清华大学

中国经济研究中心

研究动态

总字 186 期　　2020 年 3 月 20 日

****

精准扶贫中的贫困识别与解决办法

李祥瑞

“精准扶贫，重在精准”。在精准扶贫中，精准识别农村贫困人口是扶贫工作的基础和前提。只有精准识别扶贫对象，才能做到扶贫项目安排、资金使用、措施到户、驻村帮扶目标对象的精准，取得精准的脱贫成效。但是在目前精准扶贫过程中，由于信息不对称，部分贫困户低报自己的收入以享受扶贫政策的现象是困扰精准识别的重要障碍。

针对农户低报收入这一精准扶贫过程中的痼疾，研究小组[1]以独立第三方身份，历时两年收集西部、东部、东北某三县在 2014－2017 年共 66267 条数据，分析了低报收入现象的产生机理和影响，并提出相应解决方案。研究创新性地采用大数据和机器学习方法，能够有效克服当前贫困户识别中的数据失真问题，在总量和个体两个层面有效识别贫困户。

一、精准扶贫中收入低报的主要危害

准确识别扶贫对象是精准扶贫的关键前提。在精准扶贫工作中，收入低报现象主要带来三方面危害。

第一，造成扶贫资源的严重浪费。农户低报收入使得本来不是贫困户的农户获得本应用于扶持贫困户的资源，造成大量资源错配和浪费。目前，抽样调查显示，三年来西部某县因收入低报带来的扶贫金错配达 275 万元，东北某县达 404 万元，东部某县达 647 万元。以 2018 年全国扶贫支出 4770 亿元为基准，扶贫金错配比例按照三县中最低的比例约 4.86%推算，大约有 239 亿元资金被错误配置，使得原本应当得到扶贫资金的农户没有享受到应有资源。

第二，降低贫困户提高收入的积极性。低报收入现象和资源错误配置不仅不能够有效扶贫，而且会助长部分贫困户“等靠要”思想。在调研过程中，我们发现一部分贫困户养成依赖补贴的习惯，导致发展动力不足，发展能力减退。在贫困识别线附近，低报收入人群比未低报收入人群的劳动收入增长率平均低 20%左右。

第三，影响扶贫工作的公平获得感。扶贫工作本质是社会主义优越性的重要体

1研究小组主要由清华大学经济管理学院师生组成，从 2017 年开始在全国多个相对贫困地区开展精准扶贫调研。

现，是公平正义在我党工作中的集中体现。但是低报收入现象的存在，使得某些非贫困户被错误评定为贫困户，享受大量不应有的补助。这相当于否定了之前通过劳动致富家庭的努力，不利于扶贫工作达到预期政治效果。在调研工作中，我们发现“假贫困户”的存在非常容易引发村民的不公平感，甚至有部分群众将其归咎为政府（干部）处置不公，引发干群矛盾。

二、精准扶贫中收入低报的识别悖论

准确识别贫困是扶贫工作的关键和难点。现有扶贫工作中，对贫困户进行评定主要采取自上而下（指标规模控制、分级负责、逐级分解）与自下而上（村民民主评议）相结合的机制。一方面，自上而下的科层制方法难以克服逐级传达中的信息不对称性问题。在识别中机械采用文牍工作、层层填报等方法，不仅难以确保获取真实信息，还造成大量形式主义负担。另一方面，自下而上的村民民主评议面临农村人情社会的潜规则，宁多勿少、利益均沾等现象大量存在，导致贫困户存在低报收入现象。此外，由于工作量庞大导致可行性有限，独立第三方或者抽查方法也难以建立一套可供对照的可信数据。这使得决策者不仅难以从总量层面了解目前究竟有多少贫困户、其中有多少存在收入低报，而且难以清晰识别具体的贫困家庭，实现精准扶贫。

收入低报现象是扶贫工作的痼疾，传统方法在应对收入低报时面临主观悖论。现行机制下，贫困户收入低报是长期存在的痼疾。为了克服这一痼疾，避免信息失真，现有扶贫工作大量采用反复核对和多次抽查等监督方法，试图解决数据失真问题。但是这些监督仍然主要依靠人的主观观察。正如社会科学“坎贝尔定律”所指出的，“一项定量数据越被用于社会决策，它在过程中被经手人主观歪曲的可能性越大”。在精准扶贫这项规模庞大、体系复杂的工作中，这一定律体现得更为明显。扶贫工作中，越是想要通过工作人员反复核对来克服收入低报，就越有可能增加贫困识别中的主观歪曲，带来数据失真。这就形成了一个新的贫困识别与数据失真之间的主观悖论。

