

· 技术方法 ·

基于 SPSS 的农业功能聚类分区方法^{*}

吕敬堂¹, 吕大明², 张 浩²

(1. 贵州大学机械工程学院, 贵阳 550003; 2. 贵州省农业资源区划研究中心, 贵阳 550003)

摘 要 以贵州农业系统为研究对象, 应用聚类分析方法, 通过建立农业功能分区指标体系、数据标准化处理、层次聚类和 K-均值聚类分析, 将全省 88 个县分为 4 个区域; 通过计算分区指标平均值, 应用系统排序方法进行区域综合功能排序和各区域的功能单排序, 确定各个区域的农业主导功能和辅助功能, 指出各农业区域的功能拓展方向, 为分区命名提供依据。

关键词 贵州省 农业功能 聚类分析 分区方法

1 研究的背景

综合分析农业自然资源状况、发展基础和潜力、政治和社会等因素, 科学划分农业功能区, 客观评价不同区域的功能特点, 确定区域农业的主导功能及其发展方向, 对深化农业结构战略性调整, 形成布局合理、特色鲜明的现代农业空间格局具有重要意义^[1-4]。

2007 年, 全国农业资源区划办公室在四川、福建、浙江等 7 个省(市、区)开展了农业功能区划试点研究。在此基础上, 农业部办公厅于 2008 年 5 月 7 日下达了《农业部办公厅关于开展农业功能区划工作的通知》(农办计[2008]34 号), 确定了 2008 年全国开展农业功能区划研究省份名单, 共 23 个省。2009 年 4 月 9 日, 农业部计划发展司组织在贵阳召开全国农业功能区划研讨会, 总结交流研究经验。总体上看, 全国各省(市、区)的农业功能区划研究取得了较好成绩, 但在分区方法应用上还存在一些问题和困难: 一是聚类分析所得区域不连片; 二是各区域功能结构不明晰, 分区命名困难。

在贵州省农业功能区划研究过程中, 采用 SPSS13.0 统计分析软件和 Excel2003 作为分析工具, 应用聚类分析和系统排序相结合的方法, 将贵州省 88 个县(市、区)分成 4 个农业主导功能明确、辅助功能结构清晰的区域, 分区结论符合贵州实际, 较理想。该文对分区方法进行总结。

2 农业功能分区的工作程序

农业的基本功能可以分为四大类: 即农产品供给功能、就业与生活保障功能、生态调节与生态约束功能、文化传承与休闲功能^[3-8]。功能分区的思路: 首先根据农业综合功能描述指标把全省划分成几个区域; 然后分别按照农业的四大功能对所分区域进行单排序, 以确定各区域就某一功能而言在全省所处的位置; 最后按各区域对农业的四大功能进行总排序, 以确定各区域的主导功能和辅助功能, 形成不同类型的农业功能区。

分区程序包括: 分区指标确定→数据预处理→构造关系矩阵(描述样本间的亲疏关系)→聚类分析→定性分析调整→确定分区类别(分区数)→功能排序等几个主要步骤, 如图 1 所示。

整个分区工作可分为 3 个阶段: 一是分区指标形成, 主要工作内容包括收集、统计、整理、分析数据资料, 确定分区的定量指标集, 为聚类分析做准备。二是聚类分析, 采用层次聚类法和 K-均值聚类法进

收稿日期: 2009-06-30 吕敬堂为副教授 吕大明为高级工程师 张浩为科长

^{*} 基金项目: 贵州省农业功能区划项目资助(农业部办公厅农办计[2008]34号)

行分类、判断和调整，将全省划分为几个农业综合功能差异明显的区域，为农业功能区的确定奠定基础。三是确定农业功能区，采用权重系数法对农业功能作排序分析，明确各区域的主导功能及农业功能区类型。

