网络出版时间：2015-12-31 14:59:36
网络出版地址：http://www.cnki.net/kcms/detail/11.4084.s.20151231.1459.016.html

增强个体代表性：基于日志数据的长期时间利用预测①

檀学文

[摘 要] 针对时间利用日志数据存在的个体代表性不足以及统计意义上的“多零”问题，论文借鉴已有文献的两部分回归方法，从日志数据预测长期时间利用数据。结果显示，利用人口和社会经济变量以及活动参与频率变量对长期时间利用的预测结果具有较好的稳健性，分布更均匀，同时大幅度减少了“多零”问题，清除了时间利用实证分析的一大障碍。因此，未来的时间利用调查如果添加非经常性活动参加频率的问题，将会有利于提高时间利用数据的利用程度和效果。

[关键词] 时间利用福祉；社会指标；预测

社会科学的研究对象是人，主要是由个体组成的群体或社会，其中个体包括居民以及企业、社会组织、政府等法人。一般来说，研究个体的目的主要还是为了研究群体，个体主要充当样本或案例。统计学以抽样方法获取有限数量的个体作为样本，以推断其所代表的总体的情况。在意识形态领域和社会科学方法论领域均有集体主义和个体主义之争，但是后者通常还是为群体性目标服务。但是，近年以来出现了直接以个体为对象和目标的研究方法。一个典型例子是在福祉研究领域，在 OECD 创建的网站上②，网民输入自己的各项指标，便可计算出自己的福祉指数——“更好生活指数（BLI）”；澳大利亚居民在给出自己对 7 项主观满意度指标分值后，也可以得出自己的福祉指数（AUWBI）。

在样本量足够大且具有代表性的情况下，其统计特征能够用于推断总体特征。但是如果要评价样本本身，那么就需要考虑指标的样本个体代表性问题。在经济社会研究领域，一个经常性的问题是所使用的指标能在多大程度上体现该指标所应体现的含义。用统计术语来说，就是如何增强概念的名义定义和操作定义的一致性[1]。例如，在 AUWBI 指数中，福祉的含义是近期的主观福祉状况，其具体的组成变量是 7 个关于近期生活各个方面的满意度评估，这些变量的含义以及时限与近期主观福祉的内涵都是一致的。然而 BLI 指数使用了多个维度的客观指标，其指标代表性就值得讨论。例如，就业或失业都是指最近两周的情况，时间利用是指昨日的时间利用，这些指标口径对于个体样本的近期状况来说具有很大的偶然性，代表性比较差。

增强指标的个体代表性，一方面是为了顺应当前个性化的时代趋势，另一方面也是为了改进定量分析效果。如果以一日时间利用数据或一日消费数据来代表个体的时间利用特征，容易出现大量特异值，如 0 或特别大的值，损害实证分析结果的解释力。对于这种类型的指标，就存在增强个体代表性的必要性。增强指标个体代表性的方法通常可以分为三种：扩大数据记录的时间区间、使用估计的而不是记录的数据、使用替代性指标，三者各有优劣。就时间利用而言，日志记录数据准确但是代价高，如果增加记录天数则代价更高；估计数据的代表性增强，而且调查成本低，但是其准确性降低；替代性指标与原指标的一致性有时会存在问题。这就是社会科学调查研究中经常面临的数据需求与获取之间的权衡取舍问题。本文以时间利用数据为例，对此进行探索，希望为依靠调查或统计数据进行的微观研究提供有益的数据改进思路。

[收稿日期] 2015-06-14
[基金项目] 本文是中国社会科学院创新工程项目“中国农民福祉研究”的部分成果。
[作者简介] 檀学文，中国社会科学院农村发展研究所副研究员，邮编：100732。
① 吴国宝研究员组织创新团队成员对论文进行了讨论，谭清香和杨穗专门提出了修改建议，特此表示感谢。
② www.oecdbetterlifeindex.org。

本文意图利用时间利用日志调查数据，估计具有更好样本代表性的长期时间利用数据，对其统计学特征进行检验，从而对时间利用数据的获取和应用提出相应的对策建议。正文包括四个部分。第一部分是关于长期时间利用预测的理论，包括作为参考的长期食品消费预测模型以及建立在这一模型基础上的长期时间利用预测模型。第二部分利用中国农民抽样调查数据，从日志时间利用预测长期时间利用。第三部分利用统计学原理，评价估计长期时间利用数据的统计学特征，评估其样本代表性。最后一部分对本文使用的研究方法和结果进行评价，对其可能的应用价值进行了说明。