三、以大数据技术解决精准扶贫中收入瞒报低报问题

为克服贫困识别过程中的主观悖论，可以利用现代大数据技术，重新建立一套贫困识别体系。具体而言，可以分为总量估算体系和贫困户个体识别体系两部分。

第一，总量估算体系。总量估算体系主要用于宏观层面，在县级层面考察有多少贫困户，以及有多少贫困户存在收入低报和瞒报现象。这一方法的优点是不需要另外获取数据，完全基于现有数据计算可得，可以直接用于精准扶贫工作，识别整体层面的贫困户规模。具体而言，该方法主要基于聚束分析，通过对农户报告收入的整体分布情况进行考察，估计收入被低报的水平。正常情况下，如果没有低报行为，收入的整体分布应该是平滑的。但如果存在低报行为，即略高于贫困标准的贫困户低报收入以获取补助，那么将导致收入分布出现“聚束”现象：略低于贫困标

准的观测值数量大幅增加，而略高于贫困标准的观测值数量出现缺口。通过对缺口大小进行计算，可以估计农户收入被低报的比例。

基于以上方法，我们对调研得到的 2014－2017 年贫困户收入数据进行研究。研究发现，略低于贫困线的农户数目显著多于略高于贫困线的农户数目，表明贫困线附近的农户确实存在收入低报现象。而且随着补贴比例越高，在贫困线附近，略低于贫困线和略高于贫困线的农户数量差距越大，意味着农户通过低报收入获取福利补助的可能性越高。下图显示了贫困对象收入的分布情况，图中呈现明显的聚束现象：经过标准化后[2]，0 点左侧呈现观测集聚的现象，而右侧则出现大量观测缺失。也就是说，有部分略高于贫困线标准的贫困户低报了收入，导致略低于贫困标准的人群分布明显增加。