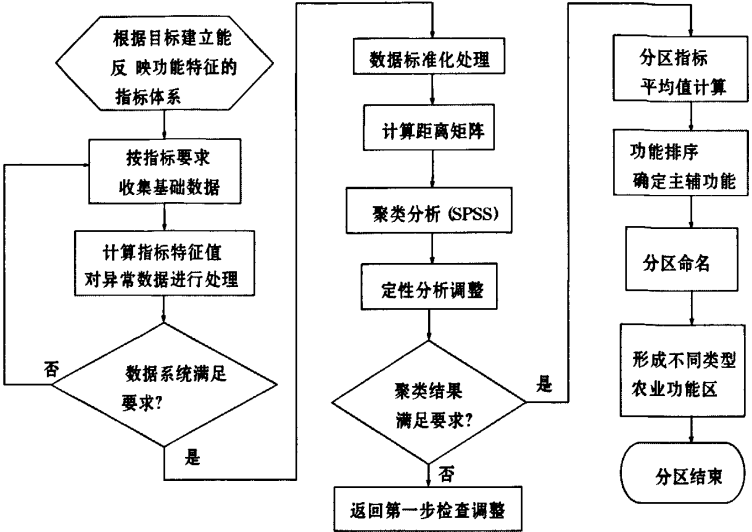


图 1 农业功能分区工作程序

3 建立功能分区指标体系

要做聚类分析，首先必须按照聚类的目的，从对象中提取能表达这个目的的特征指标，然后根据对象的亲疏程度进行分类。贵州省农业功能分区的对象是全省 88 个县（市、区），特征指标就是每个县（市、区）都具有的、能够反映农业功能特征的统计指标。各项指标的具体含义和计算方法由于受篇幅所限，在此从略。

根据贵州省农业功能区划指标体系的构建思路和设置原则，对省有关业务部门提供的统计资料进行反复研究，并多次召开专家座谈会进行咨询和论证，最终选择 4 类 10 组共 45 个具体指标作为贵州省农业功能分区的依据，力求使所构建的指标体系更加符合贵州实际，能够综合、全面地描述贵州农业功能的特点。

3.1 农产品供给功能指标

参考农业部农业综合功能区划要求，用资源禀赋、规模与结构等两组指标描述农业的供给功能，共 30 个变量。其中，资源禀赋指标组 13 个 ($X_1 \sim X_{13}$)，规模与结构指标组 17 个 ($X_{14} \sim X_{30}$)。

3.2 就业和生活保障功能指标

就业和生活保障功能指标，主要反映农业承担就业和生活保障的份额、压力和水平，受农村富余劳动力数量、城镇化水平、非农产业发展水平等因素的影响。用份额、压力、水平等 3 组指标描述农业的就业和生活保障功能，共 6 个变量。其中，份额指标组 3 个 ($X_{31} \sim X_{33}$)，压力指标组 1 个 (X_{34})，水平指标组两个 ($X_{35} \sim X_{36}$)。

3.3 生态调节功能与生态约束指标

该类指标主要反映和评价农业生态调节功能的类型，重要程度与区域生态脆弱性。用农田生态、草地生态、森林生态、生态约束等 4 组指标描述农业的生态功能，共 7 个变量 ($X_{37} \sim X_{43}$)。

3.4 文化传承和休闲功能指标

用文化传承指标组和居民休闲指标组描述农业的文化传承和休闲功能。贵州是一个多民族省份，是典

型的喀斯特岩溶山区,地形地貌复杂,民族文化多样,存在大量地域特色鲜明的农耕文化与生存方式。但由于缺乏系统、完整的统计数据,指标难以量化,该部分内容放在分区论述中定性描述。

农业休闲功能的发挥受到自然生态环境、观光农业发展规模、区域城镇人口数量、交通条件、城乡经济发展状况等诸多因素的制约。一般来说,生态环境好、交通便利、城镇人口多、经济发达的区域,农业的休闲功能就强;反之,则弱。根据调研和资料统计情况,选择两个指标来描述农业的休闲功能,即城镇人口数和县域单位土地面积公路里程数($X_{44} \sim X_{45}$)。

将 88 个县(市、区)的 45 个指标收集、整理、计算出来,就得到一个 88×45 的样本矩阵。

4 功能区划聚类分析

农业功能区划分指标确定以后,即可进入聚类分析阶段,其过程主要分为数据标准化、构造关系矩阵、聚类分析^[9-10]和确定分类数等 4 个步骤。

4.1 数据标准化处理

4.1.1 数据标准化处理的目的

指标变量大多数都有量纲,且各变量的量纲不同或数量级相差较大,没有可比性。为了能把这些数据放到一起加以比较,必须进行数据的标准化处理,消除量纲的影响。

4.1.2 数据标准化方法

贵州省农业功能区划聚类分析采用的样本数据可表示为:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & \cdots & X_{1,45} \\ X_{21} & X_{22} & \cdots & \cdots & X_{2,45} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{88,1} & X_{88,2} & \cdots & \cdots & X_{88,45} \end{bmatrix} \quad (1)$$

$$\text{均值为: } \bar{X}_j = \frac{1}{88} \sum_{i=1}^{88} X_{ij} \quad (2)$$

$$\text{标准差: } S_j = \sqrt{\frac{1}{88-1} \sum_{i=1}^{88} (X_{ij} - \bar{X}_j)^2} \quad (3)$$

$$\text{极差为: } R_j = \max_{1 \leq i \leq 88} \{X_{ij}\} - \min_{1 \leq i \leq 88} \{X_{ij}\} \quad (4)$$

