Statistics and Application 统计学与应用, 2021, 10(4), 650-656
Published Online August 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.104066

Hans汉斯

基于主成分聚类分析的不同地区贫困问题探究

赵未名

曲阜师范大学，山东 曲阜
Email: 1026438091@qq.com

收稿日期：2021年7月13日；录用日期：2021年8月6日；发布日期：2021年8月13日

摘　要

2020年是我国脱贫攻坚的收官之年。在我国脱贫攻坚战取得了全面胜利后，脱贫与返贫交叉发生是我国目前面临的主要问题。因此，深入分析造成贫困的影响因素、对不同地区根据贫困程度进行聚类，从而有针对性找到脱贫方法是十分必要的。本文选取经济、人口结构、教育、医疗、文化、交通六个维度十一个指标，利用主成分分析进行降维处理，根据累积贡献率选取前三个主成分。随后进行系统聚类，利用Ward聚类方法将31个地区按贫困成因分为四类，并根据分类提出有效的建议。

关键词

主成分分析，系统聚类，贫困成因

Research on Poverty in Different Areas Based on Principal Component and Cluster Analysis

Weiming Zhao

Qufu Normal University, Qufu Shandong
Email: 1026438091@qq.com

Received: Jul. 13th, 2021; accepted: Aug. 6th, 2021; published: Aug. 13th, 2021

Abstract

The year 2020 will be the end of China's fight against poverty. After the victory of the battle against poverty in China, the intersection of poverty alleviation and return to poverty is the main problem faced by China. Therefore, it is very necessary to analyze the influencing factors of po-

文章引用: 赵未名. 基于主成分聚类分析的不同地区贫困问题探究[J]. 统计学与应用, 2021, 10(4): 650-656.
DOI: 10.12677/sa.2021.104066

verty and cluster different regions according to the poverty degree, so as to find targeted poverty alleviation methods. This paper selects eleven indicators from six dimensions, including economy, population structure, education, medical care, culture and transportation, and uses principal component analysis for dimensionality reduction. The first three principal components are selected according to the cumulative contribution rate. Then, Ward clustering method was used to divide the 31 regions into four categories according to the causes of poverty, and effective suggestions were put forward according to the classification.

Keywords

Principal Component Analysis System, Cluster, Poverty Causes

1. 引言

随着扶贫进程的深入推进，全国各贫困地区不断脱贫，但扶贫政策的效果逐渐减弱，有关贫困问题研究开始受到社会的广泛关注。刘玉杰在多维贫困视角下进行了空间聚类与相关性研究[1]；马绍东、万泽仁利用多维贫困分析方法“A-F (双重临界值法)”，在多维贫困视角下探究了民族地区返贫成因[2]；张昭等利用 AF 方法测算了农村老年人的多维贫困，并进一步通过贫困分解的方式考察了人口老龄化对农村多维贫困的影响[3]。本文在多维贫困视角下，利用主成分分析、系统聚类方法对全国各省份进行了有效的分类，以探究各地区贫困成因的不同。

2. 材料与方法

2.1. 多维贫困指标选择

经济学家阿玛蒂亚·森提出多维贫困理论，他认为贫困包括收入、健康、教育、住房及公共物品的可获得性等多个维度的缺失[1]。基于多维贫困理论，本文在已有文献的基础上进行补充，最终选取人口结构、经济水平、教育水平、医疗、文化以及交通运输六个维度，并进一步选取多个二级指标作为衡量贫困问题的指标，如表 1 所示。

1) 人口结构和经济水平

经济水平是最传统的衡量贫困问题的指标。经济水平直接反应人们的生活水平，影响着一个地区贫困现象的发生。同时，已有文献资料显示，人口结构对经济的增长具有显著影响[4]，进而对各地区贫困程度的影响也是不容忽视的问题。在人口结构和经济水平两个维度下，本文选取平均家庭户规模、总抚养比、居民人均可支配收入、公共预算收入、人均居民消费支出以及货物进出口总额作为二级指标。其中，总抚养比指人口总体中非劳动年龄人口数与劳动年龄人口数之比。

2) 教育和文化

一个地区的教育和文化水平直接反映了该地区居民的精神文化生活。在教育和文化两个维度下，本文选取了文盲人数占 15 岁以上人口比重、文化制造业企业数、广播节目综合人口覆盖率、文化制造业企业数作为二级指标。

