DOI:10.13203/j.whugis1985.02.003

二维有序聚类方法及其在编制区划地图中的应用*

郭 仁 忠

摘 要

本文在简要分析现有的二维有序聚类方法的基础上，对地理区划问题进行了一般性分析，将其归纳为三种类型，进而给出了对三种类型普遍适用的二维有序聚类分析的一个新定义及其相应算法，并以实例说明了二维有序聚类方法在编制区划地图中的应用。

一、引 言

聚类分析为样品分类、编制类型图和区划图提供了较好的数学模型。但是通常的聚类分析没有考虑样品的地理区域分布，仅仅是根据样品的统计变量观测值进行类的划分。如图1所示，诸样品被分为三类，但同一类样品地理位置上并不一定相邻。在编制区划地图的工作中往往需要根据资料对整个制图区域进行分区，用地图表示这种区域性特征。对于这个问题在多元统计分析方法以及计量地理学中均有所提及，但尚无较好数学模型解决之。这个问题对分析的要求是：

1. 相似（统计特性相近）统计单元归为同一类；
2. 同一类样品在地理分布上是分区成片连通的（如图2所示）；
3. 由于某些用途要求区划单元在地理分布上有相对的地区集群性（图2中a比b区划单元的集群性程度高）。

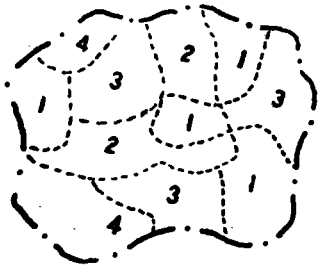
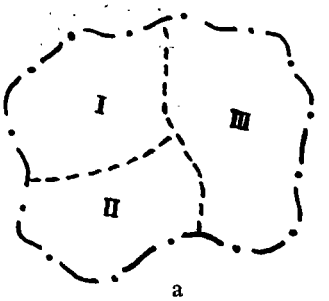
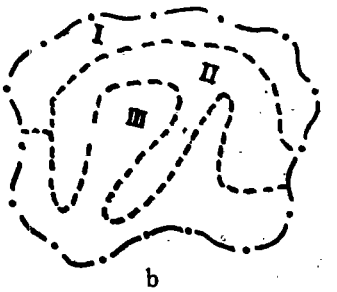


图1



a



b

图2

本文将就以上三点要求展开讨论，给出二维有序样品聚类分析的另一个定义，给出进行

本文1984年7月收到。*本文为研究生毕业论文的一部分，指导教师是张克权、徐庆荣副教授。完成过程中得到胡毓巨、刘士英、费立凡等老师的帮助，谨此致谢。

有序聚类的两个算法，最后以实例说明二维有序样品聚类分析方法在编制区划地图中的应用。

二、 現有方法的簡单分析

多元分析中二维有序聚类分析的定义首先是由中国统计学者给出的〔3〕，并提出了两个算法。当时的出发点是为了解决地质勘查中的分区问题。其定义如下：

定义一。设平面上有n个点y_1，y_2，y_3，…，y_n，如果对于任意K（$2 \leqslant K \leqslant n-1$），将其分为K类：$G_1$，$G_2$……，$G_k$，要求它们的最小支撑树不相交，则称$y_1$，$y_2$，……，$y_n$为二维有序样品，类$G_1$，$G_2$，……，$G_k$内的样品是互相邻接的。

〔3〕中根据定义一给出的两个算法从现有资料看尚无计算实例，其主要原因是算法较复杂（计算量大），程序设计比较困难，另外由方法知计算结果比较勉强。

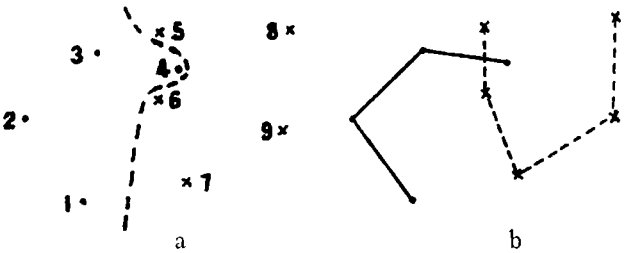


图 3

此外，定义一仍有两点需要进一步推敲：

1． 实际分析中最小支撑树不相交的要求往往可以省略。如图3所示，通过一般聚类分析，9个样品（a中所示）可以分为两类，$G_1=\{1, 2, 3, 4\}$，$G_2=\{5, 6, 7, 8, 9\}$。如果按照图3 a中短划线将两类分开，可以作为分区结果，但是从b中可以发现G_1，G_2的最小支撑树是相交的，这样根据定义一就只有改变原来分类，势必引进较大误差。