一、长期时间利用预测理论和方法

（一）居民福祉与时间利用

从传统经济研究和福利经济学角度，经济增长被视为福利改进的主要甚至唯一标志。福祉研究超越传统福利经济学的上述强假设，提出多维度、多指标表征福祉的必要性和可行性。除了用消费指标代替收入指标外，还有健康、社会联系、时间利用、主观福祉等多个领域的指标[2]。已有的多维福祉框架中，无论是社会层面还是个人层面的，大部分都包含时间利用或个人活动维度。时间利用通常情况下都是以时间在不同活动间的分配和使用状况来表征居民在这项重要资源的利用方面的福祉状况[3]。根据对福祉的不同定义，时间利用与福祉的关系大体上有三条指标选择和研究路径，即扩展的经济福祉、实时性主观福祉和多维客观福祉[4]（见表 1）。其中，后二者属于个人福祉范畴，可以分别称为主观时间和客观时间[5]。本文遵循多维客观福祉理论，将时间利用视为多维福祉的一个客观维度，与教育、经济等其他维度并列。如表 1 所示，即使在多维福祉框架下，时间利用指标也有主观指标和客观指标之分。其中，主观指标主要是对时间利用状况的主观评价，而客观指标主要是对实际时间利用的记录或回忆/估计。

表 1　对应于不同福祉内涵的时间利用指标及其数据来源

福祉类型	扩展的经济福祉	实时性主观福祉	多维客观福祉	
福祉内涵	总福祉产出不仅包括可交易的商品和服务的经济价值，也包括无酬家务劳动、志愿活动、闲暇活动等的经济价值。后者用非市场劳动影子报酬乘以该劳动持续时间表示。	福祉是人们各时刻主观心理状态的加总。将每个时段积极的和消极的主观情感评价分值分别进行累加，便可得出一定时间内加总的实时性主观福祉。	福祉是人们不同方面生活状态的加总，总体上分为客观福祉和主观福祉。其中客观福祉又可称为客观生活质量，包括健康、教育、住房、社会联系等多个维度。时间利用通常是客观福祉的维度之一，可以用时间利用指标来表征这个维度的福祉状况。	
时间指标类型	非市场劳动持续时间	各主观情感的持续时间	评估的时间利用(主观) 时间利用满意度 自我评价工作生活平衡 自我评价时间压力	实际时间利用（客观） 时间利用日志数据 跟踪或观察记录的时间利用 估计的时间利用 特定活动参与率
数据来源	时间利用调查	以时间利用日志为基础的专门主观福祉调查	问卷调查	时间利用调查/问卷调查/统计资料

资料来源：根据文献[3] [6] [7]整理。

本文的分析对象是作为客观指标的时间利用日志数据。时间利用日志调查记录受访人的基本信息以及在调查前一天 24 小时内的所有活动情况。调查表通常以 10 分钟为单位，将 24 小时划分为 144 个连续的时间单元。受访人按顺序依次填写每项活动的具体内容、持续时间、同时发生的其他活动、活动的地点以及与什么人在一起等。有时候，时间利用日志调查也通过问卷调查的方式进行，由调查员询问受

访人并填写问卷。时间利用日志调查仅调查受访日前一天发生的活动，而且按时间顺序排列，所以是最为准确的时间利用数据。时间利用日志调查表在填写、回收后，经过对具体活动内容对照时间利用同类活动分类代码表进行编码、归类，便可获得受访者的一日时间利用数据。例如，2008 年，国家统计局在 10 个省、市开展了第一次居民时间利用调查，共获得 3.7 万个居民样本[8]。这次调查的城乡居民大类平均活动时间如图 1 所示。从中可见，城乡居民时间利用有明显差别，主要体现在农民有酬劳动时间比市民长很多，而闲暇时间则短很多。

