计算机辅助面访跟踪调查中的数据管理

——以中国家庭动态跟踪调查（CFPS）为例

□ 文/任莉颖　严 洁

“中国家庭动态跟踪调查（CFPS）”是中国第一个计算机面访模式下的大型全国性跟踪调查。该项目由北京大学中国社会科学调查中心设计实施，从2010年起在全国25个省市（西藏、青海、新疆、宁夏、内蒙古、海南、香港、澳门、台湾不在其列）内的16000个家户进行跟踪访问，样本涉及个人、家庭和社区三个层次，主题包括社会、经济、人口和健康等多个领域。本文将以此项目为例，对该种调查环境下的数据管理工作进行介绍和讨论，以期得到业内人士的借鉴和交流，共同促进本土社会抽样调查的发展。

一、CAPI跟踪调查数据管理工作的特性

传统纸笔模式下的跨地区调查中，数据工作主要是调查后期的数据录入、清理，以及数据库的建立和规范。CAPI模式下的跟踪调查，访员在访问的同时将受访人答案录入电脑，采访系统在后台直接存储为数据库的形式，这样就省去了乏味的数据录入工作。此外，如果问卷设计系统化的过程中充分考虑变量取值以及变量间逻辑关系的控制，事后的数据清理工作也会负担减小，使得数据修订时间大幅度缩减，调查数据很快就可以交付使用。然而，在CAPI调查中，数据管理工作有许多突破常规的创新之处，不仅工作重点前移到调查前和调查过程中，而且工作内容变得更加繁杂，对于统计软件应用技术的要求也更高。

首先，CAPI调查中的数据管理涉及对多类型数据的同时管理。CAPI调查采集的数据内容丰富，有关于调查主题的问卷数据，有关于调查访问过程的并行数据，还有关于质量控制的核查数据等。同时这些数据也形式各异，可以有数值的，有文本的，有音频的，或有视频的。有些数据可以直接用统计软件分析，有些数据则需要进行提取、整理和加工。

其次，CAPI调查中的数据管理采用实时的管理模式。CAPI调查的数据可以及时利用通信网络传送给调查总部。在通信条件好的地方，访员可以实现在线采访，这样数据就直接存储在服务器上；在条件不太好的地方，为了防止由于网速过慢而造成的采访中断，多采用离线采访，访后将数据压缩打包发送到服务器上。这样数据管理人员就可以根据每天收到的数据，为调查执行提供必要的访问进度信息，同时也可以进行及时的数据监控和清理。

此外，CAPI调查中的数据管理还具有明显的系统性。CAPI调查采集的各类数据间具有严密的逻辑性和关联性。数据管理过程中要对这些逻辑进行系统核查，同时利用关联性为采访系统、质控系统提供必要的技术支持。

这些工作对于提高调查管理效率，保证问卷数据质量具有重要贡献。

二、CAPI跟踪调查数据管理工作的内容

CAPI跟踪调查中，调查问卷数据库的管理仍然是数据管理工作的主要任务，但在管理内容和方式上有许多创新之处。同时，数据管理工作还包括了对调查执行和质量控制等工作的数据支持。基于CFPS在过去两年的调查经验，本文将对此详细介绍。

1．调查问卷数据库的管理。对于调查问卷数据的管理可以分为多个阶段。首先，在将调查问卷电子化过程中，数据管理人员就要开始数据库结构的设计，如变量名称、变量标签，以及变量值标签的设计等。基于测试数据，还要根据调查问卷结构清点访问系统设计的变量个数，核查变量间逻辑关系等。这样，在调查问卷完善的同时，问卷数据库已完成了基础的建构，同时

也可以有效防止由于问卷设计失误而造成的数据采集问题。

下一阶段的数据修订工作包括数据清理和开放性问题编码。这两项工作得益于CAPI化调查数据传输的及时性，均可以在调查实地执行过程中就开始。这样有助于及时发现问题，并且采取有效的补救措施。

