农村贫困人口的聚类与减贫对策分析

王 瑜 汪三贵

[摘 要] 使用 K-means 聚类方法对我国农村贫困地区的贫困人口进行聚类,并进一步分析特殊类型贫困地区、集中连片贫困地区的贫困类型结构。结果表明,贫困类型的分布呈现了扶贫对象在区域间分布的不平衡性,各种贫困类型的不同特点和区域分布上的差异从一个视角揭示了收入差距特别是贫困程度差异化的来源。尤其是,少数民族地区的贫困特征和贫困人口比重都要比老区和边境县地区更加突出,而这些地区有着自身独特的特点和性质,尤其需要对少数民族地区贫困背后的形成机制开展更加深入的研究,以便提出针对少数民族地区的因地制宜的扶贫开发措施。连片特困地区的主导贫困类型各不相同,意味着片区扶贫开发需要具有片区针对性的扶贫政策。尽管聚类分析只是一种探索性分析,但是农村贫困人口的聚类仍然为我们定义各种贫困的类型、以及它们在不同区域或特定区域划分之间的内部分布结构提供了非常有价值的信息,并将为进一步的统计推断分析提供基础。

[关键词] 农村贫困人口; K-means 聚类; 特殊类型贫困; 连片特困地区; 区域分布

DOI:10.13240/j.cnki.caujsse.20140704.010

一、引言

2011 年,《中国农村扶贫开发纲要(2011—2020 年) 》的颁发,标志着我国有计划的扶贫开发进入了新的阶段,其中,“提高扶贫标准,加大投入力度,把连片特困地区作为主战场,把稳定解决扶贫对象温饱、尽快实现脱贫致富作为首要任务”表明了新阶段十年扶贫工作的重点和战略核心。连片特困地区①、特殊类型贫困地区②无疑成为了我国扶贫攻坚的主战场,因此,认识各连片特困地区和特殊类型贫困地区的基本贫困特征和内部贫困结构则是有效实施扶贫开发政策的基础。

尽管贫困地区的贫困人口人群特点各有不同,但贫困人群内部仍然具有一些类似的贫困特征:一些人口因资源禀赋的欠缺和区位因素的制约而陷于贫困,例如人均耕地不足、居住在偏远山区;

[收稿日期] 2013-12-31

[基金项目] 本文系 2010 年国家社会科学基金重大招标项目“我国特殊类型贫困地区扶贫开发战略研究”(项目号: 10zd&025) 以及“中国人民大学科学研究基金(中央高校基本科研业务费专项资金资助) 项目”(项目号: 13XNH153) 研究成果,同时感谢国家留学基金委对第一作者在美访问研究的资助(录取文号: 留金发 [2013] 3009) 。

[作者简介] 王 瑜,中国人民大学农业与农村发展学院博士研究生,中国人民大学反贫困问题研究中心助研;
汪三贵,中国人民大学农业与农村发展学院教授、博士生导师,中国人民大学反贫困问题研究中心主任,邮编: 100872。

① 以六盘山区、秦巴山区、武陵山区、乌蒙山区、滇桂黔石漠化区、滇西边境山区、大兴安岭南麓山区、燕山 - 太行山区、吕梁山区、大别山区、罗霄山区等区域的连片特困地区和已明确实施特殊政策的西藏、四省藏区、新疆南疆三地州 14 片集中连片贫困地区,是我国农村贫困人口集中分布的地区。

② 所谓特殊类型贫困地区,是指老少边贫困地区,多位于经济发展落后的中西部山区和丘陵地区。其中,老区是指在第二次国内革命战争和解放战争时期,在中国共产党领导下创立的革命根据地,它们所在的县即为老区县; 少数民族地区主要指民族自治地方 155 个(其中自治区 5 个,自治州 30 个,自治县(旗) 120 个) ,中国半数以上贫困人口在少数民族地区; 边区是指沿陆地国境线的县级行政区划单位(新疆建设兵团 56 个边境团场未在统计范围内) ,共有陆地边境县共计 134 个。

另一些人口可能是由于文化和社会环境的制约而无法摆脱贫困,比如少数民族地区特殊的生计模式、语言和文化的制约使得他们在市场经济中受益相当有限; 还有一些人口可能是由于他们家庭人口特征的制约,比如缺乏劳动力或有无法工作的成员、人力资本水平比较低、缺乏工作的技能等。

传统的基于一维的贫困地区分类方式在实践和应用中显得力不从心,因此需要一种能够反映贫困地区内在的多种特性的细分和聚类方法,来综合反映不同贫困地区多方面的特征,这便需要运用可解决多变量的、大数据量的细分的数据挖掘技术。鉴于 K-Means[6] 聚类算法在处理大数据量和多变量数据分析方面有相对优势,本文将采取 K-means 聚类方法对我国特殊类型贫困地区贫困人口进行进一步分类,或可弥补传统地域划分(贫困县与非贫困县、东中西部划分等) 的缺陷,从而有利于找出各个区域中限制发展与致贫的关键制约因素,并可以进一步深入分析这些因素相互作用的方式和路径,研究缓解这些制约因素的可能方式和先后顺序。在政策的规划和实施上,同一类地区的政策可以相互地借鉴。对农村贫困人口进行分类,可以研究形成中国农村不同类型贫困人口的深层次原因。具体来说,贫困人口可以按照不同的特点分成不同的类型,并据此识别不同地区中贫困人口类型的构成。常规的贫困分析通常在贫困与非贫困人口之间进行比较,而把贫困人口划分为不同类别,则可以为不同地区制定具有针对性的扶贫政策提供基本依据。