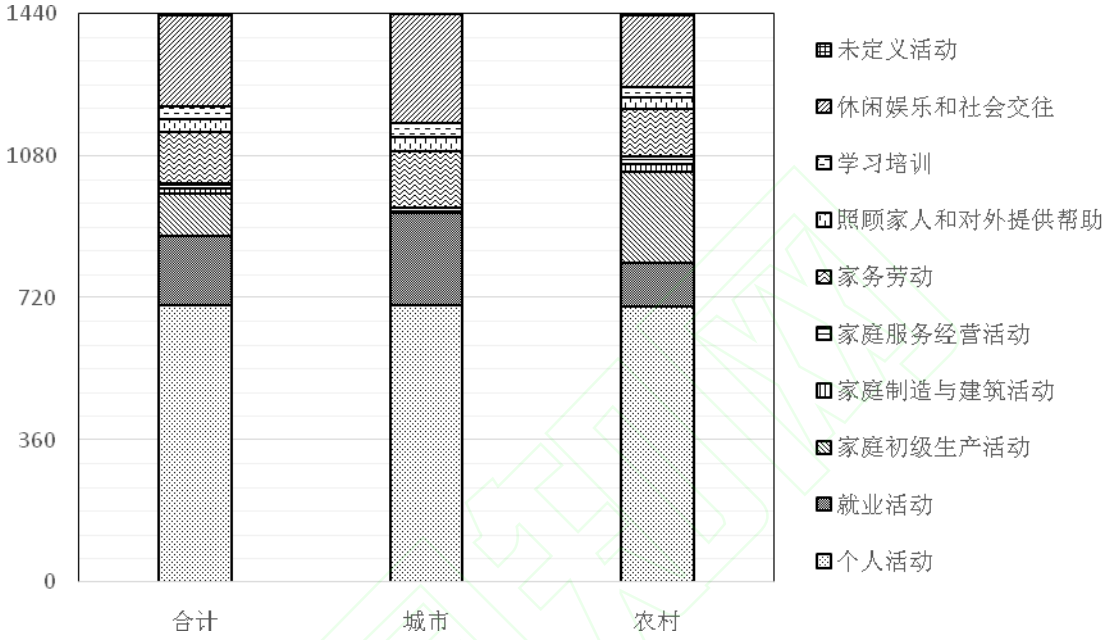

图 1　2008 年按大类划分的城乡居民时间利用状况

数据来源：《2008 年时间利用调查资料汇编》，中国统计出版社，2009 年。

（二）从日志数据预测长期数据的方法

1.通常食品消费数量预测

从随机性短期数据估计长期数据的方法较早地以及较多地用于营养和健康领域的食物消费。居民食物消费数据具有与时间利用日志数据类似的形式，即受访者对某日 24 小时内所有消费的食物的记录或根据记忆的估计数据。类似于时间利用，一日的食物消费行为具有偶发性，实际食物消费数量对于通常食品消费数量而言存在典型的测量误差，包括个人误差和人际误差，一般通过回归校正法予以调整[9]。根据消费频率，食品可以区分为日常性消费食品和偶发性消费食品。一项研究对这两类消费数据的误差修正方法进行了区分[10]。对于日常性消费食品，在传统的混合模型基础上，通过使用 Box-Cox 变换，将实际消费数据的偏态分布转换为接近于正态分布，可以估计出实际消费数据的个人误差和人际误差。对于偶然性消费，论文采用了两部分测量误差模型：第一个方程用于估计消费某类食品的发生概率；第二个方程用于估计某类食品在发生消费的情况下所存在的两类误差，与对日常性消费食品所使用的模型相同。该模型具体表达如下：

$$p_i \equiv P(R_{ij}>0|i)= H(\beta_{10}+\beta^i_{X1}X_{1i}+\mu_{1i}),j=1,\ldots,J_i \quad (1)$$

$$A_i \equiv E(R_{ij}|i;R_{ij}>0) = \beta_{20}+\beta^i_{X2}X_{2i}+\mu_{2i}+\varepsilon_{ij} \quad (2)$$

其中，公式（1）为 logistic 回归，估计第 i 种食品消费在第 j 日的发生概率 p_i，X_{1i} 为有关的解释变量，μ_{1i} 为人际误差；公式（2）为 OLS 回归，估计第 i 种食品消费在第 j 日实际发生的情况下，其预测的消费数量，X_{2i} 为有关的解释变量，随机误差 μ_{2i} 和 ε_{ij} 分别表示人际误差和个人误差。

从而，第 i 种食品的通常消费数量，也就是长期估计值，等于其发生概率以及在发生情况下的预测值的乘积，即：

$$T_i \equiv E(T_{ij}|i) = p_i A_i \quad (3)$$

该模型为混合效应模型，每个方程都包含固定效应和随机效应。两个方程存在联系，不仅两者的人际误差 μ_{1i} 和 μ_{2i} 是相关的，而且它们的解释变量中至少有部分变量是共同的。

在进行经验估计时，解释变量的选择除了人口特征变量外，还包括了食品消费频率（FFQ）变量作为补充变量。利用美国健康与营养调查数据（NHANES），该论文证明，通过将食品消费数据和食品消费频率数据结合起来，即将 FFQ_i 添加为解释变量 X_i 的一部分，能够提高通常食品消费预测以及饮食—健康关系估计的精确性。

2.长期时间利用预测

食品消费数据和时间利用数据虽然都是记录 24 小时内发生的事件，而且也都包含日常性事件和偶发性事件，但是它们实际上存在着很大差别：时间利用数据的单位是时间，如小时和分钟，受总量约束，即一天的所有活动时间加总后必然等于 1 440 分钟；食品消费数据的单位是数量，如克或公斤，加总后无总量约束。由于总量约束，一天内不同活动的时间存在替代关系，一类活动时间的增加必将导致其它某类活动时间的减少；而食品消费则不存在这种严格的替代关系，不同类型的食品消费是相互独立的。

基于每日时间总量约束以及用一系列闲暇活动的参与频率代表个人行为“习惯”的社会学理论[11]，Gershuny 提出了一种基于上述两部分模型但是相对简化的估计方案[12]，可以表达如下：

$$p_i^a \equiv P(T_i^a > 0|i) = H(\beta_{10} + \beta_1^{ia} X + \mu_{1a}) \quad (4)$$

$$t_i^a \equiv E(T_i^a|i;\ T_i^a > 0) = \beta_{20} + \beta_2^{ia} X + \mu_{2a} \quad (5)$$

