息。

其次，针对 2020 年个人库中的婚姻更正情况，对从 2018 年家庭库中获取的数据进行如下处理：（1）若受访者接受过访问，存在加载婚姻状态数据，但 2020 年在上期婚姻经历确认部分将当时汇报的婚姻状态调整为未婚/离异/丧偶，则将往期婚姻/同居相关的信息标识为不适用；（2）若 2020 年受访者首次接受访问或将往期婚姻状态调整为在婚/同居，且在婚时间早于 2020 年追踪的起始时间（2018 年 1 月 1 日或 2018 年接受访问的日期），则根据 2020

² 2020 年个人数据库中的 2018 年数据仅包括：对 2018 年调查数据的更正，以及初次访问对象对 2018 年情况的追忆。因此，2018 年调查无误的样本在此部分标记为不适用，需要向前逐期合并以填补数据。

年采集的信息补充 2018 年的婚姻信息³。

五. 合并构建用户友好婚姻数据库

为了满足用户使用需求,我们重构数据结构并合并了前期数据,将 2018、2020 年婚姻信息及两次调查间婚姻变动信息组合起来,形成一个用户友好的婚姻数据库。本部分通过调整后的 2018 年婚姻信息和两次调查间的婚姻史信息,生成 2020 年用户友好婚姻数据库。

5.1 构建婚姻状态最佳变量

结合前两部分的数据清理,按照以下步骤生成 2020 年最新婚姻信息。

第一,对于两次调查间未发生婚姻变动的受访者,将其在 2018 年时的婚姻数据顺延到 2020 年,作为当前最新的婚姻状态。

第二,对于两次调查间发生了婚姻变动,但未持续至今的受访者,依据最近的一份变更记录,调整婚姻信息。例如:A 在 2018 年时婚姻状态为未婚,但 A 在两次调查之间有一段未持续至今的婚姻,则根据该段婚姻的结束原因,将 A 在 2020 年当前最新的婚姻状态变更为离异或丧偶,将其配偶的出生年月等个人信息变更为不适用。

第三,对于两次调查间发生了婚姻变动,且其中有一段婚姻经历持续至今的受访者,将其 2020 年婚姻信息调整为持续至今的那段婚姻变动信息。例如:A 在 2018 年时婚姻状态为未婚,但在两次调查间结婚并持续至今,则将 A 在 2020 年当前最新的婚姻状态变更为在婚,将其配偶的出生年月等信息变更为该段婚姻的数据。

第四,对于两次调查间发生了婚姻变动,且其中有不只一段婚姻经历持续至今的受访者(具体参见 3.2 部分所述)。一律以最近的一段婚姻经历为准进行清理,将 2020 年的婚姻状态变更为在婚,将其配偶的出生年月等信息变更为那段持续至今的婚姻数据。

5.2 不同测量方式的矛盾数据处理

在数据处理过程中,我们发现由不同测量方式采集数据时可能存在矛盾之处。具体而言,项目组结合受访者在 2018 年的婚姻信息和两次调查间的婚姻变动信息,构建的受访者当前婚姻状态与问卷中直接询问获得的婚姻状态有少量样本存在不一致(共 110 条,占比 0.48%),

³ 编写代码及此文档时,2020 年家庭关系库尚未公布。因此,暂时通过 2020 年个人库更新信息,未处理配偶 pid。

主要有以下两种情况：

第一，受访者回答的婚姻变动为持续到现在的同居/婚姻经历，但受访者在 2020 年调查时回答的当前婚姻状态却为未婚/丧偶/离婚，此类数据共 103 条。

第二，受访者回答的 2018 年婚姻状态为未婚、离异、丧偶，且未有持续到现在的婚姻或同居经历，但受访者在 2020 年调查时回答的当前婚姻状态却为在婚或同居，此类数据共 7 条。

为了处理如上矛盾，我们在保留原有 2020 年婚姻状态变量的基础上，以婚姻变动经历数据为基础，重构当前婚姻状态的最佳变量，以获得逻辑上最顺畅的数据。需要申明的是，最佳变量只是做出了有限的更正，研究者可以根据研究性质自行决定是否使用最佳变量。

5.3 构建用户友好婚姻数据库

完成以上数据清理后，我们将生成的 2020 年及 2018 年婚姻数据纵向合并，形成用户友好婚姻数据库。数据库中包含每位受访者在 2020 年、2018 年及两年间每次变动的婚姻状态、婚姻状态最佳变量、开始年月、与配偶/同居者相识方式、配偶/同居者的 pid、配偶/同居者的出生年月、配偶/同居者的最高受教育程度、婚姻是否持续到当前、婚姻结束时间等基本信息，方便用户直接使用。

为便于用户使用，我们同步发布“CFPS2020 婚姻模块清理程序.do”（点击[此处](#)），用户可根据自己的研究需要对数据清理的流程或数据结构进行调整。