二、数据和方法

(一) 数据来源

本文使用的数据来源于国家统计局 2006 年和 2010 年的农村贫困监测调查。该调查的地域范围是分布于中西部 21 个省(自治区、直辖市) 的 592 个国家扶贫开发工作重点县(简称扶贫重点县) 。涉及调查的省(自治区、直辖市) 有: 河北、山西、内蒙古、吉林、黑龙江、安徽、江西、河南、湖北、湖南、广西、海南、重庆、四川、贵州、云南、陕西、甘肃、青海、宁夏和新疆。调查对象为全国 592 个扶贫重点县中的 5 000 多个行政村,以及 5 万多个农村常住户①。数据包括农村贫困监测调查县级统计数据、社区调查数据、住户基本情况调查数据和个人调查数据。这里主要使用其中的住户基本情况调查数据和个人调查数据。调查抽样方式是自加权随机抽样,在全部 592 个重点县,以县为总体,与人口规模成比例的两阶段抽样,先抽村再抽户。

(二) 贫困标准的选择

本文的分析同时使用了两个标准的贫困线。第一种是原有的低收入标准,即官方公布的根据历年物价指数调整的低收入标准,2006 年为 958 元,2010 年为 1 274 元,为了行文便利,简称旧贫困线; 第二种是 2011 年《中国农村扶贫开发纲要(2011—2020 年) 》中提出的将农民人均纯收入 2 300 元(2010 年不变价) 作为新的国家扶贫标准,由于本文涉及该标准在 2006 年的应用分析,所以还需将此标准倒推至 2006 年; 而由于 2009 年的农村 CPI 指数虽比上年下降(是上年的 99. 7%) ,但所公布的当年旧贫困线仍然定为与 2008 年保持不变(1 196 元) ,因此,为了保持两个标准在前后分析和比较中的一致性,本文对新标准的倒推是根据旧贫困线的历年变化指数倒推而非直接用农村 CPI 指数倒推。根据此原则,新标准的贫困线在 2006 年为 1 729 元。

(三) 数据分析方法

数据分析技术可以广义分为两种类型[5]: ①探索性和描述性的,即研究者没有预定义的模型

① 在住户调查表中,农村常住户是指在农村范围内居住或即将居住半年以上的家庭户。户口不在本地而在本地居住或即将居住半年及以上的住户也包括在本地农村常住户范围内; 有本地户口,但举家外出谋生半年以上的住户,无论是否保留承包耕地都不包括在本地农村住户范围内。

或者假设,但想要推断高维数据的总体特征或者结构; ②验证性和推论性的,即研究者想要使用可用数据来验证一个或一组假设(模型) 的有效性。数据聚类主要属于第一种数据分析技术,即探索性和描述性的分析技术,而作为一种探索性分析工具,聚类分析结果有助于进一步提出可验证的假设和模型。聚类分析是数据挖掘中的一个重要研究领域,是一种数据划分或分组处理的重要手段和方法。聚类是无监督的分类,也就是它没有先验知识可用,其聚类技术有很多类,目前聚类算法大体上分为基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法等[5-6]。聚类算法不仅可以作为发现数据库中数据分布的深层次信息的工具,还可以作为数据挖掘中的一个预处理步骤。

表 1 2006—2010 年两种扶贫标准的贫困线

年份	农村居民消费价格指数(上年 = 100)	旧标准贫困线(单位: 元)	新标准贫困线(单位: 元)
2006	101. 5	958	1 729
2007	105. 4	1 067	1 926
2008	106. 5	1 196	2 159
2009	99. 7	1 196	2 159
2010	103. 6	1 274	2 300

数据来源: 农村居民消费价格指数来源于《中国统计年鉴-2011》; 旧标准贫困线历年数据来源于《中国农村贫困监测报告-2011》; 新标准贫困线 2010 年数据来源于《中国农村扶贫开发纲要(2011—2020 年) 》,其他年份数据根据旧标准贫困线历年数据倒推得出。

K-means 算法[1]就是基于划分的聚类算法中的一个典型算法,划分式聚类算法还包括 K-modes 算法[7-10]、K-means-CP 算法[11]、FCM 算法[12]、基于图论的聚类算法[5]等。而 K-means 算法与 K-means-CP 算法使用不同数值型数据集进行的比较实验结果表明,K-means-CP 算法丝毫也不优于 K-means 算法[13]。数据聚类在诸多领域具有长久的历史,K-means 聚类方法可以说是其中最流行最简单的算法之一,它于 1955 年被首次提出,尽管在此之后有几千种其他聚类算法被提出,但是 K-means 仍然被广泛使用[6]。

1967 年,MacQueen 所提出的 K-means 算法,其聚类目标就是根据输入参数 k 把数据对象分为 k 个簇,属于动态聚类算法,也称为逐步聚类法,不断迭代调整中心直到完成。其基本思想和步骤[6,14]是: (1) 首先输入初始数据集并指定划分的簇的个数 k; (2) 任意选择 k 个数据对象点作为初始聚类中心; (3) 根据簇中对象的平均值将数据对象赋给最类似的簇; (4) 更新簇的平均值; (5) 计算聚类准则函数 E; (6) 重复步骤(3) ~ (5) 直到准则函数 E 值不再变化; (7) 输出满足平方误差准则函数收敛的 k 个簇。关于 K-means 算法的具体步骤和算法展示可以参见 Hartigan[15-16]的论述。

K-Means 算法的一个最大的优点就是操作简单、采用误差平方和的准则函数、对大数据集的处理上有较高的可伸缩性和高效性。本文选择 K-means 算法作为聚类方法,主要是基于该算法的操作简便性、应用广泛性以及聚类结果比较稳健的特点。此外,本文选择用截面数据做聚类分析,而不是面板数据的聚类分析,主要是由于截面数据的聚类分析技术成熟并被广泛使用,而面板数据的聚类分析在文献方面比较缺乏,在技术上更没有形成比较一致的业内认同①。