$$LTT_i^a \equiv E(T_i^a|i) = p_i^a A_i^a \quad (6)$$

其中，公式（4）预测第 i 个受访者的在调查日的活动 a 的发生概率 p_i^a，X 为有关的解释变量，μ_{1a} 为随机人际误差；公式（2）预测第 i 个受访者的活动 a 在调查日实际发生的情况下，其预测的活动时间，X 为有关的解释变量，μ_{2ai} 表示随机人际误差；公式（6）是估计的个体 i 的活动 a 的长期活动时间（即通常活动时间）。公式（6）的预测结果满足下述条件：

$$\sum LTT_i = 1440 \quad (7)$$

公式（4）—（6）的函数形式分别与公式（1）—（3）相同。两个模型的差别在于：

①食品消费模型使用面板数据，从而可以同时估计个人随机误差和人际误差；时间利用数据利用截面数据，只能估计人际误差。

②食品消费模型中，添加的 FFQ_i 变量是单一变量，只在估计第 i 类食品消费时使用该类食品的 FFQ；时间利用模型中，添加的习惯变量是组合变量，即一组各类闲暇活动的参与频率，对所有类型活动的估计是一样的。

③食品消费模型独立地估计各类食品的消费数量；但是时间利用模型同时估计各类活动的长期时间，结果受一日时间加总约束。

根据此项研究，上述长期时间利用估计方法至少可以解决日志数据存在的两个主要问题：一是闲暇活动等不经常发生的活动时间的“多零”问题[①]；二是数据正向偏斜和右尾极端值问题，由此增强数据的个体代表性。

二、中国农民长期时间利用预测

[①]“多零”是指在居民时间利用调查数据中，当活动分类足够细化时，很多类型的活动时间都会显示为 0，但这并不代表受访者的这些活动时间真的为 0。如果以它们作为自变量进行回归，也会对回归结果造成干扰。

（一）数据

长期数据预测的基本思路是，短期行为是长期行为的一部分，从短期行为数据一方面可以估计其长期发生的概率，另一方面估计该行为在发生情况下的数值，该估计值与估计概率的乘积即为长期估计值。利用这种方法，可以用时间利用日志数据估计长期的时间利用规律，即通常情况下个人的 24 小时都用于哪些活动。这种方法的前提是需要更多的变量支持，对于常规的时间利用日志数据或饮食日志数据是不适用的。上述 Gershuny 使用的“ Time Diary Study 2000/01 ”数据中除了日志数据，还有一系列活动参与频率变量，后者是估计所需的重要解释变量，代表着人们的行为“ 习惯 ”。借鉴上述方法，我们在调查问卷中设计了类似的活动参与频率的问题，为预测长期时间利用提供了条件。

本文使用中国社会科学院创新工程项目“ 中国农民福祉研究 ”2013 年农村居民抽样调查数据。调查内容包含家庭成员、主观福祉、劳动与就业等 12 个方面。其中，时间利用部分包括昨日时间利用日志、闲暇时间满意度以及闲暇活动参与频率三类问题。该调查在位于辽宁、江苏、湖北、宁夏和贵州 5 个省的 10 个县、市进行，在每个县、市各抽取 5 个行政村，每个村预定抽样规模为 20 人。样本省分别位于东部、中部和西部，具有一定的地域代表性。省内的样本县、市按照经济发展水平抽取，基本处于中等水平。县、市内的样本村通过随机抽样或者按照经济发展水平高低进行抽取。在样本村内，居民样本分布于不同的村民组和不同的收入和生活水平，具有一定的村庄代表性。调查问卷均由调查员提问和填写。本次调查一共回收 1 000 份有效问卷，其中 860 份问卷拥有完整时间利用数据，是本文预测长期时间利用的数据基础[①]。

（二）预测步骤与结果

借鉴 Gershuny 建立的方法，本文以时间利用日志数据为基础，预测个人的长期时间利用分布。主要预测步骤如下：

1.原始数据处理

包括时间利用活动类型重新归类、部分解释变量重新编码、缺失值处理等。各种时间利用统计活动分类都有大小不同的差别。中国国家统计局 2008 年时间利用调查将活动分为 10 个大类、66 个中类和 115 个小类。本文根据分析需要以及中国农民很多闲暇活动参与率极低的现实，将一些大类合并，将闲暇活动分为 4 种类型，合计将活动类型分为 11 类。为满足模型回归需要，对部分变量进行重新编码、缺失值处理。其中，对婚姻状况、健康状况、教育、社会身份等都进行了重新编码。

2.活动的参与概率预测

以重新归类的 11 类活动时间为基础，将其转换为以 0 或 1 表征的“ 是否参与 ”变量：若活动时间大于 0，新变量编码为 1，表示当日参与了该活动；若活动时间为 0，新变量编码为 0，表示当日未参与该活动。以该新变量为因变量，以特别选定的变量为自变量，用 logit 方程估计个人对各类活动的参与概率。自变量分为人口和社会经济特征等控制变量以及活动参与频率变量两类，前者包括年龄、年龄平方、性别、婚姻状况、健康状况、需照料家庭成员情况、教育、工作类型、调查日类型（工作日或周末）、最近一周累计工作时间以及省份虚拟变量；后者包括 14 类非经常性闲暇活动参与频率变量，代表个人活动习惯（表 2）。