公式中, $i = 1, 2, \dots, 88$; $j = 1, 2, \dots, 45$ 。

假设原数据 X_{ij} 标准化后的值为 X'_{ij} , 则可用下列方法进行标准化处理:

(1) 标准差标准化法

$$X'_{ij} = \begin{cases} \frac{X_{ij} - \bar{X}_j}{S_j}, & \text{若 } S_j \neq 0 \\ 0, & \text{若 } S_j = 0 \end{cases} \quad (i = 1, 2, \dots, 88; j = 1, 2, \dots, 45) \quad (5)$$

此法标准化后的数据均值为 0, 标准差为 1, 消除了量纲影响; 当抽样样本改变时, 它仍能保持相对稳定性。

(2) 极差标准化法

$$X'_{ij} = \begin{cases} \frac{X_{ij} - \bar{X}_j}{R_j}, & \text{若 } R_j \neq 0 \\ X_{ij}, & \text{若 } R_j = 0 \end{cases} \quad (i = 1, 2, \dots, 88; j = 1, 2, \dots, 45) \quad (6)$$

此法标准化后的数据均值为 0，极差为 1，且 $|X'_{ij}| < 1$ ，消除了量纲的影响；在以后的分析计算中可以减少误差的产生。

(3) 极差正规化变换法

$$X'_{ij} = \begin{cases} \frac{X_{ij} - \min_{1 \leq i \leq 88} \{X_{ij}\}}{R_j}, & \text{若 } R_j \neq 0 \\ 0.5, & \text{若 } R_j = 0 \end{cases} \quad (i = 1, 2, \dots, 88; j = 1, 2, \dots, 45)$$

(7)

此法标准化后的数据最小值为 0，最大值为 1，其余均在区间 $[0, 1]$ 内，极差为 1，无量纲。

除以上方法以外，还有其他的标准化方法。在这些方法中，标准差标准化法应用最广泛，故采用此法对样本数据进行处理，得标准化数据矩阵：

$$X' = \begin{bmatrix} X'_{11} & X'_{12} & \cdots & \cdots & X'_{1,45} \\ X'_{21} & X'_{22} & \cdots & \cdots & X'_{2,45} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X'_{88,1} & X'_{88,2} & \cdots & \cdots & X'_{88,45} \end{bmatrix}$$

(8)

4.2 构造关系矩阵

关系矩阵用来描述样本之间的亲疏程度，其数量指标有两种：

一是“相似系数”：性质越接近的样本，相似系数越接近于 1 或 -1；彼此无关的样本相似系数则接近于 0，聚类时相似的样本聚为一类。

二是“距离”：将每一个样本看作 m 维空间的一个点，在这 m 维空间中定义点与点之间的距离，距离较近的点归为一类。

描述样本亲疏程度的方法有欧氏距离、平方欧氏距离、夹角余弦、皮尔逊相关系数、切比雪夫距离、绝对值距离、明科夫斯基距离等几十种，但常用的只是少数。

贵州省农业功能区划采用聚类分析中用得最广泛的欧式距离来构造县与县之间的关系矩阵（距离矩阵），把每个县都看成是 m 维（ $m = 45$ ）空间的一个点，点的坐标就是标准化后的那些 X 的值。欧式距离的计算模型为：

$$d_{ij} = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2}$$

(9)

式中， $i, j = 1, 2, \dots, n$ ，是系统单元数（参与聚类的县）； $k = 1, 2, \dots, m$ ，是系统特征（指标数）。通过欧式距离模型计算，得到一个 88×88 阶的方阵 D_{ij} ，它反映了两个聚类对象（县、市、区）之间的相似程度。两个聚类对象之间的距离越小，说明它们的总体特征越相似；对角线上的数值为 0，表示聚类对象自身的距离为 0。距离矩阵的结构为：