指定表示聚类个数的 K 值,是该算法的一个重要步骤,也是可能存在挑战的方面,因为我们很

① 国内的面板数据聚类分析方法研究可以参看朱建平、陈民恳提出了针对单指标面板数据的聚类方法[3],李因果、何晓群在重构面板数据相似性测度的距离函数和 Ward 聚类算法的基础上提出的面板数据聚类方法[4],但是这些方法尚待业内讨论,也没有公开的分析程序包可供调用。

难事先确定合适的 K 值。但仍然有一些方法帮助我们找到合适的 K 值,一方面,从先验知识、理论常识和实践经验的角度,我们预设的 K 值一般是在一定可选范围内的,比如,从研究和实践的角度,我们愿意将人群分为几类,而不是几十类,我们会提出一些重要的聚类所用的变量,并且希望聚类的结果是这些特征的一些组合,因此聚类分析时一般也不会选择大于变量数的 K 值; 而另一方面,从统计的角度讲,可以按照一些统计标准对不同 K 值产生的聚类结果进行选择,例如,根据 Calinski和 Harabasz 提出的基于方差比的寻找最理想聚类数的指标[17],通过每一个 K-means 聚类结果的 Calinski-Harabasz 指数的比较来选取出最理想①的统计区分来体现这些类别。

三、农村贫困人口的聚类

聚类分析方法非常适用于分析那些自然形成的群体,例如自然形成的贫困家庭。对贫困群体进行聚类分析可以更清楚地了解贫困形成的原因和贫困的分布。贫困家庭被分为不同的类别,每一个类别内部的家庭都有着相似的特点。

(一) 贫困人口家庭特征分类

根据调查数据的基本特性、调查人口的实际特征,并结合对世行研究报告[18]中的参考,本文将贫困家庭②的特征分为 4 大类 11 种。表 2 描述了这些类型特征的具体定义。山区和远离县城从地理区位和公共资源的可获得性两方面描述了贫困家庭的劣势。土地是农村家庭重要的经济资源,劳动力的教育和家庭成员的健康是农村家庭重要的人力资本。位于少数民族地区、边境地区、革命老区的家庭可能面临更差的自然环境和社会政治环境,由此致贫。高抚养比意味着家庭中老人和儿童的数量较多,老人和儿童基本没有生产能力,并且老人容易罹患疾病,儿童需要教育投资,这会导致家庭的贫困。如果缺失经济资源和人力资本,农村家庭就很容易陷入贫困。外出务工和乡镇企业就业是农村中最主要的两种非农就业,非农就业是提高农民收入的重要途径。

表 2　农村贫困地区人口家庭特征分类

特征类型	具体特征③	描 述
资源禀赋和区位因素	山区	家庭居住在山区
	远离县城	距家庭所在村最近的县城距离超过 15 公里
	有限的土地	家庭人均土地面积小于平均水平的二分之一
老少边区	少数民族地区	家庭居住在少数民族县
	边境地区	家庭居住在陆地边境县
	革命老区	家庭居住在革命老区县
家庭人口特征	高抚养率	抚养比为家庭老人和儿童与家庭总人口的比例,比率大于 0.4 为高抚养比
	低劳动教育	家庭人均教育年限低于平均水平的二分之一
	有不健康成员	家庭中有残疾、重病和慢性病的成员
劳动就业	无乡企人员	家庭中没有成员在乡镇企业就业
	无打工人员	家庭中没有成员外出打工

表 3 是 2006 年和 2010 年不同贫困特征在农村人口和相应贫困标准下贫困人口中所占的比

① *F* 值越大,结果越理想。

② 这里的贫困家庭是按照收入贫困线确定的贫困家庭,在分析时会指出对应的贫困线标准。

③ 这些特征都是 0 - 1 变量,1 表示符合这一特征,0 表示不符合这一特征。

例。可以发现,相对于农村地区整体而言,农村贫困人口居住于山区的比例更高,居住地远离县城的比例更高,土地更加有限;更有可能居住在少数民族地区、边境地区;家庭的抚养比更高,更低的劳动教育水平以及更高的有不健康成员的比例;打工或在乡镇企业工作的机会更少。

表 3 2006 年与 2010 年农村贫困地区人口特征分类比例(%)

贫困特征	2006 年			2010 年		
	占贫困地区人口比例	占旧标准贫困人口比例	占新标准贫困人口比例	占贫困地区人口比例	占旧标准贫困人口比例	占新标准贫困人口比例
山区	63.3	65.9	66.6	63.6	61.8	65.3
远离县城	68.1	71.2	70.5	68.5	68.2	69.1
有限的土地	45.6	56.6	53.8	46.7	57.6	56.4
少数民族地区	43.0	47.5	45.7	42.8	40.4	42.3
边境地区	7.5	8.4	7.8	7.6	7.7	8.4
革命老区	18.0	17.1	17.8	17.9	14.5	15.4
高抚养比	33.2	36.0	34.9	31.0	32.6	32.1
低劳动教育	79.1	84.5	83.1	74.2	81.0	79.2
有不健康成员	10.5	13.0	11.3	9.9	11.9	10.6
无打工人员	58.5	73.4	66.1	55.4	75.8	66.2
无乡企人员	97.8	98.6	98.5	97.8	98.6	98.5

除了革命老区和 2010 年的少数民族地区特征外,几乎所有其他特征中,贫困人口都具有比农村人口总体更高比例的贫困特征。举例来说,2006 年,63.3% 的农村人口居住在山区,68.1% 居住地离县城 15 公里以上,45.6% 人口所拥有的土地面积是所有人口土地面积平均水平的二分之一以下,43% 的农村人口生活在少数民族地区,7.5% 的人口生活在边境地区,18% 的人口生活在革命老区,33.2% 的农村人口生活在抚养比超过 0.4 的家庭中,79.1% 的农村人口生活在平均受教育水平低于总体平均二分之一水平以下的家庭中,10.5% 的农村人口生活在有残疾、重病和慢性病成员的家庭中,58.5% 的农村人口生活在没有外出打工人员的家庭中,97.8% 的农村人口生活在无乡镇企业就职成员的家庭中,而除了革命老区这个特征之外,2006 年无论哪种贫困标准下的贫困人口都具有更高的特征比例。