表 2　代表习惯的不经常性闲暇活动参与频率变量

[①]通常情况下，考虑到一日时间利用数据代表性问题，时间利用调查需要考虑具体的时间选择问题，有些国家（如韩国）的时间利用调查甚至在一年内针对同一样本进行 2 次到 4 次调查，力图以此来增强其代表性。对于中国农民来说，应当考虑地域差异以及季节差异（农忙、农闲），而工作日和周末的差异是次要的。2008 年中国居民时间利用调查时间为 5 月份，各地总体上都是农忙季节，具有较好的代表性。本研究使用的时间利用调查是与农户问卷调查结合进行的，调查时间受总体调查安排的约束。不过，2013 年农村居民抽样调查是在 7 月至 9 月期间进行，总体上也都是农忙季节，但是并非最忙碌或农闲的时候，所以也具有一定的代表性。

活动编码	活动内容	参与频次	参与频率
711	阅读活动，看书、看报、看杂志等（非学习）		
719	上网		
721	一般性的散步、漫步、溜达等		
722	跑步、快走等健身活动		
724	舞蹈、健身操之类的健身活动		
725	各种球类运动		
731	棋牌游戏，如打麻将、打扑克、下棋等		
732	玩游戏（包括电脑游戏、手机游戏等）		
733	去歌厅、舞厅等娱乐场所		
741	看电影		
742	外出旅游		
745	在本村或附近观看文化表演、扭秧歌或唱歌、露天电影等		
746	不以购物为主要目的的逛街（包括赶集等）		
754	参加节庆、聚餐、红白喜事等活动		

注：表中参与频次指过去一年内的参与次数，最高为 365；参与频率分为 5 个等级：全年最多 1 次、每月不足 1 次、每周不足 1 次、每周 1 到 4 次、每周 4 次以上。

3.活动的参与者参与时间预测

以重新归类的 11 类活动时间为因变量，以上述两类变量为自变量，用最小二乘回归方程估计个人对各项活动的参与者参与时间。此处使用的自变量与步骤 2 中的 logit 回归相同。

4.活动的长期平均参与时间计算

将步骤 2 和 3 的结果相乘，得出个人各项活动的长期平均参与时间的预测值。

5.长期时间利用估计值调整

对步骤 4 的结果进行负值调整和总和调整。将小于 0 的估计值调整为 0；并以加总值与 1 440 的比值为调整因子，对预测的长期平均参与时间进行调整，使得他它们的加总值仍然为 1 440 分钟。

由此得出的估计的长期时间利用分布如表 3 所示。

表 3　长期时间利用预测结果　　单位：分钟

活动类型	日志时间		预测的长期时间			
			（总和未调整）		（总和调整后）	
	均值	标准差	均值	标准差	均值	标准差
睡觉	561.7	118.91	561.7	46.37	562.9	55.88
个人活动	137.3	82.05	137.6	27.31	137.8	27.73
工作及学习	92.1	202.54	92.1	104.18	90.8	99.17
家庭经营	259.9	268.25	260.0	159.66	259.9	156.84
家务劳动	102.2	136.37	102.2	60.22	102.4	60.62
护理与帮助	43.5	138.55	43.9	61.89	43.5	59.18
使用媒体	117.2	112.48	117.2	40.39	117.3	40.47
体力闲暇	26.0	66.45	26.0	36.71	25.9	35.56

文化娱乐	31.8	97.62	31.9	59.49	31.0	55.86
社会交往	42.3	99.53	42.3	41.35	42.2	40.37
其他活动	26.1	87.29	26.4	29.24	26.3	28.66
合计	1440.0		1441.3		1440.0	

注：此表及以后各表采用下述方案 III 的结果。

（三）预测结果可靠性与稳健性检验

与 Gershuny 的依据类似，从日志时间得到的样本总体各项活动的平均时间和样本长期时间配置的总体均值应该近似相等，本文的计算结果符合此条件。为了进检验预测结果的可靠性，我们分别对 3 套解释变量方案进行估计：方案 I 仅以上述控制变量对被解释变量进行回归；方案 II 和 III 同时以控制变量和不经常性活动参与频率变量对被解释变量进行回归，其中后者在方案 II 中采取参与频次形式，在方案 III 中采取频率形式（见表 2）。结果显示，3 套方案的预测结果都极为接近，分别是 1 440.4 分钟、1 442.0 分钟和 1 441.3 分钟，这表明模型设置具有较好的稳健性。方案 II 和 III 使用了不经常性活动参与频率变量，各方程回归结果显示，它们的 R^2 和 PseudoR^2 值都明显地大于方案 I，表明模型的解释能力得到了较大的提升（见表 4）。方案 III 的 R^2 和 PseudoR^2 值总体上稍大于方案 II，但是差别非常小，表明不经常性闲暇活动频率变量可以用分类形式代替原始频率形式且不损失效率。