县序号 $\downarrow$

$D_{ij} =$

	1	2	...	87	88
1	0	d_{12}	...	$d_{1,87}$	$d_{1,88}$
2	d_{21}	0	...	$d_{2,87}$	$d_{2,88}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
87	$d_{87,1}$	$d_{87,2}$	...	0	$d_{87,88}$
88	$d_{88,1}$	$d_{88,2}$	...	$d_{88,87}$	0

(10)

距离矩阵是关于对角线对称的方阵, 对角线上的元素全部为 0, 具有 $d_{ij} = d_{ji}$ 的特性, 即 i 县到 j 县的距离等于 j 县到 i 县的距离。

4.3 聚类分析

根据距离矩阵, 可用系统聚类法或快速聚类法将全省 88 个县分为不同的类。

4.3.1 系统聚类

开始进行聚类计算时, 首先把每个县看成一类, 即总数有 88 类; 然后把距离最小的两个县 G_i 和 G_j 归并为一个新类 $G_r (G_r = \{G_i, G_j\})$, 进一步计算这个新类 G_r 与其他未归并县 G_k 之间的距离。计算公式为:

$$d_{rk} = \min\{d_{ik}, d_{jk}\} \quad (11)$$

式中, r 为刚刚归并的新类, k 为尚未归并的类(县)。

重复以上计算, 直到全部县被归并为一类为止。将归并过程记录下来, 可以得到一个系统聚类图, 能够清楚地看出距离和类别的关系。根据系统聚类图, 就可以把 88 个县按需要分为若干类。

4.3.2 快速聚类

(1) 事先确定距离聚类。通过观察和分析距离矩阵, 可事先确定某距离范围内的县(市、区)归并成一类, 对距离矩阵进行 0-1 化处理; 然后将 0-1 矩阵中一行内有“1”的那些县(市、区)归并为一类, 即可实现快速聚类。

(2) K-均值聚类(K-means Cluster)。此法的基本思想: 先粗略地进行预分类, 然后再逐步调整, 直到分类满意为止^[9-10]。整个聚类过程如图 2 所示:

4.3.3 聚类分析结果

根据以上介绍的聚类分析方法, 运用 Excel2003 和 SPSS13.0 统计分析软件^[11-12], 对贵州省农业功能区划采集的样本数据进行反复计算、分析和比较, 最终采用 SPSS13.0 的描述统计分析模块和快速聚类分析模块, 把贵州省的 88 个县(市、区)先分为 9 类, 然后把系统输出的第 7、9 类定义为 I 类, 第 5 类定义为 II 类, 第 3 类定义为 III 类, 第 1、2、4、6、8 类定义为 IV 类, 得到将 88 个县分为四大类的初步结果。

为了得到更加科学、合理和符合贵州实际的分类, 结合定量分析结论做了大量的定性研究, 邀请长期工作在农业战线、熟悉贵州情况的专家座谈, 请专家提意见, 对聚类分析结果进行微调。

通过定性(指标选择时就进行了大量的定性研究)→定量(聚类)→再定性(调整)的研究方法, 得到贵州省农业功能区划聚类分区的最终结果为: 将全省 88 个县(市、区)分为 4 类, 即 4 个区域。其中, I 区包含 24 个县、II 区 24 个县、III 区 20 个县、IV 区 20 个县。见表 1。

表 1 贵州省农业功能聚类分析方案

类别	县市数	县市区名称
I 区	24	开阳县、息烽县、修文县、遵义县、桐梓县、绥阳县、正安县、道真县、务川县、凤冈县、湄潭县、余庆县、习水县、仁怀市、石阡县、思南县、印江县、德江县、沿河县、松桃县、黔西县、金沙县、福泉市、瓮安县
II 区	24	六枝特区、水城县、盘县、普定县、关岭县、紫云县、兴仁县、普安县、晴隆县、贞丰县、望谟县、册亨县、安龙县、大方县、织金县、纳雍县、威宁县、赫章县、贵定县、平塘县、罗甸县、长顺县、龙里县、惠水县
III 区	20	江口县、玉屏县、万山特区、黄平县、施秉县、三穗县、镇远县、岑巩县、天柱县、锦屏县、剑河县、台江县、黎平县、榕江县、从江县、雷山县、麻江县、丹寨县、独山县、三都县
IV 区	20	南明区、云岩区、花溪区、乌当区、白云区、小河区、清镇市、钟山区、红花岗区、汇川区、赤水市、西秀区、平坝县、镇宁县、铜仁市、兴义市、毕节市、凯里市、都匀市、荔波县