2． 由于最小支撑树的计算及其相交的判断极为复杂，尚无较好算法，而定义一局限于最小支撑树使问题的求解复杂化。

在计量地理学中借助于判别分析进行地理分区。但是判别分析是在已知总体的情况下进行的，因此需要分析之前先定性地将分析区域分区，然后在这个基础上进行判别分析，显然这种做法人工干预过多，结果又受定性分区的影响。总之，探讨新算法是很有必要的。

三、 基 本 原 理

实际工作中涉及二维有序聚类（亦即地理分区）的问题种类繁多，但适当归纳概括，并不难抽象为三种类型。为说明问题方便，以专题制图中容易碰到的例子说明问题。

例1．根据全国10万人口以上城镇的教育系统各类学校的统计数据进行全国范围内城镇教育结构分区，编制城市教育结构区划图。

例2．按县统计农副业生产产品单产（单位土地面积上的产量，非单位种植面积上的产量），编制全国农副业生产现状区划图。

例3．以各县气象台（站）的多年气象观测资料为原始数据编制全国气候区划图。

以上三例有一定代表性，它们的共性在于进行区划。但又各具个性，例1的统计数据是定位于点的资料，不包含除点以外的任何邻域的信息，分析的目的在于研究点的特性。例2的统计资料是来源于诸统计单元，一个样品表示一个面(一块小区域)，诸样品表示的统计单

元构成分析区域的一个划分。例 3 的统计资料定位于点但可以表示该点的某邻域内的特性，而这邻域是一个模糊集。

以上三例的各自个性代表着一种类型，故二维有序聚类的三种类型为：

设样品，y_1，y_2，……，y_n，相应地理统计单元（或点）为 $\tilde{y}_1$，$\tilde{y}_2$，……，$\tilde{y}_n$，设分析区域为平面连通点集 Ω，分区结果为 $\tilde{G}_1$，$\tilde{G}_2$，……，$\tilde{G}_k$，$\tilde{G}_j \subset \Omega (j=1,2,\cdots\cdots,k)$。

则　类型1．　$\tilde{y}_i \in \Omega$，$\tilde{G}_j \subset \Omega$，$\bigcap_{j=1}^{k} \tilde{G}_j = \phi$，$\bigcup_{j=1}^{k} \tilde{G}_j \subset \Omega$

类型2．　$\tilde{y}_i \subset \Omega$，$\tilde{G}_j \subset \Omega$，$\bigcap_{j=1}^{k} \tilde{G}_j = \phi$，$\bigcup_{j=1}^{k} \tilde{G}_j = \Omega$

类型3．　$\tilde{y}_i \in \Omega$，$\tilde{G}_j \subset \Omega$，$\bigcap_{j=1}^{k} \tilde{G}_j = \phi$，$\bigcup_{j=1}^{k} \tilde{G}_j = \Omega$

$(i=1, 2, \cdots\cdots, n;\ j=1, 2, \cdots\cdots, k)$

类型 1，2，3 用图分别说明见图 4，5，6。

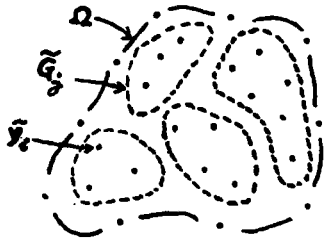


图 4

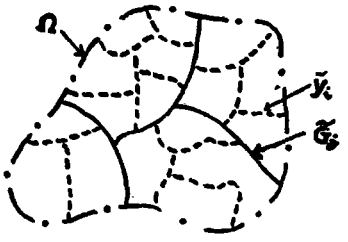


图 5

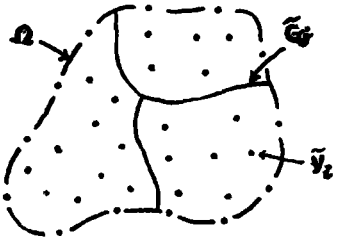


图 6

图 5 所示的类型 2 的情形是地理分区中最常见的。类型 1 和类型 3 可以转化成类型 2 来处理。对于类型 3 而言，定位于点的样品 y_1，y_2，……，y_n 分别表示各点某邻域的情况，对各点这样修改 $\tilde{y}_i$，使 $\bigcup_{i=1}^{n} \tilde{y}_i = \Omega$，$\bigcap_{i=1}^{n} \tilde{y}_i = \phi$，则就可以按类型 2 处理。