表 4　长期时间利用预测的三套方案拟合效果比较

	方案 I		方案 II		方案 III	
	Pseudo R^2	R^2	Pseudo R^2	R^2	Pseudo R^2	R^2
睡觉		0.1149		0.1446		0.1521
个人活动		0.0776		0.1108		0.1115
工作及学习	0.1952	0.2035	0.2369	0.3034	0.2255	0.3151
家庭经营	0.1643	0.3112	0.1988	0.3783	0.1965	0.3908
家务劳动	0.2914	0.0471	0.3159	0.0884	0.3075	0.1006
护理与帮助	0.1657	0.2258	0.1866	0.368	0.1954	0.3481
使用媒体	0.0974	0.0386	0.1333	0.0693	0.1431	0.0733
体力闲暇	0.1707	0.1538	0.3287	0.309	0.3303	0.2572
文化娱乐	0.1678	0.2036	0.3752	0.3783	0.3979	0.4774
社会交往	0.1187	0.2146	0.1677	0.2975	0.1815	0.3296
其他活动	0.0621	0.3438	0.0987	0.4388	0.0872	0.4664

注：睡觉和个人活动的参与频率被设定为 1，故没有为它们设立概率估计方程，从而也就不存在 Pseudo R^2。

三、预测前后的时间利用分布比较

（一）总体时间利用比较

从近期研究成果看，在统计上，中国农民时间利用具有典型的发展中国家特征，即睡眠时间足够；有酬劳动时间更长；休闲娱乐和社会交往时间更短，且以消极闲暇活动为主；无酬家务劳动时间也偏短；女性劳动时间长而闲暇时间短[4]。农民的时间利用分布在 2008 年与 2012 年以及 2013 年都比较接近，表明他们的时间利用规律是比较稳定的。对 2013 年农民时间利用日志数据和长期估计数据的统计特征进行比较显示：在总体上，两类数据的平均值极为接近，T 检验显示差异均不显著。但是预测的长期时间利用比日志时间利用的统计分布更加均匀，即估计后的标准差、偏度、峰度都比估计前大幅度下降了（表 5）。

表 5　日志时间与预测时间利用描述统计比较　单位：分钟

	均值			标准差		偏度		峰度	
	日志	预测	t 值	日志	预测	日志	预测	日志	预测
睡眠	561.7	562.9	0.29	118.91	55.88	0.67	-0.02	6.02	3.13
个人活动	137.3	137.8	0.19	82.05	27.73	3.07	0.28	20.90	3.75
工作或学习	92.1	90.8	-0.20	202.54	99.17	2.00	1.55	5.53	4.95
家庭经营	259.9	259.9	0.00	268.25	156.84	0.70	0.20	2.34	2.06
无酬家务劳动	102.2	102.4	0.04	136.37	60.62	2.06	0.66	8.23	2.76
照顾和帮助他人	43.5	43.5	0.02	138.55	59.18	3.93	2.60	19.87	11.09
使用媒体	117.2	117.3	0.02	112.48	40.47	1.43	0.09	6.74	3.06
体力闲暇	26.0	25.9	-0.07	66.45	35.56	3.67	2.05	20.27	8.05
休闲娱乐	31.8	31.0	-0.28	97.62	55.86	3.36	2.53	14.42	9.49
社会交往	42.3	42.2	-0.02	99.53	40.37	2.93	2.14	12.22	10.95
未定义活动*	26.1	26.3	0.07	87.29	28.66	4.40	1.76	24.55	6.48

注：社会和政治参与活动所发生的样本极少，为方便分类，将其并入未定义活动内。

此外，预测的长期时间利用数据大大减少了日志数据中存在的“多零”问题。尽管我们在计算中使用的简化分类已经大大减少了活动类型的数量并降低了发生零的可能性，但是日志数据中仍然有大量的零存在。除睡眠和个人活动外，其他 9 种活动时间为 0 的情况平均达到 65%之多。而在预测数据中，该比例下降到 1.9%（表 6）。

表 6 预测时间与日志时间相比含零样本数量及其变化

	日志时间为 0（例）	预测时间为 0（例）	变化率（%）
睡眠	0	0	—
个人活动	2	0	-100.0%
工作或学习	559	10	-98.2%
家庭经营	228	4	-98.2%
无酬家务劳动	263	0	-100.0%
照顾和帮助他人	592	31	-94.8%
使用媒体	179	0	-100.0%
体力闲暇	542	11	-98.0%
休闲娱乐	618	13	-97.9%
社会交往	532	8	-98.5%
未定义活动*	594	45	-92.4%

注：样本量为 702 个。

进一步地，我们可以形象地考察各类活动的时间分布特征。在图 2 所列举的三类活动中，睡眠时间的分布最为接近于正态分布，尤其是对于长期估计值而言；接下来是工作时间，其分布偏度较小，但是峰度明显比睡眠时间小，即显得更为平坦；闲暇时间分布的偏度比正态分布大，而且是向右偏，但是其峰度与睡眠时间接近。分活动的长期预测时间与日志时间相比最大特点就是其分布更加集中，不对称程度也有所下降。

