China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-23

系列编辑：谢宇　责任编辑：胡婧炜

中国家庭追踪调查
区县数据库模糊方法[1]

崔雅红　吴琼　徐宏伟　王广州

2014.3.18

[1] 谢宇教授为数据模糊工作提供了非常有用的专业建议，在此感谢。

1. 区县数据库创建背景

追踪性社会调查需要对受访者的个人信息有详细准确的记录，以便下次追踪访问时能有效地找到原受访者。但受访者的个人信息需要受到严格保护，除了调查中的必要工作人员，其他个人不得接触。为了保护受访者个人信息安全，中国家庭追踪调查（CFPS）制定了相关规则，一方面保证了在调查过程中只有相关人员在必要的时候方可对部分信息进行接触；另一方面，在数据共享时，不仅隐去个人姓名、地址、工作单位等识别信息，而且将省级以下的地址代码采取再编码化，使用户无法通过编码辨识出省级以下的地址信息，以实现对个人信息的最高程度的保护。[2]

虽然屏蔽省级以下地址代码可以起到更好的保护受访者信息的作用，但对研究可能会产生不便，特别是希望用宏观变量（譬如一个地区的经济发展水平）来解释个人差异的相关研究。

为了满足保护受访者信息安全和研究的双重需要，CFPS 提供了可供用户特殊申请使用的区县级数据库，作为对现有公开数据库的补充。区县数据库中包括 CFPS 样本区县的一系列宏观变量，如地区生产总值（GDP）、人均地区生产总值（GDP per capita）、就业率等。具体变量见表 1 的第 1 和第 2 列。

2. 区县数据模糊化

CFPS 在区县数据库中虽然不发布各样本所在区县的国标码，但如果直接发布该区县相关的宏观数据仍存在风险。因为原始的区县数据来自于《中国统计年鉴》及其它政府公开信息，用户可以通过直接比对官方发布的数据推测出具体区县。为了进一步降低风险，我们对原始的区县级宏观数据进行模糊化的处理。

CFPS 区县数据模糊化的基本原则有两条：一，模糊化后的数据与原始数据有高度相关性，不影响统计分析的结果；二，通过模糊化降低用户利用数据推测出具体区县的风险。

[2] 详细的处理方法可参见谢宇，2013，《中国家庭追踪调查（2010）用户手册（第二版）》第 61-62 页。

3. 模糊化的具体方法

我们采用的基本方法是对于某个特定变量 y，在原值的基础上加一个在特定范围内变动的随机数 u 得到模糊化后的变量 m。其中 u 呈连续型均匀分布（uniform distribution），均值为零。用公式来表达则是，

$$m = y + u$$

$$\text{u}\sim\text{uniform}(-\text{a}, \text{a})$$

表 1. 总区县库[3]中各变量的原值区间及模糊化的变化范围

变量名称	变量标签	原值大小区间	a
GDP_TOTAL	地区生产总值(万元)	[10315,59458200]	$10^{(L-2)}$
GDP_PER	人均地区生产总值(元/人)	[2543,2194718]	$0.5 \times 10^{(L-2)}$
POPULATION	总人口(人)	[444,8220207]	$10^{(L-2)}$
EMPLOY_RATE_AGE20_59	劳动年龄人口比例(%) (20-59 岁)	[43.17,91.78]	0.2
EMPLOY_RATE_AGE15_64	劳动年龄人口比例(%) (15-64 岁)	[56.37,99.1]	0.15
AGING_RATE	老年人口比例(%) (65 岁及以上)	[0.24,19]	0.075
SEX_RATIO_AGE10_19	10-19 岁人口性别比	[83.15,2900]	0.3
EDU_AVG	平均受教育年限	[2,13.14]	0.06
URBAN_PEOPLE_RATE	非农业户口人口比重(%)	[1.58,99.4]	0.5
EMPLOY_RATE	就业率(16 岁及以上各种职业人口/16 岁及以上人口)	[0.28,0.93]	0.006