$\tilde{y}_i$ 的确定可按下列两方法之一：

1．　定性确定 $\tilde{y}_i$，如例 3 可按行政区划确定。

2．　按重力模式确定 $\tilde{y}_i$：

如图 7，x 为分析区域中任一点（$x \in \Omega$），a_i 为各样品的权值，则

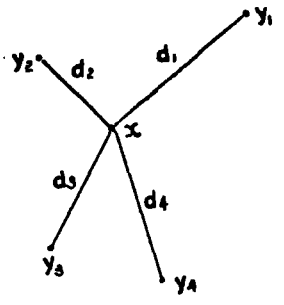


图 7

$$x \in \tilde{y}_i \Longleftrightarrow \max\left\{\frac{a_1}{d_1}, \frac{a_2}{d_2}, \cdots\cdots, \frac{a_n}{d_n}\right\} = \frac{a_i}{d_i}。$$

对于类型 1，实际上仍可按类型 3 处理，这是因为如果样品是抽样所得，那么在布点过程中总是考虑以“点”代“面”的，所谓区划是对面而言而不是对点而言。如果点是固定的，如例 1 的城市，这种分区的要求是模糊的，目的是反映一个概略的区域性、集群性规律，按类型 3 方法处理完全可以。

完成以上分析，下面给出二维有序聚类的另一个定义。

定义二．设容量为 n 的样本 $y=\{y_1, y_2, \cdots\cdots, y_n\}$ 和一个二维平面上的单连通闭集

Ω，存在Ω的一个划分$\tilde{y}=\{\tilde{y}_1, \tilde{y}_2, \cdots\cdots, \tilde{y}_n\}$，$\tilde{y}_i$都是单连通的，从y到$\tilde{y}$存在一个一对一的映射f：$\tilde{y}_i=f(y_i)$。

若求得y的一个划分$G=\{G_1, G_2, \cdots\cdots, G_k\}$和$\Omega$的一个划分$\tilde{G}=\{\tilde{G}_1, \tilde{G}_2, \cdots\cdots, \tilde{G}_k\}$，满足

$$y_i \in G_i \Longleftrightarrow \tilde{y}_i \subset \tilde{G}_j \quad (i=1,\cdots\cdots,n;\ j=1,2,\cdots\cdots,k;\ \tilde{G}_j \text{是单连通子集})$$

则称G为y的一个关于Ω的有序分类。

定义二的意义不仅在于给出了问题的确定性描述，而且在于定义本身蕴含了有序聚类的算法。根据定义二，聚类过程中只要保证两条：（1）相似样品（类）予以合并；（2）被合并的两样品（类）按f对应的Ω的两个子集的并是一个单连通子集。这样判断和计算都是比较容易的。

四、 計 算 方 法

本节根据类型2给出二维有序聚类的两个算法，同样适用于类型1和3。

在地图制图中，定义二中的Ω就是制图区域（分析区域），诸统计单元（通常是行政单元）就构成了Ω的一个划分，因而就是$\tilde{y}_i$，如图8所示。图8给出了样本容量$n=11$的例子，这不失一般性，当n为有限数时完全一样地处理。

图8

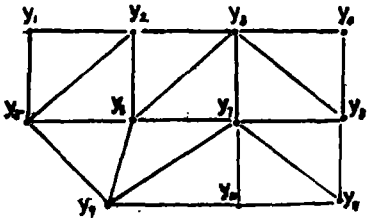


图9

根据图中的统计单元的相邻关系得到图9，图9称为关联略图，图8和图9存在如下关系：$\tilde{y}_i \ R \ \tilde{y}_j \Longleftrightarrow y_i \ r \ y_j$（R表示相邻，r表示邻接）。

设x_{1j}，x_{2j}为第j个样品的地理坐标（几何中心或行政中心或其它），y_{lj}为第j个样品的变量统计值（$i=1, \cdots\cdots, n$；$l=1, 2, \cdots\cdots, p$），n为样本容量，p为变量数。