在CFPS中，数据修订工作涉及以下几方面内容：

（1）问卷数据条目的清点。访员在结束每一份问卷的采访时，会按照指令，在访问管理系统中插入一条结果代码，指示该问卷完成。然后，在条件允许的情况下，尽快上传数据。这个过程有可能会出现三个错误：一是访员未能插入指示问卷完成状态的结果代码；二是访员忘记发送数据；三是数据发送过程中由于网络问题没能接收成功。因为合并问卷数据时要根据结果代码的情况，所以这三种原因都会造成问卷数据的丢失。此外，访员在访问过程中可能会出现访错样本的情形，这样的问卷数据需要从数据库中删除。在CAPI模式下，由于数据的关联性，这些问题可以及时发现。比如通过查看并行数据中记录的完成状态的采访样本与问卷数据库中的样本是否一一对应，可以发现访员是否已发送数据；通过查看问卷数据合并记录，发现是否出现数据接收错误；通过查看问卷间的逻辑关系，来发现是否出现结果代码没有成功插入的情形。对于访错样本的情形，则要依赖访员、督导，或质控人员的报告，通过修改结果代码，或者后期编程来删除无效的问卷数据。

（2）问卷数据点的清理。问卷数据点的清理主要针对采访过程中的三个问题：一个是受访人回答的逻辑关系错误，二是访员采访过程中的录入错误，三是访员信息记录不完全。问题跳问的逻辑关系清理在问卷设计的系统测试阶段就已检查过，虽然仍可能存在问题，但已经不是数据清理的重点。CFPS调查有一个记录家庭成员之间关系的数据库，数据清理时发现有部分数据逻辑混乱，无法据此理出家谱。对此部分数据的清理需要参考多方信息。首先要了解该问卷采访的系统设计，判断哪些环节容易造成逻辑关系应答错误，访员的录入错误，或由于系统改版造成的数据存储问题。其次要查看相关并行数据，判断访员是否曾经修改过数据，以至于出现数据值错位。同时要查看有关问卷数据，通过对姓名、年龄，以及性别的匹配来识别家庭关系。

访员录入错误的检测主要通过两个途径：一个是访员事后发现，主动报告；另一个是定期分析数据分布，查找异常值。对待访员报告的录入错误，数据管理人员要及时确认并更正，避免等到后期处理时因访员报告不清而难以查对。异常值有两种，一种是野码（wild code），即数据存储值不在题目的选项值内，或者存储值不合常理。如问去年有几个月外出，存储的数据值超出了12，这些数据值即为野码。另一种是奇异值（outlier），指取值远远大于或小于该变量的总体或组内平均值，在数据分布上处于孤立状态的一些数据点。有时，对于一些额度较大的数值型变量问题，如房产市值，访员会看错单位而出现录入错误，导致奇异值。在CAPI模式的调查中，这两种异常值可以通过建立即时的数据监控来侦测。野码可以直接判断为数据值出错，而后者则需要其他信息的辅助来判断是否为访员录入有误。一旦确定或怀疑是录入错误，如果可能的话，立即与访员联系来补救数据，同时也可以提醒访员，在后面的采访中避免类似的错误。

（3）开放性问题的即时编码。开放性问题就是研究者在问题设计时没有列举出所有可能的答案，而是由受访人自由回答，访员以文字形式记录下受访人的话语。这些文本型数据对于量化分析研究者是无法直接使用的，需要通过建立分类标准，将属于同类的答案赋予相同的代码，这个根据访员记述的受访人答案进行分类的过程就是开放性问题的编码过程。国内社会科学调查中常用的编码方法主要包括访员在实地进行手工编码，或由编码员在调查结束后集中手工编码。前者的优势是成本低，时效好，但编码质量难以控制 ；而后者的编码质量会较好，但需要在数据采集完毕后开始进行，并且在一定程度上增加了成本。在CAPI模式的社会调查中，由于数据传输的及时性，可以组织编码员在调查实施过程中就对已接收的开放性问题数据进行编码，这样不仅编码质量可以采取有效方式来控制，还提高了编码的时效性，使数据可以尽快提供给使用者。同时，编码员在编码过程中如发现一些访员记述信息不全以至于无法编码的数据条目，也可以通知督导及时联系访员补充信息，同时提醒访员遵守开放性问题答案的记录规范。CFPS在2011年追访调查中就采用了编码员即时编码的工作方式，并且在编码系统的辅助下，不仅编码质量得到控制，编码效率上也大幅度提高，降低了编码成本。