在贫困特征中,无论对贫困地区总体而言,还是对贫困地区的贫困人口而言,有几项贫困特征比例很高,其中在人口中占比超过 40% 的特征有山区、远离县城、土地有限、少数民族地区、低劳动教育水平、无打工人员、无乡企人员。这些特征方面,对于扶贫开发政策制定以及贫困人口自身的脱贫来说,都可能是巨大的挑战。

在贫困人口中,那些更贫困的人口往往具有更高比例的贫困特征。以 2006 年为例,除了山区和革命老区这两种特征之外,其他各种贫困特征中,958 元贫困标准下的贫困人口的特征比例,比 1 729 元贫困标准下贫困人口中的特征比例都要高,这意味着,更贫困的人口具有更加突出的贫困特征。比如,2006 年,66.1% 的新标准贫困线下的贫困人口生活在没有外出打工成员的家庭中,而在旧标准贫困线以下的贫困人口中这个比例是 73.4%。

在贫困人口与贫困地区人口总体的贫困特征差异方面,贫困人口在低劳动教育水平、无打工人员这两个特征方面差别尤其突出,并且有差别扩大的趋势。2006 年,在贫困地区人口总体中,低劳动教育水平的比例为 79.1%,旧贫困标准和新贫困标准的贫困人口中这个特征占的比例分别为 84.5% 和 83.1%,2010 年,在贫困地区总体人口中,低劳动教育水平的比例降低为 74.2%,而在旧

贫困标准和新贫困标准的贫困人口中这个特征占的比例分别为 81.0% 和 79.2%，尽管这种贫困特征的总体比例在下降，但是贫困人口这一特征的比例下降得更慢。作为贫困地区家庭脱贫重要渠道的外出务工，可能也因务工比例的减少而降低了贫困家庭的减贫与脱贫的机会。

比较 2010 年与 2006 年的特征分类比例可以发现，贫困人口的特征占有比例没有太大变化，这说明这些贫困人口的贫困特征是相对稳定的; 但另一方面，也有一些特征的比例在趋势上有所变动，比如山区、远离县城、少数民族地区、革命老区这几个特征，2010 年在贫困人口中的特征比例都要比总体农村人口中的特征比例低，并且比 2006 年有所下降，而这些类型往往是相互关联的，比如少数民族地区和革命老区往往也是山区和远离县城的地区，这有可能与该期间内政府对老少边区特殊类型贫困地区更多的关注和扶贫政策倾斜有关。

(二) 农村贫困人口聚类

上述贫困特征是我们对农村贫困人口进行聚类的依据。聚类是一种无监督的分类，即事先并不能先验性地知道分为几类(簇) 以及哪些类(簇) ，而在 K-means 算法中，需要预先设定类别数 K，然后由 K-means 将数据对象划分为 K 类。我们没有先验的知识，但是却可以根据已有的贫困特征以及实际意义赋予几个不同的 K 值，分别进行聚类。根据表格描述的贫困状况以及聚类的实际操作需要，太少类的划分可能针对性不足，而太多太细的聚类则没有太大区分度。从可操作性和实践需要的角度，首先将贫困人口分为典型的 5 个、7 个和 10 个类别，然后根据 Calinski-Harabasz 指数[17] 对每一个 K-means 聚类分析的比较选取出统计上最理想的类别数。

由于不同年份的聚类结果会形成不可比较的类，所以这里选择用较新的数据(2010 年) 呈现聚类结果，并且选择人均纯收入 2 300 元①(2010 年不变价) 作为贫困标准。将 2010 年的数据对象按照 2 300 元的贫困标准对人均纯收入在该标准以下的贫困人口进行聚类，其中 5 个类型分类的结果最理想。

首先，对于所有类型来说存在共同的特征，也就是，对所有贫困类型的家庭而言，劳动力的受教育水平较低，家庭中没有外出务工人员或没有乡镇企业工人，这些是比例较高的共同特征。也就是这些贫困特征几乎是贫困家庭共同面对的限制性特征，但是导致这些特征的深层次的原因需要进一步分析。

第二，除了这些共同特征之外，不同类型存在两两之间有重要差异的属性。类型 1 与类型 3 都有突出的少数民族地区特征，但是类型 1 位于山区、土地有限，类型 3 位于非山区、土地资源约束相对小，我们不妨将类型 1 称为“山区少数民族”类，类型 3 为“非山区少数民族地区”类; 而类型 2 的典型特征是位于革命老区，伴有位于山区和土地资源有限的特征，不妨将它简称为“革命老区”类; 类型 4 的突出特征是土地资源有限，不妨简称为“有限土地”类; 而类型 5 的突出特征是位于山区，不妨简称为“山区”类。从表中可以看到，这 5 个类型的贫困家庭一方面具有突出特征，另一方面又具有交叉性的共同点。比如，类型 1 和类型 2 同时伴有类型 4 和类型 5 的特征，这表明少数民族地区和革命老区往往也是位于山区因而土地资源也相当有限的地区，但是类型 3 则表明那些不位于山区、土地资源受限较少的少数民族地区家庭也是比例较高的贫困群体。

第三，从整体分类来看，5 种类型的贫困人口中，类型 1、2、3 的贫困家庭比例较高，占了总贫困家庭比例的 75% 以上，尤其是“山区少数民族”类和“非山区少数民族”类共占了总贫困家庭的将近 60%。因此少数民族地区的贫困人口在农村贫困人口占了绝大多数，更值得关注，也需要在扶贫资源分配中重点考虑这些地区存在的特有的深层次限制条件，使得扶贫政策在这些地区符合少数民族地区人们的文化习俗、生计策略。

