

空间聚类分析可以探测事件在时间、空间上集聚或分布的非随机性。事件的非随机性表明其空间分布存在着自相关性，而空间自相关性又要求在回归分析中利用特有的空间回归法。空间聚类分析和空间回归问题早在几十年前就已经提出来了，但由于其计算量较大，最初应用十分有限。随着计算能力的提高，特别是 GIS 技术的广泛运用，这方面的应用软件（包括一些免费软件）的开发也发展很快，从而大大激发了人们的研究兴趣，拓展了这类方法的应用领域。本章将讨论空间聚类分析和空间回归，介绍相关空间分析软件包的使用方法。

空间聚类分析在犯罪和健康研究中应用十分广泛。在犯罪研究方面，就是通常所说的“热点区（hot-spot）”分析。犯罪学家的研究表明，犯罪活动在空间上存在局部集中，形成“热点区”，其特征有：（1）明显表现在某些犯罪类型上，如毒品交易(Weisburd and Green, 1995)；（2）分布在某些特定地段，如贫民窟、酒吧；（3）在某些地段某类犯罪行为呈现出高峰值，如公交车站或交通中转站的盗窃行为(Block, 1995)。分析“热点区”对于警察和其他反犯罪机构大有益处，有助于他们将目标锁定在有限的区域之内。空间聚类另一方面的应用是与健康相关的研究。某种疾病是否在空间上表现出某种集聚态势？它的发病率在哪些地方高、哪些地方低？一些地区某种疾病的发病率高，可能是正常的随机波动引起的，并无一定的成因。一般只有当高发病率在空间分布上具有统计显著性时，才有一定的研究意义(Jacquez, 1998)。因此，空间聚类分析是许多探索性研究中基本的又十分有效的第一步，如果研究发现某种疾病确实存在着空间集聚，那就需要更深入细致地调查、抽样观测和开展疾控工作。

空间聚类分析可以分为基于点和基于面两种方法。基于点的方法需要事件准确的地理位置，基于面的方法用的是地区内平均发病率。到底用哪种方法，关键取决于数据，基于点的方法并不总是优于基于面的方法(Oden et al., 1996)。本章的第 9.1 节讨论基于点的空间聚类分析理论，第 9.2 节是相应的案例，分析中国南部地区泰语地名的空间分布特征。第 9.3 节讲解基于面的空间聚类分析理论，第 9.4 节中给出了相应的案例，分析伊利诺伊州各种癌症的空间分布

。当前 ArcGIS 软件中的空间统计分析功能还很有限，只能实现基于面的空间聚类分析，其它软件如 CrimeStat (Levine, 2002)也具有类似的功能。我们利用 SaTScan 软件包实现基于点的空间聚类分析。第 9.5 节介绍空间回归方法，第 9.6 节介绍在 GeoDa 软件包上如何实现，并通过研究芝加哥市谋杀犯罪率的空间差异来讲解空间回归法的应用。第 9.7 节为本章小结。

除 ArcGIS 软件外，SaTScan 和 GeoDa 软件包都可以免费下载。具体的空间聚类分析和空间回归分析方法还很多，本章仅仅通过三个案例，介绍一些常用的方法。

9.1 基于点的空间聚类分析

基于点的空间聚类方法可以归结为两大类，即全局聚类检验 (tests for global clustering) 和局部聚类检验 (tests for local clusters)。

9.1.1 基于点的全局聚类检验

全局聚类检验用于分析研究对象在整个区域内是否具有空间集聚性。将所有观察个体区分为事件和非事件两类（如疾病研究中，事件就是病例，非事件就是未患病的个体）。以怀特莫等(Whittemore et al., 1987)提出的全局聚类检验指标为例，先计算事件之间的平均距离，再计算所有个体之间的平均距离。如果前者比后者低，则表明事件在空间上存在集聚。当研究区的中心地区具有丰富的病例资料时，这个方法比较有效，但如果病例分散在外围地区，则此方法效果不佳(Kulldorff, 1998, 53 页)。另有学者(Cuzick and Edwards, 1990)提出的方法是，针对每个病例搜寻其邻近的 k 个样本，然后检验这 k 个样本中的病例数是否比随机分布状态下的病例更多。其他人(Diggle and Chetwynd, 1991; Grimson and Rose, 1991)也创立了一些基于点的全局聚类检验方法，这里就不介绍了。

9.1.2 基于点的局部聚类检验

对于大多数研究而言，确定空间集聚的具体位置或局部集聚也是十分重要的。研究区即使在全局聚类检验中没有统计显著性，也有可能存在着局部集聚的现象。

欧本休等(Openshaw et al., 1987)开发了“地理分析机”(Geographical Analysis Machine 或简称 GAM)。其分析方法是，首先在研究区中生成网格点，然后以网格点为中心划不同半径的圆

，最后寻找所含病例显著的圆，称集聚圈。GAM 方法的一个不足之处是，找到的集聚圈数量偏高，也就是含有一些“假集聚圈”(Fotheringham and Zhan, 1996)，原因是找到的集聚圈重叠性高，圈内所含的部分病例相同，样本互不独立，不适合用泊松分布（要求样本是相互独立的）来检验统计显著性。