3) 医疗和交通

一个地区的医疗和交通条件能够反映出该地区居民的生活便利度。一个医疗条件好、交通发达的地区往往经济水平更高，对一个地区的贫困程度也有着不可忽视的影响。在医疗和交通两个维度下，本文选取诊疗人次数和客运量作为二级指标。

Table 1. Selection of poverty index
表 1. 贫困程度指标选择

一级指标	二级指标	自变量符号表示
人口结构	平均家庭户规模	X_1
	总抚养比	X_2
经济水平	居民人均可支配收入	X_3
	公共预算收入	X_4
	人均居民消费支出	X_5
	货物进出口总额	X_6
教育	文盲人数占 15 岁以上人口比重	X_7
医疗	诊疗人次数	X_8
文化	文化制造业企业数	X_9
	广播节目综合人口覆盖率	X_{10}
交通	客运量	X_{11}

2.2. 数据来源

本文数据来源于 2020 年《中国统计年鉴》。

2.3. 统计分析—主成分分析和系统聚类

由于本文选取的各个指标单位不同，为避免单位量纲的影响，首先利用统计软件 R 对原始数据进行标准化处理，对每个数据用以下公式进行处理：

$$X'_{ij} = \left(X_{ij} - X_j\right) \Big/ S_j \tag{1}$$

其中，X'_{ij} 为标准化后的数据，X_{ij} 为原始数据，X_j 为第 j 个指标的均值，S_j 为第 j 个指标的样本方差。

随后，我们从相关矩阵出发对标准化后的数据进行主成分分析。根据累计贡献率选取少数几个主成分，在保留原始数据大部分信息的基础上，又很好的对原始数据进行了降维处理。

最后，利用降维后的数据进行 ward 聚类。本文选取欧式距离作为衡量样品之间的距离，通过树形图以及实际情况，最终确定分类个数，得到分类结果。

3. 结果与分析

3.1. 主成分分析降维结果

由于刚开始选入的十一个变量存在较高的相关性，观测数据中的信息在一定程度上有所重叠，因此本文利用主成分分析方法对十一个变量进行降维，从而使问题的分析得以简化。

利用 R 软件对数据进行标准化后，对十个变量进行主成分分析，分析结果如表 2 所示。

Table 2. Results of principal component analysis
表 2. 主成分分析结果

主成分因子	标准差	因子贡献率	累计因子贡献率
Comp.1	2.344370	0.4996449	0.4996449
Comp.2	1.5993376	0.2325346	0.7321795
Comp.3	1.03094041	0.09662165	0.82880116
Comp.4	0.83805119	0.06384816	0.89264933
Comp.5	0.69925256	0.04445038	0.93709970
Comp.6	0.54443113	0.02694593	0.96404564
Comp.7	0.44296471	0.01783798	0.98188361
Comp.8	0.35060837	0.01117511	0.9930873
Comp.9	0.23438662	0.00499428	0.999247968
Comp.10	0.114649860	0.001194963	0.99247968
Comp.11	0.0909524465	0.0007520316	1.00000000

由表 2 可以看出，前三个主成分的特征值均大于 1，且前三个主成分的累计贡献率达到了 82.88%，可以反映原始变量的大部分信息。同时，由陡坡图(见图 1)也可以得到，选取三个主成分有较好的结果。

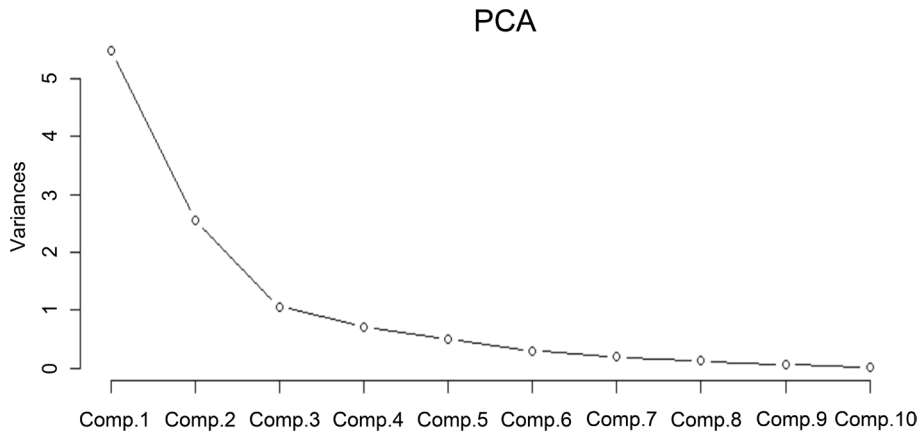

Figure 1. Steep slope map of principal component analysis
图 1. 主成分分析陡坡图

进一步，得到前三个主成分分别为：

$$y_1 = 0.257x_1 + 0.165x_2 - 0.329x_3 - 0.399x_4 - 0.333x_5 - 0.388x_6 + 0.2x_7 - 0.32x_8 - 0.329x_9 - 0.247x_{10} - 0.257x_{11}$$