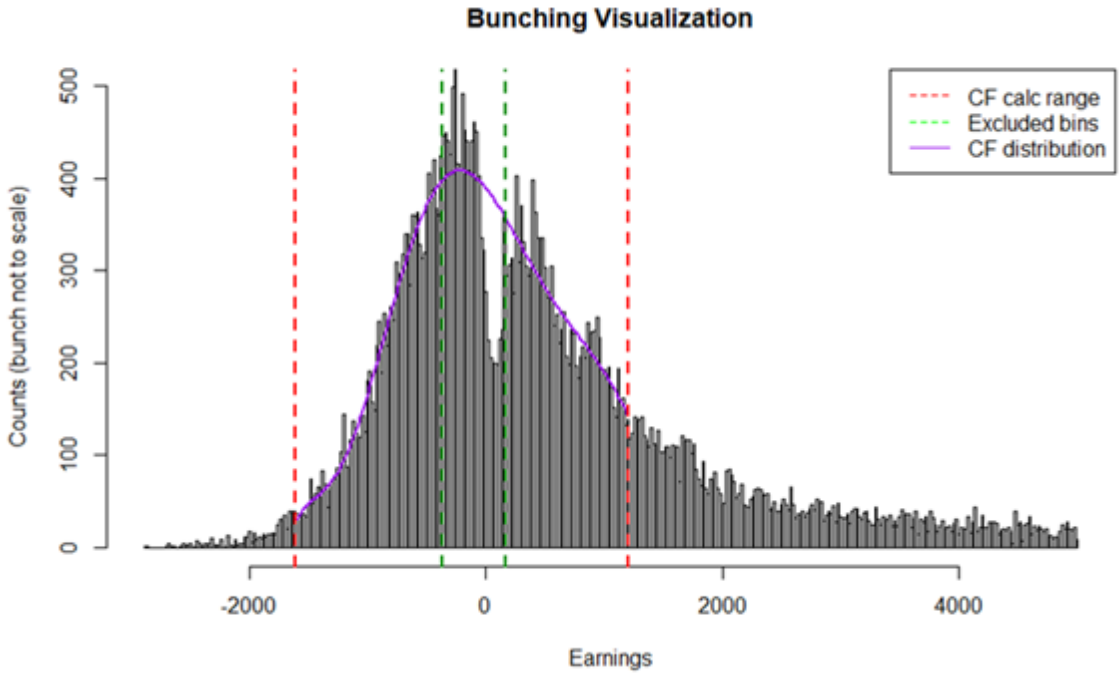

图 1　基于西部某县整体样本的聚束分析

对上述分布中的缺失部分进行估计，我们发现在现有贫困识别机制下，西部某县低报人数占建档立卡户总人数的比例接近 5%，东部某县低报人数比例为 12.78%，东北某县低报人数比例为 6.26%。结合相应贫困家庭所享受的各项补贴，估计约 4.86%的资金没有投放到真正贫困的农户手中。进一步结合补贴强度，还可以刻画出每单位补贴所对应的收入低报概率，便于在宏观层面更好地刻画和识别贫困。

第二，贫困户个体识别体系。贫困户个体识别体系主要用于辅助精准扶贫工作人员准确判断个体家庭收入，减少低报现象。这一体系主要采用机器学习的方法，通过家庭主动汇报的收入状况，结合家庭人口、住房、生产、生活条件等客观特征对贫困户的收入进行识别和判定。通过大数据机器学习，建立对贫困的预测模型，在一定程度能够避免人的主观判断失误。在具体的模型选择中，采用 SVR（支持向量回归）方法对调研得到的数据进行机器学习，预测贫困线附近贫困户的真实收入。由于收入信息只在贫困线附近存在较大扰动和误差，假定远离贫困线人群所报告的收入是准确的。因此将其作为机器学习的原始输入数据，经过机器学习训练模型，获得收入和其他家庭特征之间的关系，最后用于估计贫困线附近人群的收入情况。研究发现，针对基于机器学习所判定的“贫困户”，收入分布中的聚束效应不再存在。这表明利用机器学习校准收入，识别贫困户，能够有效克服收入低报现象。

[2]标准化后的 0 点为贫困线。

四、结论与政策建议

贫困识别的精准与否关系脱贫攻坚的成败。研究小组两年来的调查研究表明，目前低收入人群收入低报以获取政策补贴的现象仍较为普遍，不仅浪费了扶贫资源，而且打击了农户发展积极性，影响了扶贫的公平性。造成这一现象的一个重要原因是目前贫困识别中的主观因素和客观数据的结合还不够完善。采用大数据和机器学习方法可以有效克服当前贫困户识别中的数据失真问题，解决精准扶贫中收入瞒报低报问题，有必要将大数据技术有机地融入精准扶贫的各个环节，为此我们建议：

第一，推广科学评估机制，精准考核扶贫效果。精准考核是精准扶贫、精准脱贫的重要一环。目前精准扶贫、精准识别贫困户的考核主要依靠群众满意度调查，上级部门考核以及第三方机构评估。然而满意度高低取决于群众主观判断，同时部分考核评估人员是在校学生，缺乏农村生活工作经验，可能出现误判。以上因素影响考核结果的准确性，伤害扶贫一线工作人员的积极性。我们提出的经济计量模型可以更科学地评价精准扶贫的工作成效，减少主观误判，减少考核不专业现象发生。经济学分析表明，收入低报现象与对贫困户的补贴强度有关，给予贫困户的好处越多，农户就越有激励低报收入。因此，评价精准识别贫困的工作效果，应当综合考虑识别误差和对贫困户的补贴情况。基于实际数据的分析表明，借助聚束分析和断点回归等现代计量方法，只需要整体收入分布数据，就能够计算出单位补贴下的收入低报概率。在此基础上，能够以极低成本实现对精准识别工作的客观评价，为现有考核体系提供有益补充。

第二，以模型预测家庭收入，为实际收入提供参考。减少低报行为对于精准扶贫的成功具有重要意义。虽然可以通过加强审计和事后惩罚的手段减少低报，但这通常需要更多的额外投入。家庭收入状况会通过家庭的生产生活状况和其他特征显现，因此可以通过家庭各种变量推断家庭的合理收入范围。应用大数据和机器学习等技术手段，校准收入数据，可以为确保收入的真实性提供参考。

主办：清华大学中国经济研究中心　　联系电话：（010）62789695
地址：清华大学经济管理学院　　邮　　编：100084