注：L 表示原值的整数位数。

对于那些数值变化范围特别大的原始变量，采用唯一固定区间（-a，a）可能会违背前面提到的数据模糊化基本原则。举例来说，区县的地区生产总值（GDP）在五位数到八位数之间变动，如果采用固定的 a，当 a 取值过大时，会造成在低 GDP 区间模糊化后的变量区分度太小，与原值相关性太弱，违背模糊化的第一条基本原则；而当 a 取值过小时，又会造成在高 GDP 区间无法达到模糊化的效果，违背模糊化的第二条基本原则。鉴于此种情况，对于数值变化范围较大的变量，我们采取根据原值大小区间来确定梯度性 a 的方法。在具体的实施过程中，a 的确定是一个循环优化的过程，我们不断的用检验结果（检验方法见下一

[3] 总区县库包括全国所有区县。我们基于全国所有区县做了模糊化，但仅发布了 CFPS 样本所涉及到的 167 个区县（含 2012 年新增区县）的数据。

节）来确定 a 的取值是否合理。表 1 中的第 3 列和最后一列分别给出了各变量的原值区间和最终确定的 a（其中 L 表示原值的整数位数）。

4. 模糊化数据检验

4.1 与原值的描述性统计数值比较

表 2 列出了原值与模糊化后数值的分布比较。由表可见，模糊化后的数据较好的保持了原数据的平均值与标准差。其中均值变化量的绝对值最大不超过 2‰，且大部分不超过 1‰（表 2 第 3 列）；标准差变化量的绝对值最大不超过 3‰，绝大部分不超过 1‰（表 2 第 4 列）。表 2 的最后一列显示原始变量与模糊化变量的高度相关性：相关系数均在 0.99 以上。

表 2. 总区县库变量模糊化前后变量比较

变量名称	统计量	均值	标准差	相关度
地区生产总值(万元)	原始值	1478867	2814964	0.9997
	模糊化值	1480601	2820635	
人均地区生产总值(元/人)	原始值	33087.3	63723.27	0.9998
	模糊化值	33091.1	63698.07	
总人口(人)	原始值	464394	388908.6	0.9993
	模糊化值	464182	388662.9	
劳动年龄人口比例(%)(20-59 岁)	原始值	62.29474	5.795016	0.9998
	模糊化值	62.29629	5.796052	
劳动年龄人口比例(%)(15-64 岁)	原始值	74.05822	5.066114	0.9998
	模糊化值	74.05952	5.06673	
老年人口比例(%)(65 岁及以上)	原始值	8.77253	2.267862	0.9998
	模糊化值	8.773139	2.268225	
10-19 岁人口性别比	原始值	111.5593	53.01797	0.9999
	模糊化值	111.5616	53.01489	
平均受教育年限	原始值	8.714383	1.467079	0.9997
	模糊化值	8.714864	1.467436	
非农业户口人口比重(%)	原始值	29.52869	23.55246	0.9999
	模糊化值	29.53268	23.55214	
就业率(16 岁及以上各种职业人口/16 岁及以上人口)	原始值	0.686857	0.099617	0.9991
	模糊化值	0.686955	0.099717	

4.2 排序检验

排序检验的核心思想是如果将模糊化前后的数值分别排序，那么与原值相比，模糊化后的数值在整个序列中的排序变化不大。表 3 显示各个变量模糊化前后的数值的排序仅在小范围内波动，最大不超过 9（见表 3）。这表明模糊化后的数值较好的保持了原始值的相对大小。