算法一：

1.
$$d_{ij}=\begin{cases}\sqrt{(x_{1i}-x_{1j})^2+(x_{2i}-x_{2j})^2} & y_i \ r \ y_j \\ \infty & y_i \ y_j \text{不邻接}\end{cases}$$

2.
$$\tilde{d}_{ij}=\sqrt{\sum_{l=1}^{p}(y_{li}-y_{lj})^2} \qquad i, j=1, 2, \cdots\cdots, n;\quad l=1, 2, \cdots\cdots, p$$

3. $D_{ij} = \widetilde{d}_{ij} + wd_{ij}$　　$0 < w \leqslant 1$

4. 从D_{ij}出发用最短距离法聚类。

算法一符合定义二的要求是显然的，具有计算简便的优点，但聚类过程中只适宜用最短距离法。为此提出算法二。

算法二：

1. 计算距离矩阵：$D_{ij} = \sqrt{\sum_{l=1}^{p} (y_{li} - y_{lj})^2 + w \sum_{l=1}^{2} (x_{li} - x_{lj})^2}$　　$0 \leqslant w \leqslant 1$

2. 生成关联矩阵$[r_{ij}]$，根据关联略图得：

$$r_{ij} = \begin{cases} 1 & y_i r\ y_j \\ 0 & y_i\ y_j \text{ 不邻接} \end{cases}$$

3. 在满足$r_{ij} = 1$的条件下进行类的合并。

4. 如下修改r_{ij}，D_{ij}：

（1）设p，q合并为新类t，与某类S的关联性为

$$r_{tS} = r_{pS} \vee r_{qS}$$

（2）相应的D_{St}按系统聚类法中的任一种方法修改即可。

算法二比算法一稍复杂，但适合于各种系统聚类法。两个方法中的w可以通过试验确定。

顺便指出，算法一和算法二都是通过对系统聚类法加以限制而实现有序聚类的，这样势必引入较大的分类误差，也就是说同一类的样品相似性程度可能并不高。一般区划具有两个目标，一是分类误差要小，二是分区结果要使各区平面图形相对集中（图2a）。这是一对矛盾，相互制约，因此在有序聚类中笼统地说找最优解是没有意义的。

五、 算例介紹

下面用算例说明算法二在编制区划地图中的应用方法并考察分析结果的可靠性。至于算法一与此无任何差异，不另举例。

今选取湖北省分县农副业产品总产统计数据，进行湖北省农副业生产结构分区，项目如下：

水稻、小麦、棉花、油菜籽、茶、蔬、木材、鱼（原始数据略）。

根据统计数据利用有序聚类分析方法找出湖北省农副业生产的地域性特征，计算步骤如下：

1. 制作关联略图，
2. 生成关联矩阵　$r = [r_{ij}]$，
3. 计算单位面积（总产量与全县总面积之比）产量，
4. 数据标准化，
5. 计算距离矩阵$D = [D_{ij}]$。

本例采用类平方和最小增量法聚类，所谓距离矩阵实质上是类平方和增量矩阵。设样本容量为n，变量数为p，开始时各样品自成一类，类平方和为零。任意两样品i、j合并，类平方和增量为：

$$\Delta_{ij}=\sum_{l=1}^{p}\{[y_{li}-(y_{li}+y_{lj})/2]^2+[y_{lj}-(y_{li}+y_{lj})/2]^2\}$$

$$=\sum_{l=1}^{p}(y_{li}^2/2-y_{li}y_{lj}+y_{lj}^2/2)$$

$$=\sum_{l=1}^{p}\frac{1}{2}(y_{li}-y_{lj})^2=\frac{1}{2}d_{ij}^2$$

当 k，1 两类合并为类 S 时，则类 S 与任意类 t 进一步合并的类平方和增量为：

$$\Delta_{St}=[\Delta_{kt}(N_t+N_k)+\Delta_{lt}(N_t+N_l)-\Delta_{kl}N_t]/(N_k+N_l+N_t)\quad[1]$$

上式中 N_S 表示类 S 所包含的样品个数，余类推。以上是一线性表达式，因此用 d_{ij}^2 代替 Δ_{ij} 只差一个常系数$\frac{1}{2}$，并不影响分析结果，故令 $D_{ij}=d_{ij}^2$ 作为聚类统计量。其它系统聚类法都可类似处理，此不赘。此外，本例选 $w=0$ 是考虑到本例并不要求分区结果具有相对的地区集群性，如果 $w\neq 0$，以上计算方法仍适用。