$$y_2 = 0.323x_1 + 0.449x_2 - 0.272x_3 + 0.176x_4 - 0.287x_5 + 0.115x_6 + 0.208x_7 + 0.360x_8 + 0.325x_9 - 0.232x_{10} + 0.401x_{11}$$

$$y_3 = 0.4x_1 + 0.28x_3 + 0.259x_5 + 0.249x_6 + 0.722x_7 - 0.115x_8 - 0.281x_{11}$$

通过分析载荷，由于 y_1 在 x_1, x_2, x_7 上的载荷为正，其余为为负，则第一主成分可以看作度量人口结构和教育对贫困程度的正向影响变量。由于 y_2 在 x_3, x_5 上的载荷为负，因此可将第二主成分看作消费对地区贫困程度的负向影响变量。在第三主成分中，x_7 上的载荷最大，故我们将 y_3 看作教育对地区贫困程度的影响变量。

利用 R 软件分析得到各地区主成分得分如图 2 所示：

	Comp.1	Comp.2	Comp.3
北　京	-3.588	-3.031	1.117
天　津	-0.799	-2.984	0.195
河　北	0.260	0.646	-0.613
山　西	1.218	-0.996	-0.687
内蒙古	0.593	-2.204	-0.351
辽　宁	-0.330	-1.182	-1.092
吉　林	0.852	-1.670	-1.001
黑龙江	0.733	-1.884	-1.065
上　海	-3.782	-3.047	1.564
江　苏	-4.010	1.820	0.522
浙　江	-4.174	0.570	0.279
安　徽	0.118	0.712	-0.290
福　建	-0.824	-0.551	0.501
江　西	0.857	0.697	-0.369
山　东	-1.453	1.788	-0.141
河　南	-0.256	2.284	-0.787
湖　北	-0.460	0.316	-0.512
湖　南	-0.118	1.391	-0.757
广　东	-6.788	2.884	0.690
广　西	1.592	0.827	-0.605
海　南	2.010	-0.609	0.470
重　庆	0.560	-0.099	-0.724
四　川	-0.143	1.553	-0.872
贵　州	2.740	2.450	-0.488
云　南	1.393	0.165	0.088
西　藏	4.537	1.746	4.326
陕　西	0.681	-0.213	-0.704
甘　肃	2.233	0.224	0.441
青　海	2.303	-0.860	0.647
宁　夏	1.983	-0.897	0.407
新　疆	2.063	0.155	-0.187

Figure 2. Scores of the first three principal components
图 2. 前三个主成分得分

3.2. Ward 聚类结果分析

Ward 聚类法又称离差平方和法，是一种常见的系统聚类方法。在提取完主成分以后，对选择的前三个主成分进行系统聚类分析。在 Ward 聚类中，定义 G_K 和 G_L 之间的平方距离为 $D_{KL}^2 = W_M - W_K - W_L$，其中 W_M, W_K, W_L 分别为 G_M, G_K, G_L 的离差平方和，每一步合并使离差平方和增量达到最小的两个类。主要的步骤是利用上述定义的类间距离测算 31 个地区的类间距离，同时生成距离矩阵，选择使离差平方和增量达到最小的两个类进行合并。重复以上步骤，最终所有 31 个地区将合并为一个大类。聚类结果如图 3 所示。

从统计角度来看，理想的聚类结果应该是：类的个数适当，类之间较分开而类内相近。如果在(15, 25)内切一刀，则分为两类；如果在 11 附近切一刀，则分为四类。从聚类的实际意义出发，分四类似乎更加符合实际情况。北京、上海、广东、江苏、浙江由于经济发展水平高，几乎无贫困现象发生，可将其定义为无贫困地区；河北、河南、山东等地由于人口较多，劳动力充足，属于轻度贫困地区；内蒙古、吉林、黑龙江等地由于地理位置偏远，属于中度贫困地区；青海、新疆、宁夏、甘肃等地区为少数民族聚集地，经济发展较为落后，生活便利程度低，属于深度贫困地区。