表 3. 总区县库变量模糊化前后排序比较

模糊化前后排序差别	最小值	最大值	均值	标准差
地区生产总值 (万元)	-5	6	0	1.375
人均地区生产总值 (元/人)	-3	5	0	1.124
总人口(人)	-6	5	0	1.694
劳动年龄人口比例%(20-59 岁)	-6	5	0	1.547
劳动年龄人口比例%(15-64 岁)	-5	4	0	1.44
老年人口比例%(65 岁及以上)	-4	6	0	1.523
10-19 岁人口性别比	-5	5	0	1.613
平均受教育年限	-5	6	0	1.635
非农业户口人口比重(%)	-4	6	0	1.478
就业率(16 岁及以上各种职业人口/16 岁及以上人口)	-7	9	0	1.482

4.3 对分析结果的影响：以相关性分析及回归分析为例

以下我们分别以相关性分析及回归分析为例，展示采用原始变量值与模糊化后的变量值对分析结果的影响。在相关性分析中，我们将CFPS成人库中的个人受教育年限与区县库中的人均GDP、平均受教育年限和就业率这三个宏观变量进行分析，其结果展示在表4中。由表中模糊化前后所得的相关性系数可见，使用原始变量值和模糊化后的变量值所得结果的差别细微。

表 4. 模糊化前后相关度比较

	与原值的相关度	与模糊化值的相关度
人均地区生产总值 GDP per capita(元/人)	0.220***	0.218***
平均受教育年限	0.366***	0.366***
就业率 (16 岁及以上各种职业人口/16 岁及以上人口)	-0.238***	-0.237***

注：*p<.05， **p<.01， ***p<.001。

我们接着使用多元线性回归模型具体分析了个人收入（income）与年龄、性别、受教育年限及人均地区生产总值之间的关系。其中个人收入、年龄、性别、受教育年限这四个变量

来自 CFPS 2012 年成人库，人均地区生产总值来自区县库。分析样本局限在个人收入为正数的样本。表 5 分别陈列了利用原始的人均地区生产总值变量及模糊化后的人均地区生产总值变量分析的结果。由表可知，采用模糊化变量对分析结果未产生实质性影响：两套分析结果所得出的回归系数接近，相应变量的 95%置信区间高度重叠。[4]

表 5. 多元回归模型的参数估计对比图

因变量 个人收入（元/人）	原始变量模型（95%置信区间）	模糊化变量模型（95%置信区间）
截距	-2071.01(-2774.07,-1367.94)***	-2050.13(-2753.33,-1346.92)***
年龄		
20 岁及以下	-5151.06(-6796.6,-3505.53)***	-5177.01(-6822.91,-3531.1) ***
20-30 岁	1935.6(974.32,2896.89)***	1918.12(956.62,2879.62) ***
30-40 岁	5152.41(4296.23,6008.59)***	5133.09(4276.74,5989.44) ***
40-50 岁	3459.27(2639.53,4279.02)***	3441.19(2621.29,4261.1) ***
50-60 岁	1468.2(622.06,2314.35)***	1459.8(613.44,2306.16) ***
60 岁以上	0	0
性别（男）	4665.87(4148.74,5183.01)***	4659.76(4142.5,5177.02) ***
受教育年限	1082.55(1024.58,1140.51)***	1086.25(1028.3,1144.2) ***
人均 GDP(万元)	743.88(696.47,791.29)***	737.45(690.11,784.79) ***
R^2	.15	.15

注：*p<.05, **p<.01, ***p<.001

5. 结论

为了更好的保护受访户信息，CFPS 数据在发布时采用了省级以下地址再编码化的方法。但为了满足研究的需求，我们提供了可供特殊申请的区县级数据库，并对库中的变量实施模糊化处理，降低利用区间数据识别出具体区间的可能性。模糊化主要采用在原值基础上加一个固定区间内的随机数的方法。经检验，模糊化后的数值与原值分布相似且有高度相关性，这表明在确保数据安全性的同时，模糊化对数据分析的结果影响微小。

[4] 需要指出的是，此例的目的仅在于展示模糊化前后分析结果的区别，而不在于优化模型。如果我们将因变量进行 log 变换，所得的 R^2 会大幅度提高到 30%左右。