6. 从$[D_{ij}]$、$[r_{ij}]$出发用系统聚类法中的类平方和最小增量法聚类，计算框图见图10。

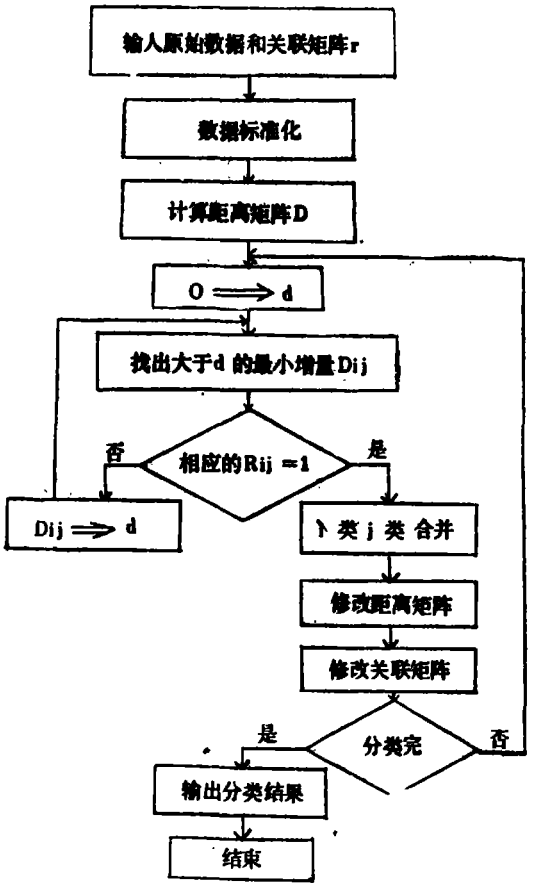


图10

本例按计算结果广济县自成一区，经人工干预将其并入第一区，这样湖北省农副业生产结构分为五区，如图11所示，各区情况简述如下：

第一区为全省主要的粮棉区，藏、油料的大部分也集中于该区，该区地理位置处于江汉平原内，水、土资源条件较好，是棉、粮的适宜产区。此外，该区水产业发达，湖北省鱼产量一半以上出于此区。

第二区为沔阳、天门、潜江三县，该区地理上也属于江汉平原，区别于第一区的主要原因在于粮棉生产比例不同，该区是全省主要的，全国有名的产棉区，三个县单位面积上的棉花产量分别为247，307，121。是邻县的几倍甚至十几倍。