---

① 由于新贫困标准下的贫困农户都应被视为扶贫对象，所以选用新的贫困线进行聚类更具有现实意义。

(三)特殊类型贫困地区的特征结构

表 5 呈现了 2006 年与 2010 年少数民族地区、革命老区和边境县地区这三类地区的人口特征分布情况,可以清晰地看到老少边区贫困特征具有高度地相似性,以及几种明显的差别。首先,老少边区贫困特征的高度相似性在于除了地区类型本身之外,其他特征的占有比例在地区类别之间是相近的,也就是他们具有类似的劣势;第二,老少边区存在重叠,主要是少数民族地区和革命老区的重叠、少数民族地区和边境县地区的重叠①,比如以 2010 年为例,革命老区的被抽样调查人口有 21.5% 也是生活在少数民族地区的,边境地区的被抽样调查人口则有 85.0% 也是生活在少数民族地区,而这种地区特征的重叠不仅解释了三类贫困地区贫困特征的相似性,也会显示他们的差异和对聚类结果的影响;第三,与革命老区相比,少数民族地区和边境地区家庭中没有打工人员的比例更高,尤其是 2010 年,差异更加明显,边境地区和少数民族地区的高度重叠性可能表明了对于少数民族地区的家庭而言,语言、文化、生计策略和习俗的差异等会限制他们外出打工的意愿和机会。

表 4 2010 年农村贫困地区人口聚类(单位:%)

贫困聚类	高抚养比	少数民族地区	边境地区	革命老区	无打工人员	无乡企人员	山区	远离县城	有限土地	低教育水平	有不健康成员	占贫困人口比
1 类	32.2	69.0	14.1	0.0	63.2	99.1	98.7	91.1	75.4	84.1	10.1	31.5
2 类	31.8	25.5	0.0	100.0	50.8	98.9	86.0	85.4	76	76.2	13.8	15.9
3 类	31.9	51.9	11.0	2.4	86.7	99.0	0.0	85.0	23.2	78.3	10.0	28.2
4 类	32.6	16.0	2.3	7.9	52.5	96.5	28.9	11.4	94.7	73.2	11.7	11.0
5 类	31.9	12.7	5.5	3.4	75.6	98.2	79.2	46.2	0.0	77.6	9.2	13.4
所有贫困	32.1	42.3	8.4	15.4	66.2	98.5	65.3	69.1	56.4	79.2	10.6	100.0

注:以上结果是对 2010 年受调查户中的贫困家庭的聚类。采用的贫困标准是人均纯收入 2 300 元,共有 19 001 户家庭属于该标准下的贫困家庭。

表 5 2006 年与 2010 年特殊贫困地区人口特征分布比例(单位:%)

贫困特征	高抚养比	少数民族地区	边境地区	革命老区	无打工人员	无乡企人员	山区	远离县城	有限土地	低教育水平	有不健康成员	样本量(户)
少数民族地区 1	37.1	100	14.7	9.3	59.1	98.4	64.4	68.5	47.2	79.3	10.4	22 926
革命老区 1	29.1	22.1	0.0	100	58.5	98.4	63.1	68.1	45.6	79.4	10.3	9 585
边境地区 1	35.6	84.9	100	0.0	60.4	98.3	64.8	69.5	44.9	78.6	10.2	3 980
少数民族地区 2	30.8	100	15.0	9.0	61.6	99.1	72.5	72.9	42.2	77.8	8.9	22 680
革命老区 2	30.1	21.5	0.0	100	38.4	98.4	68.8	75.2	58.7	69.9	12.5	9 501
边境地区 2	31.5	85.0	100	0.0	75.4	99.5	66.5	77.0	28.5	76.6	11.3	4 000

注:1 代表 2006 年,2 代表 2010 年。

在 2010 年的贫困地区贫困人口聚类结果的基础上,给出了少数民族地区、革命老区、边境县地区三类特殊类型贫困地区对应的贫困人口聚类结构。可以看出,由于地区类型的重叠,老少边三类地区的聚类结构的特征更加突出。

在少数民族地区,1 类和 3 类贫困类别(“山区少数民族地区”类和“非山区少数民族地区”类)贫困人口共占了少数民族地区贫困人口的 81.0%;在革命老区,2 类(“革命老区”类)贫困人口占

① 革命老区和边境地区是没有重叠的。

了少数民族地区贫困人口的 85. 4%; 在边境县地区 1 类和 3 类贫困类别(“山区少数民族地区”类和“非山区少数民族地区”类) 贫困人口共占了少数民族地区贫困人口的 84. 3%。尤其是,在少数民族地区和边境县地区, “山区少数民族地区”类贫困人口都占了对应所在地区贫困人口的 60% 左右,并且往往与有限的土地、低教育水平联系在一起。

(四) 不同类型贫困人口的区域分布

考虑到国内扶贫过程中往往根据地域来划分扶贫范围,而扶贫政策的开展总是需要落实到地方,所以按照区域划分来呈现区域内不同类型贫困人口的结构,可以使扶贫政策在宏观设计时便更具有区域针对性。表 7 是不同类型农村贫困人口在不同区域的绝对分布比重,图 1 展示了不同区域内农村贫困人口不同类型的结构。两图共同表明了在不同区域不同类型贫困人口的量以及区域内的贫困类型结构特征。

表 6　2010 年老少边区特殊类型贫困地区贫困人口聚类(单位: %)