通过聚类分析, 虽然把全省划分成了 4 个不同的区域, 但各个区域的主导功能和辅助功能结构关系并不清楚, 必须借助于其他方法对各区域农业的主辅功能进行准确定位。

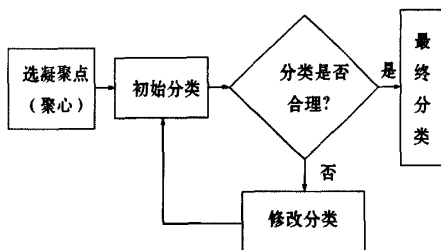


图 2 K-均值聚类过程

5 分区功能排序

有了分区就可以计算出分区指标平均值。按理，通过比较指标平均值的大小就可确定各区的主辅功能结构。但是，由于分区指标较多，仅凭观察难以做出正确判断。由于每个分区指标都有具体数值，可采用计算指标权重系数的办法，确定分区指标在各区所占的份额，从而确定各区农业功能在全省所处的地位。

5.1 分区指标权重系数计算方法

分区指标权重系数计算的基本思想是：把全省农业作为一个系统整体，而各个分区是这个系统的组成部分（子系统），分区指标就是系统功能的评价指标，对于某项指标或某项农业功能的综合评价指标来说，把全省看为 1，各分区所占份额即反映其在某项指标或某项农业功能上的强弱。

在此，指标权重系数等同于功能评价系数，包括单指标权重系数（单指标功能评价）和多指标综合权重系数（多指标功能评价）。

5.2 分区功能单排序

根据各功能的综合权重系数即可进行排序。但在计算综合功能权重系数时，必须注意各指标含义的一致性，对含义相逆的指标应进行处理后再计算。如用“单位土地富余劳动力数量”指标说明“就业和生活保障功能”强弱时，单位土地富余劳动力数量越多，就业压力越大，就业和生活保障功能越弱；用“水土流失面积占土地面积比例”和“石漠化面积占土地面积比例”指标说明“生态调节功能”时含义也是相逆的。

处理方法：在单指标权重系数计算之前，取相逆指标值的倒数，使其与其他指标的含义一致。

5.3 分区功能总排序及分区命名

5.3.1 总排序

通过农业功能单排序，各分区的农业功能在全省所处的地位已基本明晰。但它只是纵向排序的结果，当一个区的几个功能在全省排在同一名次时，孰先孰后难以分辨。因此，还需根据功能评价指标综合权重系数对每个区的农业功能做一次横向的总排序。功能权重系数可按（12）式计算：

$$g_{ij} = \frac{Q_{ij}}{\sum_{j=1}^n Q_{ij}} \tag{12}$$

式中， g_{ij} 表示第 i 区第 j 项功能的功能权重系数； Q_{ij} 为第 i 区第 j 项功能的多指标综合权重系数； i 为分区编号； j 为功能编号。

表 2 农业分区功能总排序

分区	农产品供给	就业和生活保障	生态调节	休闲功能
Ⅰ	系数 0.2910	0.2467	0.2426	0.2197
	排序 1	2	3	4
Ⅱ	系数 0.2610	0.2764	0.2566	0.1971
	排序 2	1	3	4
Ⅲ	系数 0.2521	0.2837	0.3146	0.1497
	排序 3	2	1	4
Ⅳ	系数 0.1961	0.2056	0.1998	0.3985
	排序 4	2	3	1

表 2 是按分区计算的功能权重系数及分区功能总排序结果，各分区的农业功能结构为：

Ⅰ区：农产品供给功能（主导功能）→就业和生活保障功能→生态调节和生态约束功能→休闲功能，

Ⅱ区：就业和生活保障功能（主导功能）→农产品供给功能→生态调节和生态约束功能→休闲功能，

Ⅲ区：生态调节和生态约束功能（主导功能）→就业和生活保障功能→农产品供给功能→休闲功能，

Ⅳ区：休闲功能（主导功能）→就业和生活保障功能→生态调节和生态约束功能→农产品供给功能。

Ⅳ区：休闲功能（主导功能）→就业和生活保障功能→生态调节和生态约束功能→农产品供给功能。

5.3.2 分区命名

农业功能区的命名法则，一般采用“地理方位（+地貌特征）+主导功能”模式，对功能区进行命名。但由于贵州是山区，在一个区域内山地、丘陵、河谷、盆地相互穿插，除山地以外，任何一种地貌特征面积都不超过区域面积的 50%，地形地貌极为复杂，在农业功能区名称中加入地貌特征的指导意义不大。因此，在分区命名中主要考虑 3 个因素：一是地理方位；二是主导功能；三是对今后区域农业功能拓展具有重大影响的辅助功能因素。分区命名为：