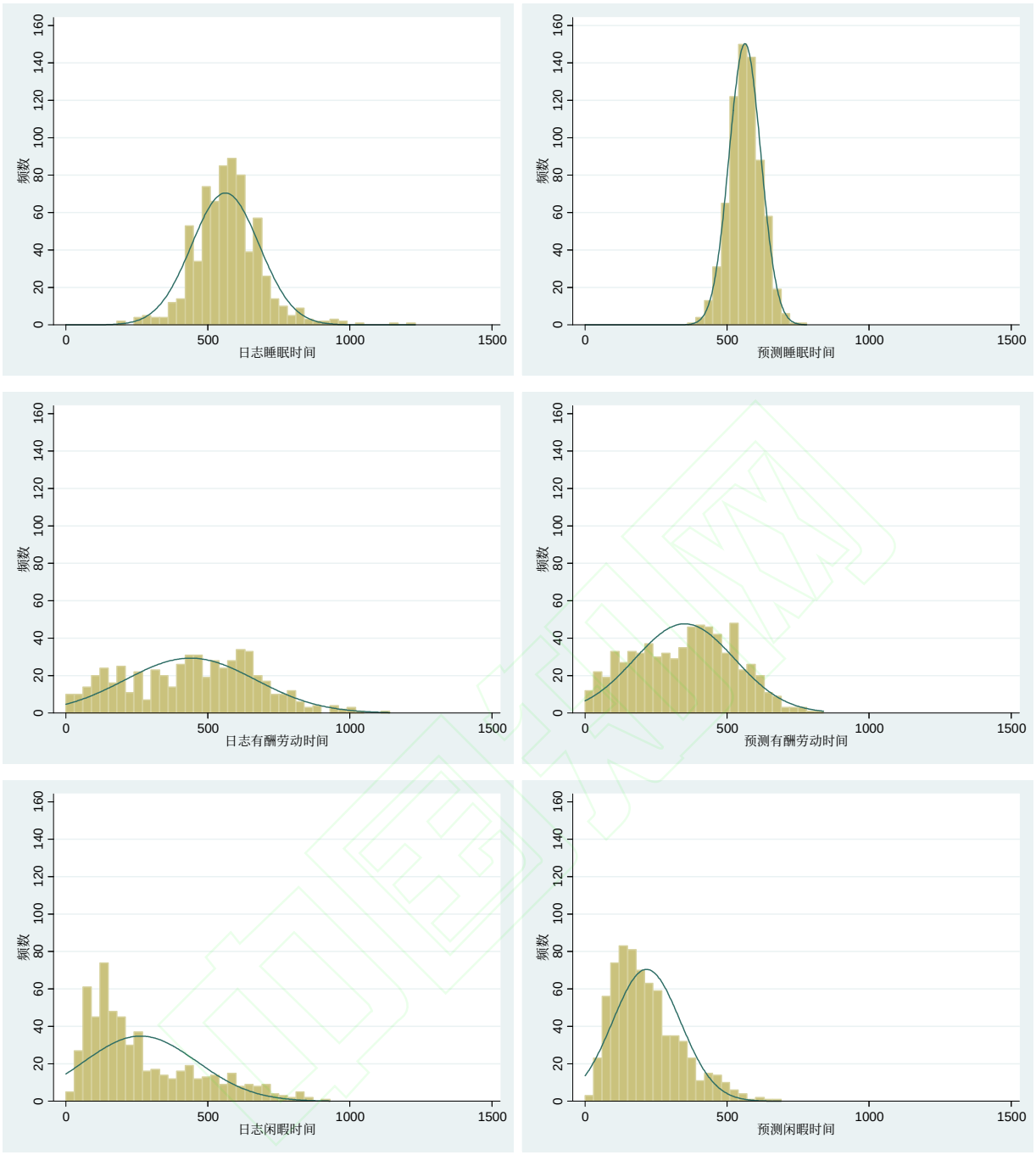

图 2　三类活动的日志时间与长期预测时间分布对比

（二）分群体时间利用比较

我们对预测前后的时间利用进行了分群体比较，包括分性别和分年龄组比较，其中按年龄分为三组：40 岁以下、40 至 59 岁、60 岁以上。对不同群体的日志时间和预测时间分别进行 T 检验，结果显示，除了极个别情况之外（40 岁以下组的个人活动时间的日志值和预测值差异显著），包括性别分类和年龄分组，几乎所有活动时间的日志值和预测值的差异都是不显著的（表 7）。为节约篇幅，分性别的结果比较省略。

表 7　分年龄组的日志时间与预测时间比较　　单位：分钟

	40 岁以下			40 至 59 岁			60 岁以上		
	日志	预测	t 值	日志	预测	t 值	日志	预测	t 值
睡眠	571.9	570.1	-0.17	544.1	551.4	1.39	589.3	580.6	-0.99