特殊类型贫困地区	贫困聚类	高抚养比	少数民族地区	边境地区	革命老区	无打工人员	山区	远离县城	有限土地	低教育水平	有不健康成员	人口占比
少数民族地区	1 类	31. 7	100	19. 7	0. 0	64. 4	98. 2	87. 0	64. 3	85. 9	9. 5	59. 3
	2 类	32. 0	100	0. 0	100	55. 7	99. 1	86. 8	70. 4	76. 5	13. 1	7. 9
	3 类	31. 0	100	16. 6	16. 1	83. 6	0. 0	71. 0	15. 5	76. 6	8. 0	21. 7
	4 类	33. 3	100	11. 8	84. 4	63. 9	11. 4	0. 0	100	71. 3	12. 9	5. 9
	5 类	29. 9	100	18. 7	2. 7	72. 1	100	0. 0	0. 0	83. 5	9. 7	5. 1
	所有贫困	31. 6	100	16. 9	84. 5	68. 3	71. 9	73. 9	53. 0	82. 2	9. 6	100
革命老区	2 类	31. 8	25. 5	0. 0	100	50. 8	86. 0	85. 4	76. 0	76. 2	13. 8	85. 4
	3 类	25. 6	34. 1	0. 0	100	89. 0	0. 0	92. 7	0. 0	78. 0	22. 0	2. 8
	4 类	32. 3	1. 7	0. 0	100	40. 0	0. 0	0. 0	91. 9	77. 4	12. 8	8. 1
	5 类	34. 5	10. 0	0. 0	100	100	49. 1	0. 0	0. 0	75. 5	19. 1	3. 8
	所有贫困	31. 8	23. 2	0. 0	100	52. 9	75. 3	75. 5	72. 2	76. 3	14. 2	100
边境县地区	1 类	32. 3	96. 6	100	0. 0	75. 6	98. 4	91. 5	54. 0	86. 8	8. 9	61. 0
	3 类	33. 4	77. 9	100	0. 0	77. 4	0. 0	73. 3	13. 2	73. 9	15. 9	23. 3
	4 类	27. 5	81. 2	100	0. 0	55. 1	5. 8	0. 0	92. 8	63. 8	21. 7	4. 3
	5 类	38. 5	43. 0	100	0. 0	94. 4	82. 7	35. 8	0. 0	73. 7	12. 3	11. 3
	所有贫困	33. 0	85. 5	100	0. 0	77. 2	69. 6	77. 0	40. 1	81. 3	11. 5	100

注: 以上结果是依据表 4 的聚类结果对少数民族地区、革命老区、边境县地区三类特殊类型贫困地区分别呈现贫困家庭人口聚类的结构,而不是分地区分别进行的聚类。

表 7　2010 年农村不同类型贫困人口按区域的绝对分布比重

贫困类别	显著特征	华北	东北	华东	中南	西南	西北	合计
1 类	山区少数民族	1. 3	0. 0	1. 5	7. 5	9. 8	4. 3	24. 5
2 类	革命老区	7. 8	2. 5	2. 0	3. 9	5. 0	4. 5	25. 7
3 类	非山区少数民族	5. 5	0. 9	0. 0	1. 9	9. 0	4. 5	21. 8
4 类	有限土地	3. 2	0. 6	0. 8	1. 5	1. 3	4. 2	11. 6
5 类	山区	4. 7	1. 5	1. 6	2. 6	3. 1	3. 0	16. 4
合计		22. 5	5. 5	5. 8	17. 4	28. 2	20. 5	100. 0

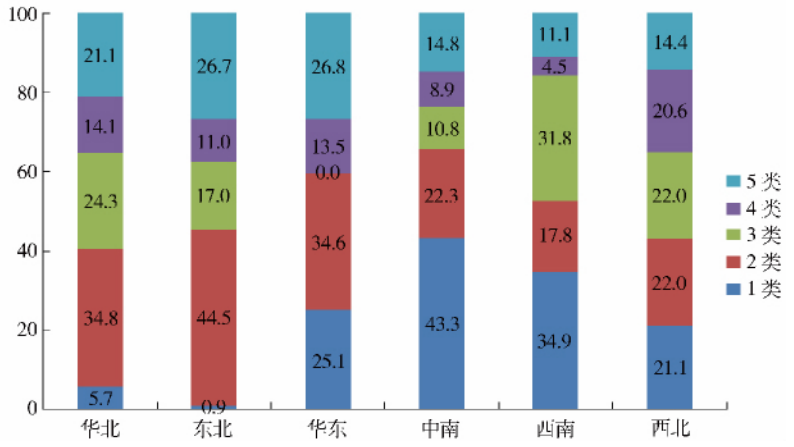

图 1 2010 年农村不同类型贫困人口区域内相对分布比重

注: 以上图表结果依据表 4 的聚类结果对不同区域测算贫困人口类型的结构,而不是分区域分别进行的聚类。样本地区中,华北地区包括河北、山西、内蒙,东北地区包括吉林、黑龙江,华东地区包括安徽、江西,中南地区包括河南、湖北、湖南、广西、海南,西南地区包括重庆、四川、贵州、云南,西北地区包括陕西、甘肃、青海、宁夏、新疆。从华北到西北不同地区的贫困户(人均纯收入 2 300 元以下) 样本量(户) 分别为 4 268、1 047、1 101、32 89、5 352、3 894。

首先,从表 7 可以看出,华北、西南和西北地区的贫困人口数量各占总量 20% 以上,三个地区的贫困人口总计占农村贫困人口的 70% 以上。这表明了从受益面来看,缓解贫困、减少贫困人口的主战场是这些地区。在区域间的扶贫资源分配可能需要考虑贫困人口比重的分布。

第二,从图 1 可以看出,不同区域农村贫困人口类型的结构具有明显差别。从区域内贫困人口贫困类型的结构看,华北地区以革命老区、非山区的少数民族、山区这几类贫困类型为主,东北地区以革命老区、山区为主,华东地区以山区少数民族、革命老区和山区为主,中南地区以山区少数民族、革命老区类为主,西南地区以少数民族类型包括山区的少数民族和非山区的少数民族为主,西北地区则 5 种贫困类型比较均匀。但在所有地区内部,少数民族类型的贫困人口都占有相对高的比重。以上这些贫困结构显示了在不同区域之内,哪些是更为主要的贫困类型。

