第八章　地理空间相关性

Tobler 提出的地理学第一法则就说明了地理空间中任何事件（物）之间是相关的，距离越远，相关程度就越弱，距离越近，相关程度就越强。因此，研究地理空间的相关性是非常重要的，也是地理空间推理过程中经常需要考虑的问题，地理空间相关性主要表现在地理空间的自相关性，以及特征之间在区域或时间上的相关性。

第一节　地理空间自相关

一、地理空间自相关的定义

地理空间自相关反映的是一个区域单元上某种地理现象或某一属性值与邻近区域单元上同一现象或属性值的相关程度，可以简称为“空间自相关”。它可以检验一个位置上的一个变量的观测值与邻近位置上同一变量的值是否显著相关。它也与地球表面某一位置上的目标或活动与邻近位置上的目标或活动的相似程度有关。因此，当同时处理位置信息和属性信息时，空间自相关是一种特别的、非常有效的分析技术，它作为一个描述性指标，提供了某一现象的空间分布信息，反映了一个观测对其周围观测的影响度。

二、地理空间自相关的特性

空间自相关可以分为两种不同等级的指标：全局指标和局部指标。当需要反映整个区域的全局模式时，通常采用全局度量指标，习惯上，将全局相关指标称为空间自相关统计。局部相关指标称为局部空间统计。

空间自相关可以分为正相关和负相关。空间自相关是大多数地理过程和空间分布的一个基本特征，在地理应用中通常是正的空间自相关。当属性值与位置无关时，空间自相关为零。当位置相似的观测单元的属性倾向于相似时，存在正的空间自相关；当空间紧密相连的观测单元的属性比更远的属性倾向于更加不相似时，存在负的空间自相关。

空间自相关度量方法大多数情况下可以写成一个标准化的交叉积统计。交叉积统计指出了两个矩阵对应项之间的相关度，一个矩阵确定 n 个位置之间的空间连接，另一个矩阵反映 n 个位置上某一属性变量 X 的值之间明确的相似性定义。

空间自相关分析中所关心的空间目标相当于度量区、统计报告区或采样点，例如点、线、面和网格。同一个空间目标也可以用不同的几何类型描述。

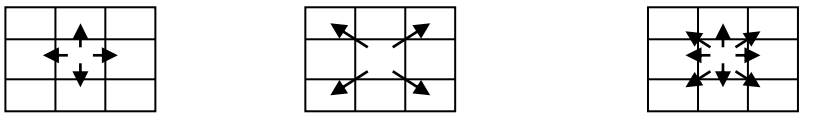
在空间自相关分析中必须确定一些位置邻近关系的度量规则。大多数空间自相关分析遵守的邻近关系是：直接 4 邻域邻近（Rooks）、对角线方向 4 邻域邻近（Bishops）、8 邻域邻近（Queen’s 或 Kings），如图 8-1 所示。对于矢量数据格式或不规则间隔点而言，可以使用从目标单元到 4 个甚至 N 个最邻近单元之间的距离，或者使用一个随机单元 X 和所有邻近单元之间的距离。

任何空间自相关的度量与空间数据的尺度是密切相关的，不同尺度的情况下，空间目标之间的邻近关系会产生明显的变化。

空间自相关的大多数指标通常指定一个二元权重矩阵（$W_{n\times n}$）来表达 n 个目标单元的空间邻近。根据邻接标准，当目标 i 和目标 j 邻接时，空间权重矩阵的元素$\{W_{ij}\}$为 1，否则为 0，所有对角线元素$\{W_{ij}\}$设为 0。如果两个目标不邻接，也可以定义二阶邻接关系，或定义高阶邻接关系。对于面状目标，可用一个标识点表示，例如面的质心。由于空间排列不规则，因此可将权重设置为某一递减函数，例如负幂数（$w_{ij}=d_{ij}^{-b}$）、负指数（$w_{ij}=\exp(-bd_{ij})$，其中，b 可解释为影响权重变化速度的参数）。Cliff 等（1973 年，1981 年）引入了面状目标单元之间潜在相互影响的最一般形式：

$$w_{ij}=[d_{ij}]^{-a}\bullet[\beta_{ij}]^{b}$$

其中，d_{ij} 表示空间目标单元 i 和 j 之间的距离，β_{ij} 为单元 i 与单元 j 公共边界的长度占单元 i 的边界总长度的比例，a 和 b 为参数。也可以建立加权的空间权重矩阵，或对空间权重矩阵进行标准化。



（a）Rooks 情形　　（b）Bishops 情形　　（c）Queen’s(Kings)情形

图 8-1　不同情形的位置邻近关系示意图

三、空间自相关的度量与检验

1、通用交叉积统计

通用交叉积统计可以被看作空间自相关的一般模型，所有交叉积空间自相关的方法都具有一个矩阵交叉积，或者说都具有一个通用交叉积统计，其公式表示如下：

$$\Gamma=\sum_{i}\sum_{j}W_{ij}C_{ij}$$

其中，W_{ij} 矩阵称为连接矩阵、邻近矩阵或空间权重矩阵，矩阵的元素值是度量原始数据中邻近关系的一些方法的函数（例如 Rooks、Bishops 或 Queen’s 情形）；C_{ij} 是对应空间位置（i，j）的数值的邻近性（或距离）的一个度量。

2、Moran’s Ⅰ系数

Moran（1950 年）首先提出了度量空间自相关的方法。目前，几乎在所有涉及空间自相关的研究中都应用 Moran’s I。Moran’s I 是通用交叉积统计的一个特例，Moran 系数（MC）的形式定义如下：

$$MC=\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}}\bullet\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_j-\bar{x})^2/n}$$

其中，w_{ij} 是通用交叉积统计中二元空间权重矩阵 W 的元素，反映了空间目标的位置相似性；x_i、x_j 分别为位置 i 和位置 j 的某一属性值，$\bar{x}$ 为 n 个位置的属性值的平均值。很明显，