个人活动	116.2	125.9	2.39*	140.4	134.6	-1.33	148.6	154.6	0.95
工作或学习	117.9	128.9	0.69	117.6	109.7	-0.77	18.3	20.2	0.30
家庭经营	241.2	227.2	-0.74	264.1	267.3	0.29	266.9	272.1	0.35
无酬家务劳动	100.5	98.0	-0.23	103.4	102.7	-0.12	101.1	105.4	0.49
照顾和帮助他人	87.7	85.3	-0.17	36.2	34.6	-0.29	21.3	26.9	0.77
使用媒体	110.5	110.9	0.04	118.0	120.1	0.37	121.0	116.8	-0.59
体力闲暇	12.1	14.2	0.58	26.1	25.2	-0.37	37.5	36.9	-0.11
休闲娱乐	26.6	23.5	-0.53	25.4	26.7	0.36	49.1	45.9	-0.49
社会交往	31.8	32.5	0.11	37.3	41.2	0.96	61.2	52.3	-1.00
未定义活动	23.6	23.5	-0.02	27.4	26.5	-0.20	25.7	28.5	0.51

注：*表示在 5%水平上差异显著，双尾检验。

四、总结与讨论

作为一项衍生性或工具性任务，本文致力于从日志时间利用数据预测长期时间利用数据，其目的是提高时间利用数据的个体代表性。我们借鉴一项英国学者的研究成果，利用课题组的抽样调查数据，估计了长期时间利用的预测数据。从预测数据的统计学特征看，预测数据具有较好的稳定性和比日志数据具有更好的个人代表性。对于日志数据存在的“多零”问题，预测结果对其有了很大的弥补。从而，预测的长期时间利用数据可以更好地用于时间利用指标构建以及福祉决定的实证研究。

如表 1 所示，时间利用指标有主观指标和客观指标之分，类型很多，通过比较判断各类指标的优缺点以及选择更好的指标是时间利用研究的一项有价值的任务。本文对长期时间利用的预测可以对这项工作有所贡献，可以用预测的长期时间利用指标与其他类型指标进行比较。长期时间利用预测对数据要求比较高，除了日志数据还需要大量的个人特征变量以及闲暇活动频率变量，对问卷长度和调查成本形成挑战。但是无论如何，该投入对于增加时间利用数据的整体价值是有利的。中国到目前为止只开展了一次官方时间利用调查。我们预期未来中国必将进行更多的时间利用调查。从而我们建议在未来的调查中对全体样本或者部分样本收集更多的信息，例如预测长期时间利用所需的控制变量以及活动频率变量，以便于更好地开展时间利用数据分析和研究。

[参考文献]

[1] 巴比.社会研究方法（第 10 版）.邱泽奇译.北京：华夏出版社，2005:122

[2] Stiglitz J E,Sen A, Fitoussi J-P. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. http://www.stiglitz-sen-fitoussi.fr/, 2009

[3] Gershuny J. *Time-Use Surveys and the Measurement of National Well-Being*. Swansea, UK: Office for National Statistics, 2011

[4] 檀学文.时间利用对个人福祉的影响初探——基于中国农民福祉抽样调查数据的经验分析.中国农村经济,2013 (10)：76-90

[5] Robinson J P. Using Time as Social Indicator. *Social Indicators Network News (SINET)*, 2013(114－115):1-7

[6] Bloom N, KretschmerT, Van Reenen J. Work Life Balance, Management Practices and Productivity//Freeman, Shaw (ed.). *International Differences in the Business Practices and Productivity of Firms*. The University of Chicago Press,2009

[7] 檀学文，吴国宝.福祉框架下时间利用研究进展.经济学动态，2014(7):151-158

[8] 安新莉，殷国俊.2008 年时间利用调查结果简介.国家统计局网站（http://www.stats.gov.cn），2008-11-21

[9] Carroll R J, RuppertD, StefanskiL A et al. *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. Boca Raton, Florida: Chapman and Hall CRC Press, 2006

[10] Kipnis V, Midthune D, Buckman D W. Modeling Data with Excess Zeros and Measurement Error: Application to Evaluating Relationships between Episodically Consumed Foods and Health Outcomes.*Biometrics*, 2009(65): 1003-1010

[11] Bourdieu P. *Distinction*. London: Routledge and Kegan Paul, 1984

[12] Gershuny J. Too many zeros: a method for estimating long-term time-use from short diaries. *Annals of Economics and Statistics*,2012(105/106): 247-271

Enhancing Individual Representativeness：Predicting Long-term Time Use Based on Diary Data

Tan Xuewen

Abstract: As for problems of weak sample representativeness and "*too many zeros*" in statistical sense, this paper uses two-part modeling methods of existing literature to predict long-term data from diary time use data. The results show that, the predicted outcomes of long-term time use are statistically robust and more evenly distributed. Moreover, the large-extent reduction of "*too many zeros*" problems can be realized at the same time, which helping clearing a major obstacle involving in the time use of empirical analysis. Hence, there is no doubt that adding questions like non-recurring event participation frequency in the future time use survey, would improve the efficiency of the utilization of the time use data.

Key words: Time use；well-being；social indicators；prediction