(五) 连片特困地区的贫困人口类型与分布

《中国农村扶贫开发纲要(2011—2020 年) 》明确了扶贫攻坚的主战场,将六盘山区、秦巴山区、武陵山区、乌蒙山区、滇桂黔石漠化区、滇西边境山区、大兴安岭南麓山区、燕山-太行山区、吕梁山区、大别山区、罗霄山区等区域的连片特困地区和已经明确实施特殊政策的西藏、四川藏区、新疆南疆三地州确立为扶贫攻坚主战场。除了西藏、四川藏区、新疆南疆三地州之外,11 个连片特困地区都在农村贫困监测调查范围之内。在已有聚类分析的基础上,对各连片特困地区的贫困人口进行贫困类型识别有助于了解连片特困地区的主要限制特征,为连片特困地区的扶贫开发提供一定参考依据。

表 8 呈现的是 2010 年连片特困地区的基本贫困概况。根据 2010 年贫困监测的样本数据结果可以发现,根据两种贫困标准,从贫困发生比率看,大兴安岭南麓山区、吕梁山区、燕山-太行山区的贫困发生率高出总样本平均水平许多; 从贫困人口总量来看,燕山-太行山区、秦巴山区、六盘山区、滇黔桂石漠化区、滇西边境山区、武陵山区的贫困人口数量较大。尤其是燕山-太行山区,贫困发生比率和贫困人口数量都在 11 个片区位于前列。

根据前文表 4 的总体聚类结果,将 11 个连片特困地区中 5 个贫困类型的结构比例列出在表 9 中。从表 9 中可以看出,11 个连片特困地区有各自不同的主要贫困类型,滇黔桂石漠化区、滇西边

境山区、乌蒙山区以“山区少数民族”类型的贫困为主导,大别山区、罗霄山区、武陵山区以“革命老区”类型的贫困为主导,大兴安岭南麓山区则以“非山区少数民族”类型的贫困为主导,六盘山区、吕梁山区以“山区”型贫困类型为主导,而秦巴山区、燕山-太行山区没有一个绝对主导的贫困类型,属于多种贫困类型区域。

由此可见,作为新纲要中确定的扶贫攻坚主战场的 11 个连片特困地区,其主要的贫困类型各不相同,当然,这里的基础聚类指标中没有考虑特殊的自然环境因素,在片区扶贫开发过程中,实际上还需要将贫困类型与当地特殊的自然环境因素结合考虑。

表 8　2010 年连片特困地区的贫困概况

连片特困地区	旧标准贫困线			新标准贫困线			不同片区样本量比例
	贫困发生率	占总贫困人口比例	占总样本人口比例	贫困率	占总贫困人口比例	占总样本人口比例	
大别山区	6.9	5.0	0.6	23.2	5.5	2.0	8.7
大兴安岭南麓山区	26.7	5.6	0.7	56.7	3.9	1.4	2.5
滇黔桂石漠化区	8.8	11.8	1.4	31.8	14.1	5.2	16.2
滇西边境山区	9.7	10.0	1.2	36.4	12.4	4.5	12.4
六盘山区	14.7	12.7	1.5	38.6	11.0	4.0	10.4
吕梁山区	23.0	8.3	1.0	54.1	6.4	2.3	4.3
罗霄山区	8.7	3.2	0.4	32.9	4.0	1.5	4.5
秦巴山区	11.3	13.4	1.6	34.6	13.5	4.9	14.3
乌蒙山区	9.0	6.4	0.8	34.3	8.1	2.9	8.6
武陵山区	10.4	10.0	1.2	35.5	11.3	4.1	11.6
燕山-太行山区	25.5	13.7	1.7	54.7	9.7	3.5	6.5
合计	12.1	100.0	12.1	36.5	100.0	36.5	100.0

注: 11 个连片特困地区的总样本为 35 511 户。

表 9　2010 年连片特困地区的贫困类型结构

贫困类型	大别山区	大兴安岭南麓山区	滇黔桂石漠化区	滇西边境山区	六盘山区	吕梁山区	罗霄山区	秦巴山区	乌蒙山区	武陵山区	燕山-太行山区
1	1.4	1.8	83.0	74.7	32.2	3.8	9.8	33.3	68.5	33.2	26.4
2	42.7	0.0	4.1	0.0	11.8	17.5	66.0	28.0	0.0	43.5	0.0
3	16.6	77.8	4.6	9.3	4.0	8.9	5.4	4.0	2.8	5.1	27.6
4	33.4	1.6	2.8	6.6	11.4	13.1	16.9	16.4	13.6	12.0	19.5
5	5.9	18.8	5.6	9.4	40.6	56.7	1.9	18.3	15.2	6.2	26.5
合计	100	100	100	100	100	100	100	100	100	100	100

四、基本结论与建议

首先,贫困类型的分布,既呈现了扶贫对象在区域间分布的不平衡性,也表明了区域内部扶贫需要提高针对性。各种贫困类型的不同特点和区域分布上的差异从一个视角揭示了收入差距特别是贫困程度差异化的来源。除了受教育水平低、没有乡镇企业务工机会和家庭缺乏外出打工成员

这些共同特征之外,不同的区域内不同类型的贫困类型组合结构,意味着各个地区贫困形成的深层次原因可能各不相同。因此,扶贫政策可能需要针对不同地区的贫困类型组合对不同地区的贫困人口所面临的共同障碍和特殊劣势做出合理的调整。