分母不能为零，如果为零，那么使用该系数来度量空间自相关是不适当的。只有当（$x_i-\bar{x}$）（$x_j-\bar{x}$）等于零时，该系数才等于零。

Griffith（1987 年）认为，当 *MC*= -1/(*n*-1)时，表示一种随机的地理分布模式；当 *MC*＞-1/(*n*-1)时，且 *MC* 是显著的时，表示相似的属性值倾向于聚集在一起的地理模式（正的空间自相关）；当 *MC*＜－1/(*n*－1)，且 *MC* 是显著的时，表示不同的属性值倾向于聚集在一起（负的空间自相关）。当 *n* 是一个比较大的数值时，*MC* 的期望值收敛于 0。

Moran's I 的统计显著性通常采用随机试验方法。在正态假设条件下，Moran's I 的期望值 $E_N(I)$、方差 $Var_N(I)$ 分别（Goodchild，1986 年）为：

$$E_N(I)=-1/(n-1)$$

$$Var_N(I)=\frac{1}{w_0^2(n^2-1)}(n^2w_1-nw_2+3w_0^2)-E_N^2(I)$$

在随机假设条件下，Moran's I 的期望值 $E_R(I)$、方差 $Var_R(I)$ 分别（Goodchild，1986 年）为：

$$E_R=-1/(n-1)$$

$$Var_R(I)=\frac{n((n^2-3n+3)w_1-nw_2+3w_0^2))-K((n^2-n)w_1-2nw_2+6w_0^2)}{w_0^2(n-1)(n-2)(n-3)}-E_R^2(I)$$

其中，$w_0=\sum_i^n\sum_j^n w_{ij}$ 表示空间权重矩阵的元素值之和；

$w_1=0.5\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij}+w_{ji})^2$，如果权重矩阵是对称的，那么 $w_1=2\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}=2w_0$；

$w_2=\sum_{i=1}^{n}(w_{i\bullet}+w_{\bullet i})^2$，$w_{i\bullet}$ 表示空间权重矩阵第 *i* 行之和，$w_{\bullet i}$ 表示第 *i* 列之和。如果权重矩阵是对称的，那么 $w_2=4\sum_{i=1}^{n}w_{i\bullet}^{\ 2}$；

$K=m_4/m_2^2$ 为样本峰态系数，其中，$m_4=\sum_i^n(x_i-\bar{x})^4/n$，是第四样本矩，

$m_2=\sum_i^n(x_i-\bar{x})^2/n$，是第二样本矩。

Moran's I 的标准差 *SD* 和标准的 *Z*-值的计算如下：

$$SD_N(I)=\sqrt{Var_N(I)}$$
$$Z_N=[I-E_N(I)]/\sqrt{Var_N(I)}$$
$$SD_R(I)=\sqrt{Var_R(I)}$$
$$Z_R=[I-E_R(I)]/\sqrt{Var_R(I)}$$

3、Geary's C 比率

Geary's C 比率（Geary，1954 年）是另一个可以用来度量面状目标和间隔量表数据的空间自相关的全局指标，其形式为：

$$GR=[(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-x_j)^2]/[2(\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij})\sum_{i=1}^{n}(x_i-\bar{x})^2]$$

Geary's C 的期望值为 1。当 *GR*=1 时，表示一种随机的地理分布模式，当 0＜*GR*＜1，且显著时，表示相似的属性倾向于聚集在一起的地理分布模式（正的空间自相关）；当 2＞*GR*＞1，且显著时，表示不同的属性值倾向于聚集在一起（负的空间自相关）。

Geary's C 的显著性检验方法与 Moran's I 的显著性检验方法相同，方程式如下：

$$E_N(C)=1，\ E_R(C)=1；$$

$$Var_N(C)=\frac{1}{2w_0^2(n+1)}((2w_1+w_2)(n-1)-4w_0^2)；$$

$$Var_R(C)=\frac{w_1(n-1)(n^2-3n+3-k(n-1))-\frac{1}{4}w_2(n-1)(n^2+3n-6-k(n^2-n+2))+w_0(n^2-3-k(n-1)^2)}{n(n-2)(n-3)w_0^2}$$

Geary's C 的标准差 *SD* 和标准的 *Z*-值的计算公式与 Moran's I 的相同。

4、局部空间相关的统计方法

局部空间统计相关的计算公式如下：

$$\Gamma_i=\sum_j W_{ij}C_{ij}$$

其中，W_{ij} 和 C_{ij} 分别为矩阵 *W* 和 *C* 的元素。假设有一个划分为 n 个区域单元的地区，每个区域单元由一相关点 *i* 确定，其地理坐标已知，*i*=1，2，…，*n*。每个位置 *i* 与一个值 x_i 相联系，表示随机变量 *X* 的实际观测值，其它位置{*j*}上的变量值表示为{x_j，$j\neq i$}。局部空间相关指标 G_i（*d*）的计算公式如下：

$$G_i(d) = (\sum_{j,j\neq i}^{n} w_{ij}x_j) / \sum_{j,j\neq i}^{n} x_j$$

其中，$\{W_{ij}\}$是一个对称的二元空间权重矩阵，当位置 j 位于位置 i 的某一给定的距离 d 范围内时，空间权重矩阵的元素 W_{ij} 为 1，否则 W_{ij}=0。

空间相关局部指标表示围绕这个观测的相似值的显著性空间集聚程度，空间相关局部指标之和与对应的空间相关全局指标成比例。将单元 i 上观测到的一个变量 x_i 的空间相关局部指标也可以表示为：

$$L_i = f(x_i, \{x_j\})$$

其中，f 是一个函数，$\{x_j\}$是单元 i 的邻近单元的观测值的集合。

作为空间相关局部指标的一个特例，观测单元 i 的局部 Moran 统计可以定义为：

$$I_i = z_i \sum_{j}^{n} w_{ij} z_j$$

式中，z_i 和 z_j 是观测值与均值的偏差，即 $z_i = (x_i - \bar{x}), z_j = (x_j - \bar{x})$。局部 Moran 之和为：

$$\sum_{i}^{n} I_i = \sum_{i}^{n} z_i \sum_{j}^{n} w_{ij} z_j$$

而 Moran's I 是

$$I = (n/S_0) \sum_{i}^{n} \sum_{j}^{n} w_{ij} z_i z_j / \sum_{i}^{n} z_i^2 = \sum_{j}^{n} I_i / [S_0 (\sum_{i}^{n} z_i^2 / n)] = \sum_{j}^{n} I_i / [S_0 m_2]$$

其中，$S_0 = \sum_{i}^{n} \sum_{j}^{n} w_{ij}$，$m_2 = \sum_{i}^{n} z_i^2 / n$ 为第二样本矩。因此，局部 Moran 之和与全局 Moran 的比例因子 γ 表示为：$\gamma = S_0 m_2$。对于一个行标准化空间权重矩阵而言，S_0=n；对于一个标准化的变量而言，m_2=1。由于对所有观测单元而言 m_2 是一个常量，局部 Moran 统计也可以定义为：

$$I_i = (z_i / m_2) \sum_{j}^{n} w_{ij} z_j$$

在一个随机分布假设下，I_i 的期望值、方差分别为：

$$E[I_i] = -w_i / (n-1)$$

$$Var[I_i]=\frac{w_{i(2)}(n-b)}{(n-1)}+\frac{2w_{i(kh)}(2b-n)}{(n-1)(n-2)}-E[I_i]^2$$

其中，$b=m_4/m_2^2$，$m_4=\sum_i^n z_i^4/n$ 表示第四样本矩，$w_i=\sum_j^n w_{ij}$，$w_{i(2)}=\sum_{j,j\neq i}^n w_{ij}{}^2$，

$w_{i(kh)}=\frac{1}{2}\sum_{k,k\neq i}^n\sum_{h,h\neq i}^n w_{ik}w_{ih}$。

I_i 的一个合理的局部 Moran 显著性检验形式为：

$$Z(I_i)=(I_i-E[I_i])/\sqrt{Var[I_i]}$$

观测单元 *i* 的局部 Geary 统计可以定义为：$C_i=\sum_j^n w_{ij}(z_i-z_j)^2$，符号的含义同上。不失一般性，所有观测的 C_i 统计之和为：

$$\sum_i^n C_i=\sum_i^n\sum_j^n w_{ij}(z_i-z_j)^2$$

而 Geary's C 统计为：

$$C=[(n-1)/2nS_0]\bullet[\sum_i^n\sum_j^n w_{ij}(z_i-z_j)^2/m_2]$$

由于对所有观测单元而言 m_2 是一个常量，局部 Geary 统计也可以定义为：

$$C_i=(1/m_2)\sum_j^n w_{ij}(z_i-z_j)^2$$

则有：

$$\sum_i^n C_i=[\sum_i^n\sum_j^n w_{ij}(z_i-z_j)^2]/m_2$$

在一个随机分布假设下，C_i 的期望值、方差分别为：

$$E[C_i]=w_i t_i^2[n/(n-1)]$$

$$Var[C_i]=[(n-1)S_{1i}-w_i^2][E_{2i}-F_i^2]/(n-2)$$

其中，$w_i=\sum_j^n w_{ij}, S_{1i}=\sum_i^n w_{ij}{}^2, t_i^2=[\sum_j^n(z_j-z_i)^2]/(n-1)=F_1, m_r=\sum_i^n z_i^r/n,$

$$E_{2i}=[\sum_{j}^{n}(z_j-z_i)^4]/(n-1)=n[m_4-4z_im_3+6z_i^2m_2+z_i^4]/(n-1)$$

同理，C_i统计的显著性检验为：

$$Z(C_i)=(C_i-E[C_i])/\sqrt{Var[C_i]}$$

第二节　地理空间相关场

一、地理空间相关场的定义

判断在地理空间上两个或多个地理现象是否相关是地理分析中经常遇到的问题，也是进行地理空间推理时必须考虑的问题。郭仁忠（2001 年）、张克权（1991 年）把地理空间相关场分类为：时间域上的地理空间相关场和空间域上的地理空间相关场。

地理空间相关场就是分布在同一个地理区域的不同地理要素之间的关联关系，可以是两个或多个地理要素在区域或时间分布上的相互关系。在地理分析和专题地图制图中常用该方法分析地理要素在区域或时间分布上的相互影响程度，以及绘制相应的专题地图。这种相互关系模型可以分为：两个指标之间的相互关系模型、多个指标之间的相互关系模型。根据地理要素之间的关系和观测数据的类型，可以建立不同的空间相关模型，得到多种相关系数。

地理空间相关可以从不同的空间抽象层次上进行描述，例如：局部范围内，可以计算一个点与邻域的相关性，或者两个相邻点或区域的相关性；一个地理区域与其上一个更抽象的地理区域的相关性。从时序上，可以计算不同时间段之间两种地理要素之间的相关性，也可以计算一个地理要素在不同时间点之间的相关性。

二、单相关系数

单相关系数用于表达两种地理现象之间在空间上的线性相关程度。假设在同一个空间上按照一定的规律分布了 n 个抽样点，每个点上两种地理现象的观测值分别 x_i，y_i，那么这两种地理现象之间的空间相关系数为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{ns_xs_y}$$

其中，$\bar{x}$、$\bar{y}$ 分别为两种地理现象的观测值的平均值；s_x、s_y 分别为两种地理现象的观测值的标准差；$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$；$s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$；$\sum_{i=1}^{n}(x_i-\bar{x})^2$、

$\sum_{i=1}^{n}(y_i-\bar{y})^2$ 分别为两种地理现象的观测值的离差平方和。r 的值在[-1，1]之间，当$|r|\geqslant 0.7$时，可以认为这两种地理现象有实质性的线性关系。

二、偏相关系数

偏相关系数也称为局部相关系数，用于表达在多种相互有影响的地理现象中两种地理现象之间的相关程度，该系数删除了其他地理现象的影响。这里，假设有 3 种相关的地理现象 *A*、*B* 和 *C*，那么，地理现象 *A* 和 *B* 的偏相关系数为：

$$r_{AB/C}=\frac{r_{AB}-r_{AC}r_{BC}}{\sqrt{(1-r_{AB}^2)(1-r_{BC}^2)}}$$

其中，r_{AB}、r_{AC} 和 r_{BC} 为单相关系数。

三、复相关系数

复相关系数也称为全相关系数，用于说明多种地理现象对一种地理现象的影响程度，这里，假设有 3 种相关的地理现象 *A*、*B* 和 *C*，那么，地理现象 *B* 和 *C* 对 *A* 的复相关系数为：

$$r_{A\bullet BC}=\sqrt{\frac{r_{AB}^2+r_{AC}^2-2r_{AB}r_{AC}r_{BC}}{1-r_{BC}^2}}$$

其中，r_{AB}、r_{AC} 和 r_{BC} 分别为 *A* 和 *B* 的单相关系数，*A* 和 *C* 单相关系数，*B* 和 *C* 单相关系数。

四、秩相关系数

秩相关系数用于处理观测值为等级数据时的两种地理现象的相关程度，常用的方法有 Spearman 方法和 Kendall 方法。

Spearman 方法的等级相关系数的计算公式为：

$$S_r=1-\frac{6d^2}{n^3-n}$$

其中，*n* 为样本的总数或地理区域内子区域的总数（例如，统计地图的行政区域总数、栅格单元的个数等），$d=\sum_{i=1}^{n}(p_{ai}-p_{bi})^2$，$p_{ai}$ 和 p_{bi} 分别为地理现象 *A* 和 *B* 的等级编号。S_r 在[-1，1]之间。

假设一个地区有 *n* 个地理子区域，有两幅分别描述同一个地区的两种地理要素（*A* 和

B）的地图，*A* 和 *B* 的数据分别用 5 个等级和 4 个等级来描述。每个子区域的原始等级如表 8-1 所示，这里，假设 *n*=6。按照 *n* 个地理子区域的编号，若每个地理子区域属于不同的等级，*A* 和 *B* 的数据都是如此，那么就可以直接使用前面的公式进行计算。但是，在实际应用中，等级数会小于 *n*，并且，*A* 和 *B* 的等级数不同，如表 8-1 所示，那么，就必须调整 *n* 个地理子区域的等级，使地理子区域的等级在[1，*n*]之间，但是等级数不变，调整的方法是：当几个地理子区域的等级相同时，用平均等级编号表示。可以从应用的角度出发调整等级的顺序，同时尽量保证 *A* 和 *B* 的等级的数值从小到大，例如：1-2、2-4、4-6。对于表 8-1 而言，6 个地理子区域的 *A* 要素的等级调整过程如下：

表 8-1　Spearman 方法中的等级调整

区域编号	1	2	3	4	5	6
A 的原始等级	1	4	2	2	3	5
B 的原始等级	1	3	2	1	2	4
A 的等级 p_{ai}	6	2	4.5	4.5	3	1
B 的等级 p_{bi}	5.5	2	3.5	5.5	3.5	1

A 的原始等级序列为：1、4、2、2、3、5
调整等级的顺序为：5、2、4、4、3、1
第 1 步转换：6、2、4(5)、4(4)、3、1
第 2 步转换：6、2、4.5、4.5、3、1
无相同等级，结束转换。

调整过程的说明：先把等级序列按从大到小排列，然后把最高等级定义为区域总个数 *n*，因此在第 1 步把第 1 个单元的等级变为 6。在第 1 步转换时，第 3 个和第 4 个地理子区域的等级相同，都为 4，若不同时，应当分别是 5 和 4，因此按照调整方法，在第 2 步转换时就转换为（5+4）/2。

6 个地理子区域的 *B* 要素的等级调整过程如下：

B 的原始等级序列为：1、3、2、1、2、4
调整等级的顺序为：4、2、3、4、3、1
第 1 步转换：4(6)、2、3、4(5)、3、1
第 2 步转换：5.5、2、3、5.5、3、1
第 3 步转换：5.5、2、3(4)、5.5、3(3)、1
第 4 步转换：5.5、2、3.5、5.5、3.5、1
无相同等级，结束转换。

调整过程的说明：在第 1 步转换时，第 1 个和第 4 个地理子区域的等级相同，都为 4，若不同时，应当分别是 6 和 5，那么，在第 2 步转换时就转换为（5+6）/2。在第 3 步转换时，第 3 个和第 5 个地理子区域的等级相同，若不同时，应当分别是 4 和 3，那么，在第 4 步

转换时就转换为（4+3）/2。利用转换后的等级编号就可以计算 S_r 的值。

在 Spearman 方法中，地理子区域的等级在[1，n]之间，若 A 和 B 的地理子区域编号按升序（或降序）排列，同时它们的等级编号也是按升序（或降序）排列，那么相关程度最好。但是，现实中，它们的等级编号顺序是变化的，从前面的例子就可以看到这一点。也就是说，地理子区域的等级编号顺序的变化决定了 A 和 B 的相关程度。

Kendall 方法的等级相关系数的计算公式为：

$$K_r = \frac{2T}{n(n-1)}，其中，T = \sum x_{ik} y_{ik}$$

这里，同样假设一个地区有 n 个地理子区域，有两幅分别描述同一个地区的两种地理要素（A 和 B）的地图，A 和 B 的数据分别用 5 个等级和 4 个等级来描述。每个子区域的原始等级和调整后的等级如表 8-1 所示，这里，假设 n=6。

理论上，它们的等级编号是按降序排列，但是，它们现在的排列分别为：6、2、4.5、4.5、3、1；5.5、2、3.5、5.5、3.5、1。为了描述它们的排列顺序的变化，可以对它们的等级编号进行一一对比。例如，设序列{6、2、4.5、4.5、3、1}={$x_1, x_2, x_3, x_4, x_5, x_6$}，每对等级编码的比较结果可以表达为：

$$x_{ik} = \begin{cases} +1; & x_i > x_k \\ 0; & x_i \equiv x_k \\ -1; & x_i < x_k \end{cases}$$

其中，$i<k$，也就是说，比较结果为：{ $x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{23}, x_{24}, x_{25}, x_{26}, x_{34}, x_{35}, x_{36}, x_{45}, x_{46}, x_{56}$ }={1，1，1，1，1，-1，-1，-1，1，0，1，1，1，1，1}。同样道理，设序列{5.5、2、3.5、5.5、3.5、1}={$y_1, y_2, y_3, y_4, y_5, y_6$}，每对等级编码的比较结果可以表达为：

$$y_{ik} = \begin{cases} +1; & y_i > y_k \\ 0; & y_i \equiv y_k \\ -1; & y_i < y_k \end{cases}$$

比较结果为：{ y_{12}，y_{13}，y_{14}，y_{15}，y_{16}，y_{23}，y_{24}，y_{25}，y_{26}，y_{34}，y_{35}，y_{36}，y_{45}，y_{46}，y_{56}}={1，1，0，1，1，-1，-1，-1，1，-1，0，1，1，1，1}。

T=1×1+1×1+1×0+1×1+1×1+（-1）×（-1）+（-1）×（-1）+（-1）×（-1）+1×1+0×（-1）+1×0+1×1+1×1+1×1+1×1+1×1=1+1+0+1+1+1+1+1+1+0+0+1+1+1+1=12。

K_r=(2×12)/(6×5)=0.8。

五、多成分标志

在进行两种地理要素的叠置分析时，两种要素的值属于间隔量表或名义量表，若需要计算它们之间的相关系数，可以采用多成分标志来描述。先对描述该两种要素的地图（空间数据）进行栅格化处理，然后，用一个矩阵来描述两种要素在不同等级上所共有的栅格数（频数），如表 8-2 所示，假设有两幅分别描述同一个地区的两种地理要素（A 和 B）的地图，A 和 B 的数据分别分为 a_1、a_2、a_3 和 b_1、b_2 几类。F_{ij} 表示既是 b_j 又是在 a_i 的栅格数，

k_a、k_b 分别为 A 和 B 的数据的等级数。$n_{bj}=\sum_{i=1}^{k_a}F_{ij}$，$n_{ai}=\sum_{j=1}^{k_b}F_{ij}$。多成分标志的计算公式为：

$$p=\sqrt{\frac{T-1-\frac{(k_a-1)(k_b-1)}{n}}{\sqrt{(k_a-1)(k_b-1)}}}$$

其中，$T=\sum_{i=1}^{k_a}(\frac{1}{n_{ai}}\sum_{j=1}^{k_b}\frac{F_{ij}^2}{n_{bj}})$；

表 8-2　在不同等级上所共有的栅格数的计算方法

		要素 A			
		a_1	a_2	a_3	n_{bj}
要素 B	b_1	F_{11}	F_{21}	F_{31}	$n_{bj}=\sum_{i=1}^{k_a}F_{ij}$
	b_2	F_{12}	F_{22}	F_{32}	
	n_{ai}	$n_{ai}=\sum_{j=1}^{k_b}F_{ij}$			

六、四分相关系数

若同一个地理区域有两种地理要素（A 和 B）只需要以名义量表来描述（例如范围法表示的地图），那么这两种地理要素之间的相关程度可以用四分相关系数表示。四分相关系数为：

$$R_{++}=a/\sqrt{(a+b)(a+c)}$$

其中，a 是两种地理要素都存在的范围的大小；b 是 A 存在而 B 不存在的范围的大小；c 是 A 不存在而 B 存在的范围的大小。

七、基于信息熵的相关系数

若同一个地理区域有两种地理要素（A 和 B）以名义量表或比率量表来描述，那么这两种地理要素的相关系数可以表示为：

$$K=\frac{H(A)+H(B)-H(AB)}{H(AB)}$$

其中，$H(A)$、$H(B)$分别为两种地理要素（A 和 B）的独立事件信息熵；$H(AB)$为两种地理要素（A 和 B）同时存在的信息熵。K 的范围是[0，1]。

若现象 A 和 B 分别表示为 n 和 m 个级别，那么现象 A 和现象 B 的不同等级（第 i 级和第 j 级）所占的面积与区域总面积的比率分别为 W_{ai}、W_{bj}，现象 A 的第 i 等级和现象 B 的第 j 等级同时出现的面积与区域总面积的比率为 W_{abij}，其中 i、j 的范围是[1，n]和[1，m]。这两种地理要素的相关系数可以进一步表示为：

$$K=\frac{\sum_{i=1}^{n}W_{ai}\log_2 W_{ai}+\sum_{i=1}^{m}W_{bi}\log_2 W_{bi}-\sum_{i=1}^{n}\sum_{j=1}^{m}W_{abij}\log_2 W_{abij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}W_{abij}\log_2 W_{abij}}$$

当 $n=m$ 时，该公式的计算结果比较理想，否则，就需要进行修正。因为当 $n\neq m$ 时，K 有一个最大值的限制，该值为：

$$K_{\max}=\frac{H(A)_{\max}+H(B)_{\max}-H(AB)_{\min}}{H(AB)_{\min}}$$

其中，$H(A)_{\max}=\log_2 n$，$H(B)_{\max}=\log_2 m$，

$$H(AB)_{\min}\approx-(n\times\frac{1}{m}\log_2\frac{1}{m}+n\times\frac{m-n}{m}\times\frac{1}{n}\log_2\frac{m-n}{m}\times\frac{1}{n})。$$

所以，修正后的 K 为：$K'=\frac{K}{K_{\max}}$

第三节　空间自相关在统计地图数据分级中的应用

统计数据分级是专题地图综合中的一个重要问题，它的目的是在尽可能少地丢失原始信息的基础上，将大量的观察数据进行归纳合并，并且保证数据分级后在统计分布和空间分布上尽可能反映出现象的本质。国内外已提出了很多种分级方法和分级质量评价指标，在质量评价方面，目前主要以视觉变量和空间认知为基础，在保证分级视觉效果的前提下，根据统计学原理对分级的数据精度进行评价，很少考虑数据的空间分布特征。Armstrong M. P.，Xiao N.和 Bennett D. A.（2004 年）把遗传算法用于统计地图的数据分级，分析了多个质量评价指标两两结合用于评价分级质量的方法，以便找到多标准下的最优解，他们也认为在分级中应当考虑数据的地理特征。当然统计数据千差万别，地理特征很多，但是在分级过程中，数据的空间分布规律是必需考虑的，也就是说我们不仅要考虑数据的统计精度、图面视觉效果和空间认知，还要考虑分级区域破碎程度和空间自相关程度等。

一、分级质量评价指标

一、分级质量评价指标

统计制图中数据分级的主要原则包括：统计学方面的要求，即保持数据的主要统计特征；制图学方面的要求，即尽量保持原始数据的空间分布特征；增强地图信息的传输效率。从人的视觉效果和空间认知能力来看，一般认为 4～7 级是合适的分级数。那么，具体哪一个分级数比较合理？分级界线如何确定？这就需要通过相应的评价指标来确定。陆效中（1989 年）提出了 6 个评价数据分级精度的指标，指标值越大，精度越高。这些指标是总偏差误差 EC_1、加权总偏差误差 EC_2、平均偏差误差 EC_3、加权平均偏差误差 EC_4、综合误差 *SumEC* 和分级匹配精误差 *MEC*。

它们的计算公式依次为表达式：

$$EC_1=\frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}\left|x_{ij}-\overline{x_j}\right|}{\sum_{i=1}^{N}\left|x_i-\bar{x}\right|} \quad ; \quad EC_2=\frac{\sum_{j=1}^{k}\frac{1}{\overline{x_j}}\sum_{i=1}^{N_j}\left|x_{ij}-\overline{x_j}\right|}{\frac{1}{\bar{x}}\sum_{i=1}^{N}\left|x_i-\bar{x}\right|}$$

$$EC_3=\frac{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{N_j}\sum_{a=1}^{N_j}\left|x_{ij}-\overline{x_j}\right|}{\frac{1}{N}\sum_{a=1}^{N}\left|x_i-\bar{x}\right|} \quad ; \quad EC_4=\frac{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{\overline{x_j}N_j}\sum_{a=1}^{N_j}\left|x_{ij}-\overline{x_j}\right|}{\frac{1}{\bar{x}N}\sum_{a=1}^{N}\left|x_i-\bar{x}\right|}$$

$$SumEC=(EC_1+EC_2+EC_3+EC_4)/4$$

$$MEC=(\sum_{j=1}^{k}\left|x_j^{'}-\overline{x_j}\right|/x_j^{'})/k$$

其中，k 为分级数，N 为数据总数，N_j 为第 j 级中的数据的个数，x_i 为原始数据中第 i 个数据的值，x_{ij} 为 j 级第 i 个数据的值，$\bar{x}$ 为所有数据的平均值，$\overline{x_j}$ 为第 j 中级数据的平均值，$x_j^{'}$ 为第 j 级的中点值。

这些指标的作用各不相同，EC_1 越小说明各等级内部统计数据越均匀一致；对于分布范围较大的数据，各级数据的偏差大小起的作用不一样，EC_2 很好地反映了偏差的相对大小；数据很多时，聚集在同一级内的许多数据就会产生较大的误差积累，这时用 EC_1 的计算精度较低，EC_3 能很好反映出分级精度在这方面的要求；EC_4 反映数据平均偏差的相对大小；*SumEC* 是前面 4 种分级误差的平均值；一组数据一经确定分级之后，各级的数据平均值反映了各级的数据重心，各级中点反映了各个等级的重心，分级系统中这两种重心越吻合，表明分级效果越好，MEC 反映这两种重心的匹配误差。

原始数据本身就有潜在的空间分布规律，这种空间分布特征在地理研究中是需要考虑的重要因素。对于以区域为统计单位的数据而言，相邻区域之间的数据差别和数据的空间自相关程度都是用来描述空间分布特征的非常重要的指标。数据分级后，区域的属性值就

被人为抽象了，相邻区域之间的差别就会发生变化，同样道理，空间自相关程度也会变化。也有学者认为，数据分级后，各级所占的区域面积应当符合一定的统计规律，只有当统计数据与统计单元的面积密切相关时，才需要考虑，例如，符合正态分布规律；各级所占的区域面积应当尽量相等，以便读图者能快速的从地图上获取数据在空间分布上的差异，也可以从视觉上达到一种平衡。但是，正态分布规律的保持需要足够数量的样本数，统计数据所关联的区域面积的统计规律有时也并不符合正态分布规律，作者认为，在选择分级方法时需要考虑这个问题。下面研究有关空间分布特征的评价指标。

1、区域边界误差 *BE*（Boundary Error）

该参数说明了数据分级后，新的不同级别之间两边的数据差异变化情况。理论上，空间上相邻的数据如果差别不大，就应当归为一级，数据分级后，区域边界两边的统计值就发生了变化，我们当然希望这种变化越小越好，其表达式如下：

$$BE = 1 - \frac{\sum_{i=1}^{\|H\|} \sum_{(r,l)\in H} \left| x_{ir} - x_{il} \right|}{\sum_{i=1}^{\|H\|} \sum_{(r,l)\in G} \left| x_{ir} - x_{il} \right|}$$

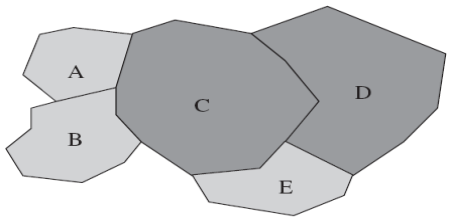


图 8-2　拓扑关系示意图

其中，G 为空间上相邻的区域单元对，如图 8-2 所示，区域单元为 A、B、C、D 和 E，其中 C、D 为一级，A、B、E 为一级，则 G 为（A，B）、（A，C）、（B，C）、（C，D）、（C，E）、（D，E）；H 为属于不同级别的相邻单元对，对于图 8-2 而言，H 为（A，C）、（B，C）、（C，E）、（D，E）；$\|H\|$ 为 H 的个数，图 8-2 中的 $\|H\|$ 为 4；l、r 分别表示相邻单元对的左单元、右单元；x_{il} 表示不同级别的相邻单元左边单元值，x_{ir} 表示不同级别的相邻单元右边单元值。*BE* 的取值范围为[0，1]，在分级数确定的情况下，*BE* 越小，表明在地理空间上相邻的不同级别的区域单元差别越大，该分级方案很好地反映了事物在地理空间分布上的实际差异，分级效果越好。

2、地理面积均等程度 *GAE*(Geographical Area Equalization)

不同级别之间的地理面积均等程度是从视觉重量感的观点出发，用于检测各级区域的不一致性。该指标表达式如下：

$$GAE = \max\left|\frac{i}{k} - \frac{1}{A}\sum_{j=1}^{i} A_j\right| \qquad i = 1，\cdots，k$$

其中k为分级数，A为所有级别的区域面积之和，A_1，…，A_k按升序排列，A_j为第j级所包含的区域面积总和。*GAE*的取值范围为[0，1]，*GAE*越小，各级区域面积越均衡一致，分级的视觉效果越好。

3、空间自相关系数*MIC*(Moran's I statistic Coefficient)

空间自相关系数是检测邻近单元相似性的重要指标，可以用于检验空间变量的取值是否与相邻空间上该变量取值大小有关。*MIC*取值范围为[-1，1]，当*MIC*=0时，代表空间不相关，*MIC*取值为正数时，表明空间变量在一点上的取值与相邻点的取值变化趋势相同，被称为空间正相关，相反则被称为空间负相关。空间自相关分析首先要对所检验的空间单元进行配对和采样，本文对直接相邻的群进行配对，全部采样。

$$MIC = \frac{M\sum_{(i,j)\in C}(x_i - \bar{x})(x_j - \bar{x})}{\|P\|\sum_{i=1}^{M}(x_i - \bar{x})^2}$$

其中，群为属于同一级的邻近的单元集（在地图上表现为同一级相邻接的单元所构成的多边形），图8-2中，A和B为一个群，C和D为一个群，E为一个群，M为群的个数，P为邻近群对的集合，$\|P\|$为*P*的个数，$\bar{x}$为所有群的平均值，x_i、x_j分别为两相邻的群*i*、群*j*对应的值或所在分级的均值。Armstrong M. P.等（2003年）认为，为了使各评价指数的取值范围保持一致，以方便比较与计算，规定*MIC*1=(1-*MIC*)/2，这里*MIC*1的取值范围为[0，1]，MIC_1越小表明级间空间自相关越大。但是，原始数据本身就存在着空间自相关性，若一味追求该值的最小化，就会与原始数据的空间自相关情况相违背，因此尽量保持分级后数据的空间自相关程度与原始数据空间自相关程度一致，才是合理的。该指标对于划分为一个等级的相邻区域集合的局域空间自相关的评价是比较合理的。下面分析*MIC*1与分级数之间的关系。

二、*MIC1*与分级数之间的变化规律

以某地区的31个县人口数为例，见表8-3，为了简化原始空间数据，这里以图8-3的子区域为统计单位，图中的号码代表子区域的统计序号。

采用系统聚类法将表8-3中的数据逐步从31级归类为1级，分级数从1到31的*MIC*1的值依次对应为：0、0.92747、0.59646、0.59646、0.4667、0.4341、0.373、0.3245、0.3245、0.3245、0.335、0.3281、0.3281、0.3251、0.3251、0.3171、0.3171、0.3171、0.3171、0.3171、0.3658、0.3171、0.3189、0.3084、0.3992、0.3004、0.3004、0.3004、0.275、0.2836、0.2836，分级数所对应MIC_1的变化趋势如图8-4所示。

表8-3　统计数据表

序号	数据	序号	数据	序号	数据	序号	数据
1	202083	9	346484	17	479824	25	588006
2	232678	10	354847	18	500899	26	609555
3	264514	11	357756	19	503124	27	610998
4	280498	12	369194	20	512166	28	820142
5	301779	13	373105	21	517093	29	951806
6	312889	14	374483	22	534594	30	958281
7	320982	15	390706	23	541679	31	1016065
8	343486	16	473305	24	561782		

3	2	4	27	31	29
1	6	28	26	17	30
11	12	25	16	19	18
13	23	20	21	8	7
5	15	14	24	22	9
10					

图 8-3　统计单元

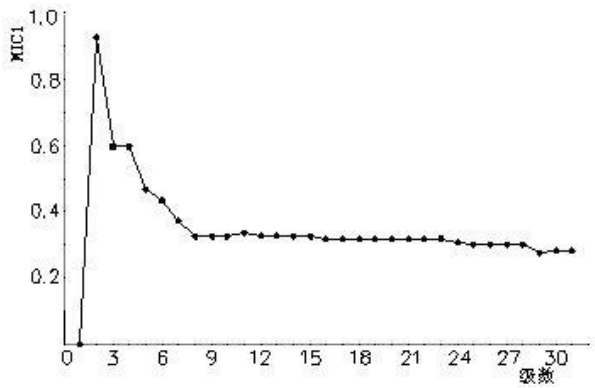


图 8-4　分级数与 *MIC*1 的关系

从图 8-4 可以看出：级数为 1，即整个区域分为一级时，*MIC*1 取最小值 0，对应空间自相关系数 *MIC* 取最大值 1，若只单独考虑该指标的最小化，那么可以认为，从分为 1 级的效果最好，但显然无意义。从分级数为 2 到分级数为 8，*MIC*1 迅速变小，从分级数为 8 到分级数为 31 变化幅度很小，实际上，分级数为 31 时的 *MIC*1 就说明了原始数据的空间自相关系数。同时，分级数为 2、3、4 时，*MIC* 为负值，是空间负相关，应该排除，分级数为 8 级以后，*MIC*1 变化很小，因此，从空间自相关来看，5～8 级是合适的。

三、综合性评价

统计制图中数据分级的方法很多，这里选择八种方法：最优分割法、逐步聚类分级法、模糊聚类分级法、逐步模糊模式识别分级法、任意数列分级、算术级数分级、几何级数分级、任意级数分级，以上各方法均分为 7 级，对表 8-3 中的数据进行分级的结果如表 8-4 所示，表 8-5 列出了各分级方法所得分级结果的 9 个评价指标值。

表 8-4　不同方法的分级结果

方法	等级	分级界线	数据个数	方法	分级界线	数据个数
最优分割分级	1	200000－250000	2	逐步聚类分级法	200000－250000	2
	2	250000－290000	2		250000－330000	5
	3	290000－330000	3		330000－430000	8
	4	330000－430000	8		430000－530000	6
	5	430000－540000	7		530000－720000	6
	6	540000－700000	5		720000－890000	1
	7	700000－1020000	4		890000－1020000	3
模糊聚类法	1	200000－250000	2	逐步模糊识别法	200000－270000	3
	2	250000－330000	5		270000－330000	4
	3	330000－430000	8		330000－380000	7
	4	430000－490000	2		380000－430000	1
	5	490000－720000	10		490000－520000	6
	6	720000－890000	1		520000－720000	6
	7	890000－1020000	3		720000－1020000	4
任意数列分级	1	200000－240000	2	算术级数分级	200000－250000	2
	2	240000－290000	2		250000－330000	5
	3	290000－340000	3		330000－420000	8
	4	340000－420000	8		420000－540000	7
	5	420000－540000	7		540000－670000	5
	6	540000－720000	5		670000－830000	1
	7	720000－1020000	4		830000－1020000	3
几何级数分级	1	200000－270000	3	任意级数分级	200000－250000	2
	2	270000－340000	4		250000－330000	5
	3	340000－430000	8		330000－430000	8
	4	430000－540000	7		430000－550000	8
	5	540000－670000	5		550000－690000	4
	6	670000－830000	1		690000－850000	1
	7	830000－1020000	3		830000－1020000	3

表 8-5　分级评价指标

分级方法	*GAE*	*BE*	*MIC*1	EC_1	EC_2	EC_3	EC_4	*SumEC*	*MCE*
取值范围	[0，1]	[0，1]	[0，1]	[0，1]	[0，1]	[0，1]	[0，1]	[0，1]	[0，1]
逐步聚类	0.2212	0.1060	0.3730	0.1133	0.1265	0.1041	0.1194	0.1158	0.0373
模糊聚类	0.2212	0.1659	0.3478	0.1334	0.1442	0.1008	0.1167	0.1237	0.0388
任意数列	0.3456	0.2028	0.3678	0.1316	0.1229	0.1294	0.1294	0.1283	0.0427
算术级数	0.2212	0.1567	0.3678	0.1125	0.1259	0.1050	0.1200	0.1159	0.0464
几何级数	0.2028	0.1567	0.3714	0.1103	0.1239	0.1047	0.1191	0.1145	0.0422
任意级数	0.2212	0.1705	0.3608	0.1129	0.1267	0.1050	0.1199	0.1161	0.0421
最优分割分级	0.3456	0.2028	0.3992	0.1316	0.1229	0.1294	0.1294	0.1283	0.0407
逐步模式识别分级	0.2304	0.0876	0.4115	0.1359	0.1326	0.1297	0.1308	0.1323	0.0391

前面已分析了各个评价指标的作用，除 *MIC*1 外，其它的 8 个指标都是值越小越好，同时 EC_1、EC_2、EC_3、EC_4 这四个指标与 *SumEC* 有关联，*SumEC* 是它们的平均值。当 *MIC* 只是局域空间自相关系数时，也可以认为 *MIC*1 的值越小越好。从理论上讲，能保证这些评价指标同时最小的分级无疑是最好的分级方法，然而，由于这些指标之间的相互制约性（例如，追求各级多边形的面积均衡，就很难保证各级内多边形群的邻近性），这样的分

级实际上是不存在的。从表 8-5 来看，根据指标 *GAE* 进行权衡，几何级数分级的效果最好；根据 *BE* 指标权衡，逐步模式识别分级法的效果最好。然而，为了保证分级的整体效果最好，我们就需要兼顾多个指标，综合性评价公式如下：

$$A = a_1 \times GEA + a_2 \times BE + a_3 \times MIC1 + a_4 \times EC1 + a_5 \times EC2 + a_6 \times EC3 + a_7 \times EC4 + a_8 \times SumEC5 + a_9 \times MEC$$

其中，a_1、a_2、a_3、a_4、a_5、a_6、a_7、a_8、a_9 为各指标对应的权重，且 $\sum_{i=1}^{9} a_i = 1$，a_i 取值范围为[0，1]，其取值可根据制图目的、要求并结合经验进行调整。例如，若强调空间分布规律，则可以取值 a_2=a_3=0.3，a_1=a_4=a_5=a_6=a_7，a_8=a_9=0.2，用户可以通过人机交互的方式得到满意的结果。

实际操作中，制图者可以在适当兼顾其它指标值相对小的情况下，选择自己特别关注的指标值，获得满意的分级结果，这种评价方法就为制图者提供了较多选择，以便找到更适合的分级。也可以把 *MIC*1 单独列出，在统计误差评价方面甚至可以只考虑 *SumEC* 和 *MEC*，那么评价指标公式就变为：

$$\begin{cases} \left| MIC_k - MIC_n \right| \le \varepsilon \\ Min(SumEC) = a_1 \times GEA + a_2 \times BE + a_8 \times SunEC + a_9 \times MEC \end{cases}$$

其中，MIC_k 是分为 k 级后的空间自相关系数，MIC_n 是 n 个原始数据的空间自相关系数，ε 为给定的阈值，a_1＋a_2＋a_8＋a_9＝1.0。