2．调查管理的数据支持。CAPI调查中数据管理的另一个重要任务就是为调查执行提供必要的数据支持，包括访问进度报告，辅助信息咨询，以及样本状态审核等。

（1）访问进度报告。调查管理的一个重要内容就是要把握访问的进展情况，而且对于不同层面的管理人员，需要了解的进度情况也有所不同。调查的执行主管较为关注整体的进展情况，各级督导则需要知道本辖区内访员的采访情况。在纸笔模式的调查中，这些情况需要逐级汇报，不仅要花费管理人员大量的时间和精力，而且报告的结果在准确度和时效性上都较差。而在CAPI调查中，由于可以直接采集到管理数据，并且数据可以及时上传，这些工作完全可以由数据管理人员直接承担。CFPS中为调查管理制作了多个进度表，如针对项目研究人员制作了各类问卷访到率、拒访率的报告；针对执行主管制作了各类问卷的访问进度报告，全国完成访问的家户比例等；针对各级督导制作了每日完访家户比例，每日完访各类问卷总数，采访结果频数统计，日访问量和周访问量等，并且这些结果在省、村居和访员三个层面汇总报告。此外，CFPS还开发了SAS EBI为基础的报告系统，实现报表每日更新，并且多用户可以同时查看并下载表报。为此，数据管理人员不仅要编程制作这些报表，还要对报告系统进行维护，设定用户权限，确保各类报表的正常展现。

（2）辅助信息咨询。在调查实施过程中，访员常常会遇到一些困难，需

要一些辅助信息的帮助。例如，不能确定受访户时需要询问该受访户其他家庭成员的姓名；无法联系受访户时需要查询受访户的补充联系方法；有时为了博取受访人的信任，还要查询上一波的调查数据以证明自己的身份。另外，督导们在分析受访户访问难度时，也需要一些关于受访户的基础信息。这些信息往往有多个数据来源，这时从各类数据中提取重要变量，建构一个信息完备，可供查询的数据库，可以显著提高工作效率，并且保证信息提供的完备。

（3）访问状态审核。在实地调查结束前的一个重要收尾工作就是要对所有样本的访问状态进行审核，确认是否将样本全部发给访员，是否有访员未联系过的样本，或者没有达到规定的联系次数或据访次数就放弃采访的样本，以及其他不是最终访问状态的样本。数据管理人员要将这些访问状态有问题的样本提取出来，并将其对应的管理数据中的联系记录等数据信息整理好，发给相应的督导，由督导根据这些信息与访员沟通，确定最终的访问状态，反馈给数据管理部门，再由数据管理人员对这些联系状态的结果代码进行清理。

3．质量核查的数据支持。有了并行数据的支持，CFPS 可以对访员行为进行有效的监控和核查。CFPS 质量核查的样本主要有两个选择方法：一个是按照一定比例从已经完成访问的样本中随机抽选；另一个是对可能存在采访问题的访员的所有完成访问的样本进行核查。对于有受访录音的首选录音核查；对于没有采访录音，但存有受访样本电话信息的采用电话核查；对于没有以上两种信息的，考虑使用实地核查。那么，如何发现可能存在采访问题的访员呢？CFPS 的经验是主要通过两个指标来进行过滤：一个是整份问卷的不合理采访用时比例；另一个是整个问卷的问题无回答率。CAPI 采访管理系统可以记录下采访过程中的痕迹数据，其中包括每道问题的采访用时。通过将访员采访的实际采访用时与合理采访用时比较，来考察访员是否按照采访规范进行提问并记录答案。如果在一份问卷中，不合理采访用时的问题过多，则将此访员列为“可能存在采访问题”的访员，并对其所有完成访问的样本进行核查。同理，我们也计算出每份问卷的没有实质性答案的问题比例，并根据经验确定一个标准，对于无回答比例高出这个标准的样本及其该访员完成访问的其他样本进行核查，这样可以用来防范访员采访时的一些不负责行为。

此外，在调查执行过程中还有一些指标的观测可以对访员采访质量进行监控，如访员完访问卷的录音比例，问题的跳问比例等。这些指标性数据不是访员直接采集或者系统直接存储，而是需要数据工作人员根据经验，从各类数据中提取相关变量进行加工，然后提供给执行管理或质量核查人员使用。

4．数据加载信息的整理和预备。数据加载（preload）指在调查实施前将一些数据信息提前导入到采访管理系统中，这是 CAPI 调查的一项重要环节，对于跟踪调查来说，数据加载工作更是必不可少。一般来说，需要加载的数据信息有两类：一类是受访样本的姓名、地址，以及联系方式，访员将借此找到正确的受访家户或个人；另一类是前一波调查的问卷数据，本波调查的一些问题需要据此来决定受访人回答问题的路径。

