

# 用聚类分析法划分山东省农业 气象产量气候区

孙仁邦

(山东省气象局, 济南 250031)

## 提 要

详细介绍了用聚类分析法划分山东省农业气象产量气候区的方法。用该方法制作产量分区预报, 不仅考虑了分区内农业气候条件的影响, 还考虑了分区内的生产力水平和行政区的划分, 从而使农业气象产量预报更具体、更准确且比较简便。

**关键词:** 聚类分析 划分 农业气象产量气候区

## 引 言

山东省是一个农业大省, 粮、棉、油生产在国民经济中占有举足轻重的地位。但目前全省的产量预报仍停留在对全省平均单产和总产的宏观预报上, 这不但影响了预报准确率的进一步提高, 而且也满足不了政府部门预先掌握全省各地区丰歉状况的需要。为使预报值具体、准确, 又不致使计算量过大, 按一定规则, 将全省划分成若干个农业气象产量气候区, 再分区建立预报模式, 是解决这一矛盾的有效途径。

在以往的农业气候区划成果中, 只考虑农业气候条件, 没有考虑生产力水平和行政区划等因素, 且目前尚没有与农业气候区划相对应的作物产量资料, 无法按划分的农业气候区建立农业气候因子与作物产量间的定量关系。为此, 本文本着“农业气象产量预报模式应根据气候条件、生产力水平、作物品种类型等因素, 适当兼顾行政区划, 分区建立”的原则, 以棉花为例, 用聚类分析方法, 探讨山东省农业气象产量气候区的划分。

## 1 区划因子的选择与处理

用于农业气象产量气候区区划的因子,

必须是对作物产量形成有明显影响且各地差异较大的因子。

对山东省而言, 4—10月份降水量反映了棉花生长季内的水分条件, 它的多寡对棉花生长发育和产量形成至关重要;  $\geq 10^{\circ}\text{C}$  的活动积温和4—10月份的日照时数则是衡量棉花生育期内光、温条件好坏的重要标志。这3个因子基本上反映了山东省棉花生长季内的农业气候条件, 且各地差异较大(见表1), 故选其为划分农业气象产量气候区的气候因子。从表1可见, 各地市的棉花单产大不相同, 这除了表明气候条件的差异外, 还表明各地的生产力水平不同, 在农业气象产量气候区划分时, 必须考虑这一因素。这也是区分农业气象产量气候区区划与常规的农业气候区划的主要标志。

在山东省现有的16个地市中, 烟台、威海、日照3市基本不种棉花, 故在区划时不予考虑。

表1中的气候因子是现行地市内所含气象台、站自建站至1980年的平均值; 为了较好地反映各地市近代生产力水平, 本文采用近11年(1981—1991年)各地市棉花单产平均值作为产量因子。

表1 划分农业气象产量气候区的因子值

台 站	滨州	德州	聊城	菏泽	泰安	临沂	济宁	潍坊	淄博	枣庄	青岛	济南	东营
4-10月降水量/mm	568	557	552	604	643	776	645	635	600	743	684	589	552
≥10℃活动积温/℃	4306	4380	4464	4864	4565	4310	4560	4205	4299	4568	4040	4516	4254
4-10月日照/小时	1709	1726	1684	1611	1556	1560	1581	1643	1661	1480	1630	1697	1709
棉花产量/kg·亩 <sup>-1</sup>	54	66	62	52	49	44	61	58	57	59	52	56	54

我们可将表1看作由  $m(=4)$  个变量、 $N(=13)$  个样本构成的样本矩阵

$$X = (X_{ij}) \quad \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2, \dots, 13 \end{cases}$$

矩阵  $X$  中各元素的观测值因量纲或量级不同,可能对数据分类带来不合理影响。为了消除这些影响,本文对矩阵  $X$  中的原始数据进行标准化处理,即将矩阵中  $X_{ij}$  变换为

$$X_{ij}^* = \frac{X_{ij} - \bar{X}_i}{S_i} \quad (1)$$

其中  $\bar{X}_i$  为第  $i$  个因子的均值,  $S_i$  为其标准差,形成新矩阵  $X^* = (X_{ij}^*)$ 。

矩阵  $X^*$  中各因子的平均值均为0,标准差都为1,且与因子的量纲没有关系。

### 2 样本间的相似系数

矩阵  $X^*$  中的每一列表示一个样本(本文中则代表一个地市的状况),任两个样本  $X_j^*$  与  $X_k^*$  之间的相似性,可通过矩阵中第  $j$  列与第  $k$  列的相似程度来表示。本文选用相似系数作为衡量两个样本间相似程度的统计量。即把任两个样本  $X_j^*$ 、 $X_k^*$  看成  $m$ (变量个数)维空间的两个向量,用两个向量的夹角余弦(即相似系数)  $\cos\theta_{kj}$  表示其相似程度。

$$\cos\theta_{kj} = \frac{\sum_{a=1}^m X_{ak}^* \cdot X_{aj}^*}{\sqrt{\sum_{a=1}^m X_{ak}^{*2} \cdot \sum_{a=1}^m X_{aj}^{*2}}} \quad (2)$$

$$Q = \begin{bmatrix} \cos\theta_{11} & \cos\theta_{12} & \dots & \cos\theta_{1N} \\ \cos\theta_{21} & \cos\theta_{22} & \dots & \cos\theta_{2N} \\ \dots & \dots & \dots & \dots \\ \cos\theta_{N1} & \cos\theta_{N2} & \dots & \cos\theta_{NN} \end{bmatrix}$$

相似系数矩阵  $Q$  是一个实对称矩阵,其中  $\cos\theta_{11} = \cos\theta_{22} = \dots = \cos\theta_{NN} = 1$ 。

矩阵中  $-1 \leq \cos\theta_{kj} \leq 1$ , 当  $\cos\theta_{kj} = 1$  时,

说明  $X_k^*$  与  $X_j^*$  完全相同,  $\cos\theta_{kj}$  愈接近1,说明  $X_k^*$  与  $X_j^*$  愈相似;愈接近-1,说明两个样本差异越大,当  $\cos\theta_{kj} = -1$  时,说明  $X_k^*$  与  $X_j^*$  完全不一样。这样我们便可根据矩阵  $Q$  对  $N$  个样本进行分类,把比较相似的样本归为一类,不怎么相似的归于不同的类。

### 3 形成聚类图

依据相似系数矩阵  $Q$ ,对全省13个产棉地市进行聚合归类,最后形成聚类图。聚合分类遵从以下原则:

- 3.1 聚合分类按样本间相似系数的大小,由大到小依次进行。即先选相似系数最大的两个样本,将其归为一组,再选次大的。
- 3.2 若选出的一对样本在已分好的组中都未出现过,则把它们形成一个独立的新组。
- 3.3 若选出的两个样本中,有一个是在已划分好的组中出现过,则把另一个样本也加入到该组中。
- 3.4 若选出的一对样本,都分别出现在已划分好的两组中,则把这两个组连结在一起。
- 3.5 若选出的一对样本都出现在已划分好的同一组中,则这对样本就不用再分组了。

按以上5条原则反复进行,直到把所有的样本都聚合成一个大组为止,得出样本聚类表2。

表2中  $X_1^*$ 、 $X_2^*$ 、 $\dots$ 、 $X_{13}^*$  为样本矩阵中13个样本,依次代表表1中的13个产棉地市。

根据表2可作出13个产棉地市棉花农业气象产量气候区的区划聚类图(见图1)。

若取相似系数  $\geq 0.450$  作为归类分组的下限指标,可将全省13个产棉地市归类为5个大区。即:

