数据的分级处理

05级　李爱华

主要内容:

1. 数据分级处理的必要性

2. 数据分级处理的基本原则

☆3. 数据分级的方法

分级处理的必要性：

- 原始数据不能直观地反映： a现象在空间分布上的规律性；b由于数量差异而产生质量差异感、特殊的水平或集群性
- 数据一旦分级，级内数据的数量差别消失，造成一些信息损失，但是，它也为读者提供了更加直观的信息，把同质区域作为一个等级表达出来，提供集群概念
- 分级的重要任务：找出关键的临界值，增强同级间的同质性和各级间的差异性

分级的基本原则：

A 分级处理主要包括：

1.分级数量的确定：受地图用途、地图比例尺、数据分布特征、表示方法、数据内容实质、使用方式等多种因素的制约

- 要做到详细性与地图的规律性、易读性的统一
- 分级数与采用的表达方式有着密切的关系

艺术符号表示：宜3级

几何符号表示：5-7级

线状符号表示：宜3级

分级统计图用面积色表示：同色表达最多5级，两种颜色表达可以区分7-8级

分区统计图表：粗略的3级，最多5-7级

2.分级界限的标定：常见图例中分级界限标定比较混乱，正确的标定方法是左闭右开或者左开右闭。

3.分级界限的确定：保持数据分布特征和分级数据有一定的统计精度

- 按数据的分布特征分级的原则，适合于任何要素和现象（最常见）
- 按各分级单元的个数分布的原则，适合于按一定指令性标准反映单元个数的统计
- 按地图上各级面积分布的原则，适合于与实地面积有关的分布现象，如人口密度

分级的基本原则（续）：

B 分级的基本原则是各种分级方法的基础

1. 客观反映数据的分布特征，以数据的集群性作为分级数的重要依据
2. 分级界限应该在数据变化显著特征上，使各级内部差异尽可能的小，等级之间的差异尽可能大
3. 分级的结果：一般是中间级别包含的单元多，两端级别包含单元较少。也有要求分级单元数近似相等
4. 根据地图的用途和要素特征，要保留个别的特征级别和分级界限
5. 为了用途的方便，应适当地保持凑整地分级界限
6. 对于离散分布的现象，且物理个数不多，相邻级别的分级界限可以断开；对于连续分布现象的分级，其界限必须是相互连接的，并要正确处理分级点的所属关系

☆数据分级方法

从数据的特征，主要有以下三类:

1.考虑数据类型及其分布特征的分级方法

2.按分级数据单元物体的个数进行分级的方法

常用于依据给定的某一级别或几级数量指标标准的社会经济现象的分级。分级简单，适合于绝对数量和相对数量指标的分级

3.按地图上各级分布面积对比的分级方法

主要用于反映与面积相关的数量指标的分级，通常为相对指标

考虑数据类型及其分布特征的分级方法

—既适合于绝对数量的分级，也适合于相对数量的分级；既适合于点状分布要素，也适合于线状和面状分布要素

—分为两类：

- 按照简单的数学法则，主要有数列分级法、级数分级法等
- 统计学分级方法，即按某种变量系统确定间隔的分级，主要有统计量分级法、自然裂点法、自然聚类法、迭代法、逐步聚类和模糊聚类等法

按照简单的数学法则分级方法			统计学分级方法									
数列分级方法			级数分级方法		按某种变量系统确定分级间隔的分级方法						聚类分级方法	
等差数列分级	等比数列分级	倒数分级方法	算数级数分级	几何级数分级	自然裂点法	按正态分布参数分级	按嵌套平均值分级	按分位数分级	按面积等梯度分级	按面积正态分布分级	逐步聚类分级方法	模糊聚类分级方法

- 用于具有均匀变化的制图现象，其特点是差级相等便于比较
- 设H为数列的最高值，L为数列的最低值，N为预分的级数

$$A_i = L + \frac{i-1}{K}(H-L) \qquad (i = 1,2,...,K+1)$$

实际使用中，K, A_i 都应当凑成整数

返回

- 等比数列分级

$$\lg A_i = \lg L + \frac{i-1}{K}(\lg H - \lg L)$$

$$A_i = L(\frac{H}{L})^{\frac{i-1}{K}} \qquad (i = 1, 2, \ldots, K+1)$$

返回

- 倒数数列分级

$$\frac{1}{A_i}=\frac{1}{L}+\frac{i-1}{K}(\frac{1}{H}-\frac{1}{L})$$

$$A_i=(\frac{1}{L}+\frac{i-1}{K}(\frac{1}{H}-\frac{1}{L}))^{-1} \qquad (i=1,2,\ldots,K+1)$$

返回

- 直接对分级间隔进行选择
- 设H为数列的最高值，L为数列的最低值，Y为级差基数，B_i为某级所需级差基数的倍数值，在数列中为第i项

$$Y = \frac{H - L}{\sum_{i=1}^{K} B_i}$$

- 在等差数列中 $B_i = a + (i-1)d$
- 在等比数列中 $B_i = gr^{i-1}$
- 在采用的级差为算术级数或者几何级数时，也可以采用以下的六种变化方法来确定分级间隔：①按某一恒定的速率递增②按某一加速递增③按某一减速度递增④按某一恒定速率递减⑤按某一加速递减⑥按某一减速度递减

返回

- 自然裂点法：某种现象的观测值或者统计值可能不是均匀分布的、有自然裂点
- 按正态分布参数分级:首先计算数列的平均值Z和标准差S。可以分为
- 按嵌套平均值分级：首先计算整列平均值，然后把数列分成2^n个等级
- 按分位数分级：将数列分成分成若干段，每分段中的个数相等
- 按面积等梯级分级：当数据表上具有制图区域各统计单元的面积时，按其统计面积的大小排序；累加面积值作为分段依据，依据需要分级；每个等级中的样本数不一样，但各级面积基本一致
- 按面积正态分布分级：同按面积等梯级分级，但是按正态分布的规律是中间级别所占的面积较大，往两端依次减小；每个等级中的样本数不一样

分级结果的检验：一般以以下两个标准来检验

- 各级中样本数成正态分布或均匀分布

- 同级区域的连通性

  优良的分级应当使分级后产生的区域数相对较少，即连通性较大，通常用破碎指数来衡量：

$$F = \frac{m-1}{n-1}$$

m为分级后产生的区域数，n为地图上表示的单元总数

F=1　　没有任何两个单元连通

F=0　　所有单元连通为一个区域

0〈F〈1　　一般情况

返回

第一步：数据排序

为了便于制图聚类图和确定分级界限，数据从小到大排列

第二步：建立相似矩阵

1.关键是确定样本间的相似性，常用的相似性统计量是相关系数、夹角余弦和距离系数等

2.常见的计算相似系数方法：最大最小法、算术平均法、几何平均法

第三步：聚类分级的逐步计算

1.求出相似矩阵中的最大元素rij

2.划去矩阵中的第i行、第j列

3.将原始数据中的第i个和第j个数据　　加权平均后代替第j个数据

4.计算除去第i个数据以外的其余数据的相似系数矩阵

5.要了解数据间的自然聚合情况，重复以上计算

返回

模糊聚类分级法

- 依据逐步聚类法确定分级的基本思想，按照数据之间的相似程度确定分级时，一个数据属于哪一个等级并不是绝对的，而有一定的模糊性
- 计算步骤

1.数据排序

2.建立相似矩阵

3.相似矩阵转化为等价矩阵

模糊相似关系一定是满足自反性和对称性，但一般而言，它并不一定满足传递性，也就是说它不一定是模糊数学等价关系。因而，需要采用传递闭合的性质将模糊相似性关系通过自乘改造为模糊等价关系。

4.由等价矩阵进行聚类分级

end

<100　100-300　300-500　500-700　700-1000　>1000(元/人）

<100　101-300　301-500　501-700　701-1000　>1001(元/人）

0-99　100-299　300-499　500-699　700- 999　>1000(元/人）

正确的标定方法是左闭右开或者左开右闭：

≤100　100-300　300-500　500-700　700-1000　>1000(元/人）

<100　101-300　301-500　501-700　701-1000　≥1001(元/人）

返回