第二,少数民族地区的贫困特征突出,尤其以西南、西北、华北为多数。老少边区特殊类型贫困地区中,尤其以少数民族地区、以及少数民族和革命老区或者边境地区的重叠地区的贫困人口比重大。少数民族地区的贫困特征方面,除了一些共性,还表现在外出务工比例低和受教育水平低。从目前的扶贫实践体系来看,主流的一些扶贫政策,例如异地扶贫搬迁、整村推进、以工代赈、就业促进等方式未必能在少数民族地区奏效,尚需对少数民族地区贫困的形成机制和脱贫方式开展更加深入的研究,以便提出针对少数民族地区的因地制宜的扶贫开发措施。

第三,对于一些贫困人群共有的脱贫障碍,比如教育水平低下、外出务工成员少、没有乡镇企业工作的机会,即使他们在特征表现上相同,但是这些特征本身的形成也具有不相同的原因。而另一些障碍也与相应的区域环境有密切的相关,比如少数民族人口比较多的区域,受教育水平和外出务工的比例都更低,山区的人口因为离市场远、土地更加有限(以及可能更恶劣的生存环境) 而更容易陷入贫困。家庭人口特征和地区的环境因素可能同时对贫困的发生造成影响。结合地区特征来实施提高农村贫困人口素质和能力的措施可能更加有效。

第四,贫困类型在区域内分布的不同结构,为不同区域的扶贫政策制定和资源分配提供了参考。比如华北地区的贫困人口以革命老区(34.8%) 和山区(21.1%) 、非山区的少数民族(24.3%) 为主,其中,革命老区和山区这两类都具有山区地域的限制特征且外出务工的比例又低,对于这些地域的贫困人口,不仅需要提高土地的产出水平,更需要提高当地贫困人群的生计能力,通过提高农业生产技术、外出务工能力来提高他们的就业能力,而对于该地区的非山区的少数民族贫困人口,或许提高外出务工能力并不是合适的政策措施,而主要应当根据他们的生计策略和文化习俗,提高特色经济的发展能力,比如发展乡土教育、特产经济、文化旅游等。

第五,11 个连片特困地区的主导贫困类型各不相同,而贫困类型之间又有一些相同特征,这意味着片区扶贫开发需要具有片区针对性的扶贫政策,并需要结合当地特殊的自然环境因素综合考虑,同时,主导贫困类型相同的片区之间的扶贫开发策略似乎在一定程度上可以互相借鉴。

总而言之,尽管聚类分析只是一种探索性分析,但是农村贫困人口的聚类仍然为我们定义各种贫困的类型、以及它们在不同区域间、特定区域划分、片区间的分布结构提供了非常有价值的信息。同时,这种探索性分析也将为进一步的统计推断分析提供基础。

[参考文献]

[1] MacQueen J. *Some methods for classification and analysis of multivariate observations: Proc. 5th Berkeley Symp. Mathematical Statist*. Probability, 1967

[2] 朱建平, 陈民恳. 面板数据的聚类分析及其应用. 统计研究, 2007(4):11 - 14

[3] Tukey J. Exploratory Data Analysis. Addison-Wesley, 1977

[4] 李因果, 何晓群. 面板数据聚类方法及应用. 统计研究, 2010(9):73 - 79

[5] Jain A K, Murty M N, Flynn P J. Data clustering: a review. *ACM Comput. Surv.*, 1999,31(3):264 - 323

[6] Jain A K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010,31(8):651 - 666

[7] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 1998,2(3):283 - 304

[8] Huang Z, Ng M K. A fuzzy k - modes algorithm for clustering categorical data. *Fuzzy Systems, IEEE Transactions on*, 1999,7(4):446 - 452

[9] Huang Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.: Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97), 1997
[10] Chaturvedi A, Green P E, Caroll J D. K-modes clustering. *Journal of Classification*, 2001, 18(1): 35 - 55
[11] Ding C, He X. K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization: Proceedings of the 2004 ACM symposium on Applied computing, 2004
[12] Chuang K, Tzeng H, Chen S, et al. Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 2006, 30(1): 9 - 15
[13] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究. 软件学报, 2008(1): 48 - 61
[14] Khan S S, Ahmad A. Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters*, 2004, 25(11): 1293 - 1302
[15] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1979, 28(1): 100 - 108
[16] Hartigan J A. Clustering algorithms. John Wiley & Sons, Inc., 1975
[17] Caliński T, Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*, 1974, 3(1): 1 - 27
[18] Mundial B. From poor areas to poor people: China's evolving poverty reduction agenda. An assessment of poverty and inequality in China. Washington DC: World Bank. Poverty Reduction and Economic Management Department East Asia and Pacific Region, 2009

Clustering Analysis of the Rural Poverty Population and Poverty Reduction Strategies

Wang Yu　Wang Sangui

Abstract　Using the method of K-means clustering, this paper makes the classification poverty population in rural China and thus the analysis of structure of poverty types in areas of special types of poverty and in contiguous poverty areas. The outcomes show that the targeted poor are disproportionately distributed among regions and the features of different types and their regional distribution can be treated as sources of income inequality especially the poverty levels. In particular, poverty characteristics are more notable and the poverty is lager in population in ethnic minority areas than those in old revolutionary base areas and border regions, which implicates that further research is required to explore the hiding mechanism causing poverty in ethnic minority areas so as to put forward poverty alleviation and development measures accommodating to local condition. Also, the leading poverty type is different among contiguous poverty-stricken areas, so that targeted policies are needed. Though clustering is mainly deemed as exploratory analysis, the clustering of rural poverty population still helps to make classifications and definitions of various types of poverty and the internal structure and regional distribution of these poverty types, which can contribute to further statistical inferences and causal analysis.

Key words　Rural poverty population; K-means clustering; Special types of poverty; Contiguous poverty-stricken areas; Geographical distribution

(责任编辑: 陈世栋)