第三区由鄂东南五县组成，该区是重要的茶叶、藏类产区，茶藏生产占全省的75%左右，其余各类农副产品的生产也都有发展。

第四区由江汉平原边缘的低山、丘陵、岗地诸县组成，该区的农副业生产各业相对均衡，粮、棉、油、茶都有相当规模。

第五区为鄂西山地，该区为全省木材、茶的主要产区，粮、棉、水产、藏的生产水平较低，产量也不高，油料生产具有一定规模。

以上分析是完全根据统计数据作出的。图12是传统的地貌分区，将图 11 和图 12 作一比

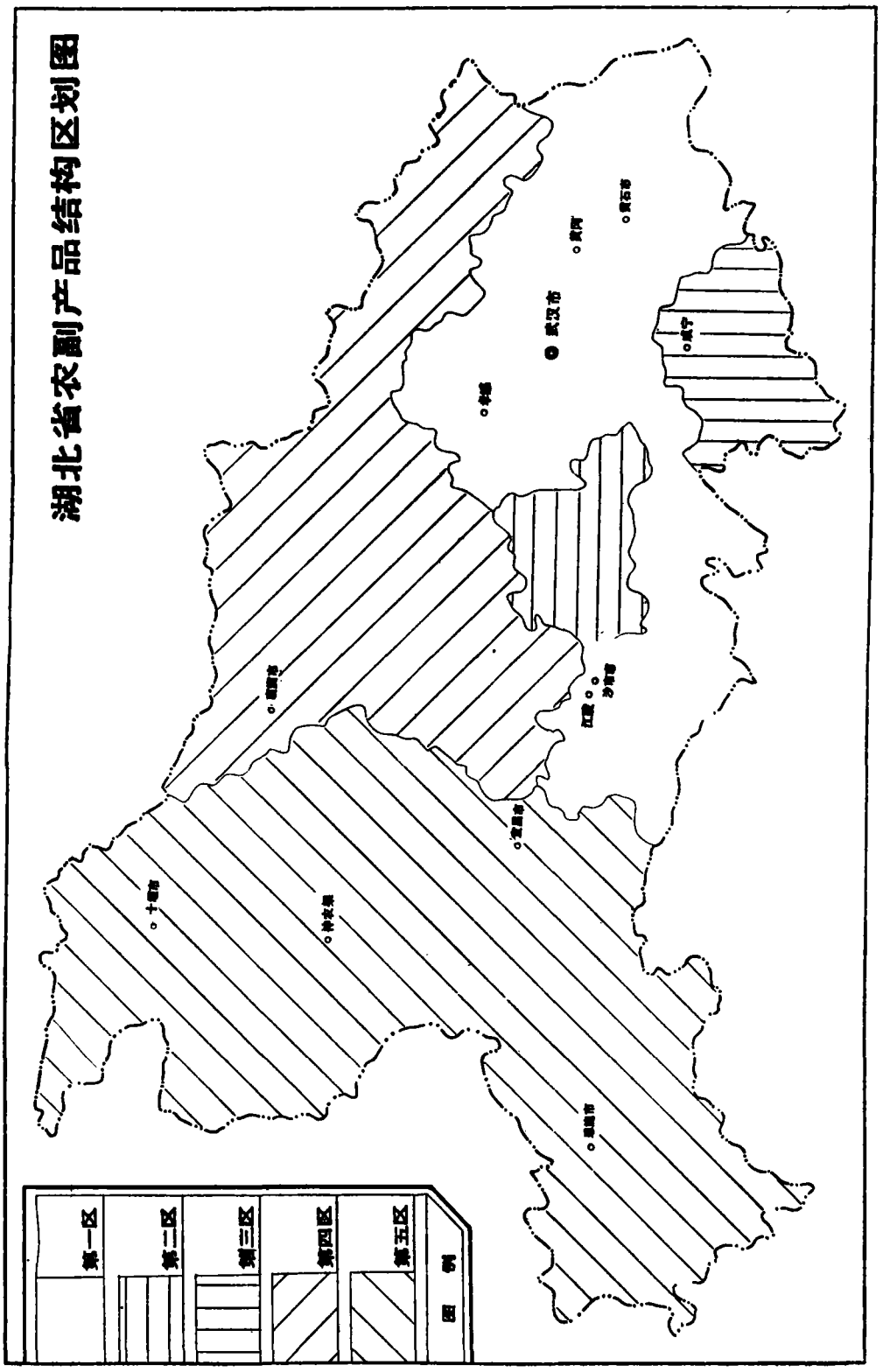


图11

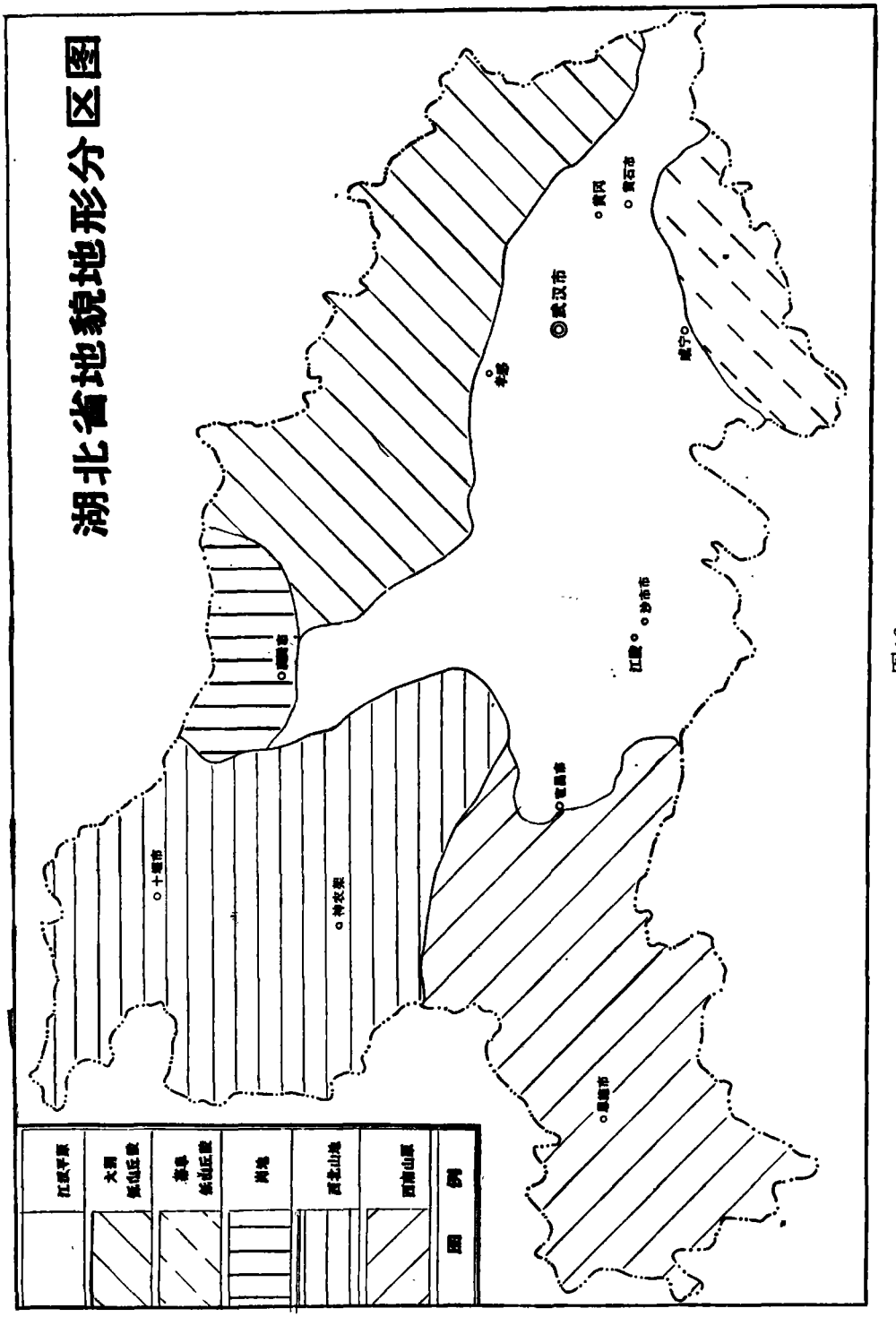

湖北省地貌地形分区图
武汉市
黄冈
黄石市
咸宁
孝感
荆州市
江陵
沙市市
襄樊市
十堰市
神农架
恩施市
图例
江汉平原
大别山区
武当山区
岗地
西北山地
西南山区

图12

较，容易发现两种分区结果在一定程度上相协调，主要差异是鄂西山区，图12将其分为两区，而图11将其合为一区。鄂北岗地在图10中也没有得到相应的体现。这些差异的产生应从两方面去分析，一是这些地区的农副业生产尚未充分做到因地制宜；二是这些地区地貌上的差异对农业生产的影响还不够大，因此从产量上还不能反映出来。图11和图12的总体协调说明湖北省农副业生产的发展在一定程度上受地形地势的影响，从这点上来说，全省农副业生产是基本符合因地制宜发展生产的策略的。在进行农业生产的区域化、专门化规划时，通过各种专题的分区图的比较分析，结合国家计划，人民需要，经济效益等因素综合平衡，就可以得出一个较优的农业综合区划方案。

顺便指出，在实际分析中，统计单元应更小些，本例中以县为统计单元显然太粗了，这也是图11和图12分区差异的一个因素。

参 考 文 献

〔1〕 张尧庭、方开泰，多元统计分析引论，科学出版社，1982。

〔2〕 Kang-tsung Chang, Multi-component Quantitative Mapping, The Cartographic Journal, Vol. 19 No.2 December 1982.

〔3〕 方开泰、潘恩沛，聚类分析，地质出版社，1982。

〔4〕 唐文雅、叶学齐、杨宝亮，湖北自然地理，湖北人民出版社。

Clustering Method for 2—Dimensional Ordered Samples and Its Application to the Compilation of Maps of Regional Division

Guo Renzhong

Abstract

In this paper, on the basis of simple study of the existing methods for 2-dimensional ordered samples, the problems of geographic regionalization are generally analysed and concluded into 3 groups, further more, a new definition of clustering analysis for the 2-dimensional ordered samples and its corresponding algorithms which are generally applicable are carried out. Finally, an example is offered to illustrate the use of the given method in the compilation of maps of regional division.