CFPS 的数据加载工作十分复杂。首先要涉及对受访样本的姓名、地址和联系方式的整理。在初访调查中，这类信息主要来源于地址绘图抽样的记录。但绘图阶段采集的信息差错很多，在初访调查时访员将对此进行修正，在跟踪调查中访员又会对这些信息进行补充或验证。然而，在调查过程中也会把这些信息弄得混乱，如访员用一个家户的问卷采访了另一个家户，这样联系信息与家户信息就无法对应了。此外，为了保证追踪调查中找到准确的家户，对于一些没有门牌号码的住宅，还要请访员加备注指示查找路线。同时，CFPS 在两波调查之间还会有各种形式的样本维护活动，也会采集到一些需更新的联系方式。由于这多种情形并存，貌似简单的样本联系信息的加载工作变得复杂，而这项工作上的失误将直接导致访员实地采访时无法联系上受访样本，甚至造成样本流失。

然后，要根据问卷设计的要求来整理问卷数据加载信息。这些问卷数据加载信息往往有多个来源，需要从家庭、个人等多个数据库中提取。有些信息可以直接提取变量，如受访人的基本背景信息，及对一些问题的回答选项；而有些信息则需要稍微加工，如对某些问题是否作答的判断。由于 CFPS 是以家户为对象的跟踪调查，这些信息都要整合在家户层面上，以循环加载的方式统一在同一个家户样本代码下面。所以对数据管理的统计技能要求也比较高，需要应用到矩阵拼接的方法来实现。

这两大类信息整理完毕后，最后就是要把这些信息存储为系统加载特定的文件格式，并且分门别类，以保证系统顺利加载。

三、CAPI 跟踪调查数据管理面临的挑战

CAPI 跟踪调查在提升数据质量的同时，也对数据管理提出了更多的挑战。首先是来自于技术上的挑战。前面提到 CAPI 跟踪调查采集到不同类型的并行数据，例如录音文件、录入痕迹文件，视频文件，GPS 数据等，这些文件的内容需要及时被转化成可进行统计分析的数字代码，目前对这些文件内容的编码标准、编码技术还在积极探索之中。除了编码技术以外，前文所述的矩阵拼接技术（用于数据加载），访问过程中的实时报表技术，动态化网络管理技术，亦或无回答权重调整时的多水平模型分析技术（并行数据和受访者数据多数不在同一层面），都对数据管理技术提出了新的要求。

其次在数据管理程序上带来了新的挑战。传统的数据管理程序通常是单项流程化的垂直管理。CAPI 跟踪调查中的数据管理因拓展到研究设计和数据采集过程，因此将在数据管理过程中面临多用户，多维、多向流程，以及动态网络化管理的局面。例如，设计访问系统的信息技术人员，质控人员，实地执行管理人员均已成为数据管理面临的服务对象；在数据加载维度，将与信息技术人员，实地执行管理人员构成两个循环的信息沟通流程，在数据清理阶段又要面对数据本身，又要面对访员，甚至还要涉及受访者，在这个维度上，将需要

多个信息循环过程。为此，探索出一套行之有效的科学管理程序是数据管理工作的必然之选。

伴随技术上和程序上的挑战，迅速而来的是对人员专业素质上的挑战。CAPI 跟踪调查要求数据管理人员不再局限于只了解数据本身，不再局限于传统的数据整理技术，不再局限在统计分析软件中，而是要拓展到研究设计者的思维，拓展到数据采集中的质量控制思维；要开发对录音、视频的文件的编码技术，提高动态化多维多方向科学管理的技术，甚至要求对 Oracle、SQLserver，SAS EBI 等软件有所了解和应用。 CS

参考文献：

[1] Couper, M. and Lyberg, L.: The use of paradata in survey research [J]. In Proceedings of the 55th Session of the international Statistical Institute, Sydney, Australia, 2005.

[2] 丁华等编著，地图地址抽样框制作手册[C]。北京：北京大学出版社，2011.

[3] 任莉颖,计算机辅助面访跟踪调查的数据特征与应用[J],中国统计 2012.2.

[4] 北京大学中国社会科学调查中心网站：www.isss.edu.cn.

作者单位：北京大学中国社会科学调查中心