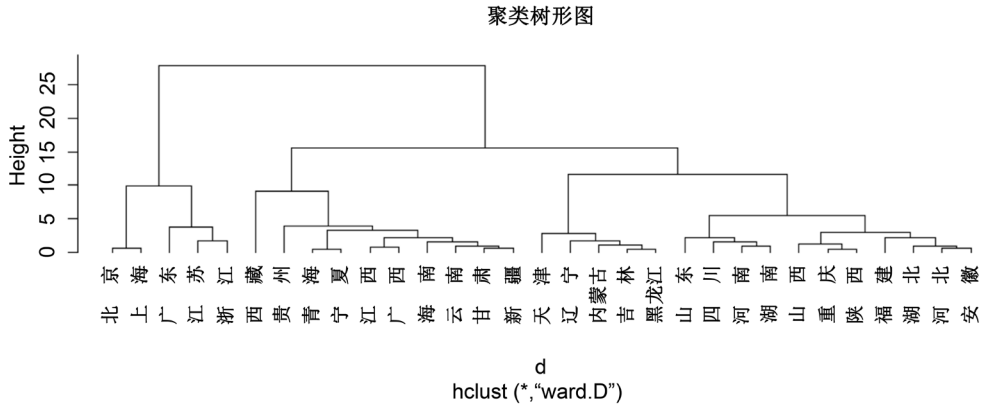

Figure 3. Tree diagram
图 3. 树形图

4. 讨论

4.1. 结论分析

1) 经济发展水平、教育、医疗水平、人口结构、交通发达程度均对各地区贫困程度有影响。通过主成分分析的载荷可知，不同主成分分别代表了不同维度指标对各地区贫困程度的影响。由主成分得分，我们可以得到各地区的贫困影响因素，从而有针对的提出建议。

2) 根据贫困程度不同，本文将 31 个地区共划分为四类。第一类：北京、上海、广东、江苏、浙江。第二类：福建、河北、河南、山东、重庆、安徽、陕西、山西湖北、湖南、四川。第三类：内蒙古、吉林、黑龙江、辽宁、天津。第四类：西藏、贵州、青海、宁夏、江西、广西、海南、云南、甘肃、新疆。

4.2. 建议

1) 政府应当通过生育政策的制定来控制当地的人口结构。由本文分析可知，人口结构是影响各地区贫困的重要原因，社会抚养比过大、平均家庭户规模过大都是造成贫困的重要原因。因此，制定一个适合的生育政策对于保持脱贫成果是十分必要的。

2) 根据分类结果，政府应当分区域采取不同的扶贫措施，而不是对所有地区采取相同的扶贫政策。基于本文聚类结果，西藏、贵州、青海等地划分为一类，说明少数民族地区仍然是我国扶贫的重点关注地区，政府应当充分分析当地实际情况，保证该地区脱贫不返贫，有效维护好脱贫的成果。

参考文献

[1] 刘玉杰. 多维贫困的空间聚类与相关性研究[J]. 云南农业大学学报(社会科学), 2021, 15(3): 48-54.

[2] 马绍东, 万仁泽. 多维贫困视角下民族地区返贫成因及对策研究[J]. 贵州民族研究, 2018, 39(11): 45-50.

[3] 张昭, 杨澄宇. 老龄化与农村老年人口多维贫困——基于 AF 方法的贫困测度与分解[J]. 人口与发展, 2020, 26(1): 12-24+11.

[4] 李豫新, 陈琨. 民族地区人口结构的经济增长效应分析[J]. 北方民族大学学报(哲学社会科学版), 2019(5): 138-145.

附录：R 程序代码

```
data <- read.table("D:/data/data3.csv", header = T, sep = ",") #读入数据
data1 <- scale(data[, -1])
#----------------------主成分分析----------------------#
round(cor(data1), 3) #计算相关矩阵，保留 3 位小数
PCA<-princomp(data1, cor=T) #从相关矩阵出发进行主成分分析
PCA
summary(PCA, loadings=T) #列出主成分分析的结果
screeplot(PCA, type="lines") #陡坡图，用直线图类型
scores<-round(PCA$scores, 3) #主成分得分，保留 3 位小数
scores<-cbind(data[, 1], scores[, c(1, 2, 3)]) #将地区名与前 3 个主成分得分合并
scores
#-----------------------ward 聚类-----------------------#
d <- dist(data1, diag = T)
#离差平方和法#
hc<- hclust(d, "ward.D")
cbind(hc$merge, round(hc$height, 2))
plot(hc, hang=-1, labels = data[, 1]) #聚类分析树形图
```