- I 区: 黔北农产品供给与就业生活保障功能区,
- II 区: 黔西就业生活保障与农产品供给功能区,
- III 区: 黔东南生态调节与就业生活保障功能区,
- IV 区: 区域中心城郊休闲、旅游与就业生活保障功能区。

6 结论

通过农业功能区划分方法的研究, 得出以下结论:

(1) 聚类分析是农业功能区划分的科学方法, 能够客观地把一个农业系统划分成若干综合功能特征差异明显的区域, 但并不能直接求出各区域的主辅功能结构关系, 需采用其他辅助手段, 如系统排序等方法, 才能阐明系统区域的功能结构。

(2) 在聚类分析基础上, 根据分区指标平均值计算功能权重系数, 对系统进行纵向和横向排序, 是识别系统区域功能结构关系的适用方法。该法简单、科学、合理, 便于应用。

(3) 建立指标体系是农业功能区划分的基础。要求指标体系能够综合、全面地反映农业系统的功能特征。

(4) 收集完整、可靠的指标数据, 是决定分区结果是否符合实际的重要前提。残缺和异常数据不能进入聚类分析系统, 否则会造成所分区域不连片而偏离实际。

参考文献

- 1 段学军, 陈雯. 省域空间开发功能区划方法探讨. 长江流域资源与环境, 2005, 14, (5): 540~545
- 2 陈艳梅, 戴茜. 河北省沙河市域生态功能区划和主导功能分析. 安徽农业科学, 2006, 34, (18): 4631~4633
- 3 罗其友, 陶陶, 姜文来, 等. 我国东北地区农业功能区划研究. 农业现代化研究, 2005, 26, (6): 407~412
- 4 吕晓芳, 王仰麟, 张镜铨, 等. 西部生态脆弱地区农业功能区划及模式探析. 自然资源学报, 2007, 22, (2): 177~183
- 5 刘卫东, 石承苍, 任国业. 四川省农业功能区划研究. 中国农业资源与区划, 2008, 29, (3): 27~32
- 6 包宁智. 农业资源区划的功能定位与发展对策. 中国农业资源与区划, 2006, 27, (5): 55~56
- 7 刘军萍, 荣文芳, 卢宏升. 北京农业功能区划研究. 中国农业资源与区划, 2006, 27, (5): 49~54
- 8 李树德, 李瑾, 黄学群, 等. 天津沿海都市型现代农业功能区划及发展重点研究. 中国农业资源与区划, 2008, 29, (3): 36~41
- 9 方开泰, 潘恩沛. 聚类分析. 北京: 地质出版社, 1982
- 10 杨善林, 李永森, 胡笑旋, 等. K-means 算法中的 k 值优化问题研究. 系统工程理论与实践, 2006, 26, (2): 97~101
- 11 郝黎仁, 樊元, 郝哲欧等. SPSS 统计分析. 北京: 中国水利水电出版社, 2003
- 12 吴明隆. SPSS 统计应用实务. 北京: 中国铁道出版社, 2001

A CLUSTER DISTRICT METHOD FOR AGRICULTURAL FUNCTION BASED ON SPSS

Lv Jingtang¹, Lv Daming², Zhang Hao²

(1. School of Mechanical Engineering of Guizhou University, Guiyang 550003;

2. Agricultural Resources and Regional Planning Research Center of Guizhou Province, Guiyang 550003)

Abstract Taking Guizhou agriculture system as the research object, this paper divided the entire province 88 counties into 4 regions by the application of cluster analysis method, and the establishment of agriculture function district indicator system, data standardization processing, the level cluster and the K-means cluster analysis. Through calculating mean value of partition indicator, and the application of system sorting method to carry on region synthesis function sorting and various regions function list sorting, this paper determined the agricultural dominant function and the auxiliary function, pointed out the direction for function expanding in each region, and could provide the basis for subarea naming.

Keywords Guizhou Province; agricultural function; cluster analysis; function sorting; subarea naming