169

比撒格和纽维尔(Besag and Newell, 1991)的方法，是在病例周围查找是否存在集聚。假如 k 为局部聚类最小病例数，模型首先从每个病例点出发，识别邻域中含有 $k-1$ 个病例的区域（不包括中心点的病例），然后分析这些区域中的病例总数是否明显地高于其期望值。通常 k 值在 3 到 6 之间，也可先用不同 k 值进行敏感度分析，再选择合适的 k 值。与 GAM 方法类似，这种方法找到的集聚圈也会重叠，但它产生“假集聚圈”的可能性比 GAM 要小，同时计算强度也小一些(Cromley and McLafferty, 2002, 153 页)。还有其他人（如 Rushton and Lolonis, 1996）也提出了一些基于点的空间聚类分析方法，这里就不介绍了。

下面讨论库多夫(Kulldorff, 1997)提出的空间扫描统计法(spatial scan statistic)及其相关的 SaTSCa 软件。SaTScan 是免费软件，由 Kulldorff 的研究组在美国卫生部的资助下开发完成，下载和使用说明请参考网址 <http://www.satscan.org>。主要用来分析疾病在空间上或时空上的集聚分布，检验其分布是否具有统计显著性。

与 GAM 相似，空间扫描统计法使用圆作为扫描窗口，搜索整个研究区，但它避免了上述的“假集聚圈”问题。扫描窗口半径大小的选取，以圈内样本数占总样本数的比例来确定，从 0%到 50%逐步上升。针对每个圆，比较窗口内和窗口外的发病风险，寻找窗内风险统计上明显高的圈，定义为空间集聚。空间扫描统计法使用泊松(Poisson)分布或伯努利(Bernoulli)分布来判断统计显著性。如果可感人群的数据是基于面的（如各地区的总人数），则选用泊松分布。需要输入的数据是各地区病例和人口总数，以及这些地区（以中心点代表）的座标。如果是二项分布的数据（即病例与非病例的个体数据），则选用伯努利模型，它要求所有样本的地理坐标，是病例的样本记为 1，非病例的样本记为 0。

例如，在伯努利模型中窗口 z 的似然函数(likelihood function)计算如下：

$$L(z, p, q) = p^n (1-p)^{m-n} q^{N-n} (1-q)^{(M-m)-(N-n)} \quad (9.1)$$

其中, N 为研究区中的总病例数, n 为窗口中的病例数, M 为研究区中的非病例数, m 为窗口中的非病例数, $p = n/m$ 为病例在窗口中的概率, $q = (N-n)/(M-m)$ 为病例在窗口外的概率。

对每个窗口, 求似然函数的最大值, 最可能的集聚圈就是窗口内最不可能为随机分布的圆。这里, 最大似然的统计检验是基于蒙特卡洛法。这种方法找到了最可能的第一级集聚圈后, 还可能找到与之不重叠的次一级集聚圈。

9.2 案例 9A: 中国南部地区泰语地名的空间聚类分析

170

本案例是对第三章第 3.2 节和第 3.4 节所述中国南部地区泰语地名研究的扩展。第三章利用空间平滑和空间插值的方法, 对泰语地名的空间分布进行了可视化显示。图像显示还只是对现象的描述性研究, 不能区别泰语地名在区域上的分布到底是随机的还是存在集聚性。要回答这个问题, 还得依靠严格的空间统计分析。本案例采用基于点的空间聚类分析方法, 选用 SaTScan 软件(版本 5.1)来完成。

这个案例使用的数据与案例 3A 和 3B 相同, 主要是点图层 qztai。图层属性表中的变量 TAI 值为 1 代表泰语地名, 值为 0 代表非泰语地名。另外, shape 文件 qzcnty 用作绘制背景地图。

1. 用 ArcGIS 准备 SaTScan 软件的数据

在 SaTScan 软件平台下, 用伯努利模型执行基于点的空间聚类分析需要定义三个数据文件, 即事件文件(包含区位 ID 和每个区位的事件数)、非事件文件(包含区位 ID 和每个区位的非事件数)和坐标文件(包含区位 ID 和对应的笛卡尔坐标或经纬度坐标)。这一步就是在 ArcGIS 中定义好相关属性, 然后用 SaTScan 软件的 Import Wizard 读入, 产生上述三个文件。

图层 qztai 属性表中的变量 TAI, 已经定义了每个区位的事件, 其值为 1 正好表示该区位事件数为 1, 为 0 表示该区位事件数为 0。如何定义非事件呢? 对于这个案例来说很简单,

每个区位的非事件数正与事件数相反。在 ArcGIS 中，打开图层 qztai 属性表，增加一个新的变量 NONTAI，并计算 $NONTAI = 1 - TAI$ 。如何定义坐标呢？根据如下步骤：ArcToolbox > Coverage Tools > Data Management > Tables > Add XY Coordinates，这样 qztai 的属性表中就添加了两个新变量 X-COORD 和 Y-COORD，代表坐标。将属性表输出成 dBase 文件格式 qztai.dbf。

2. 用 SaTScan 软件执行空间聚类分析

运行 SaTScan 软件，选择 Create New Session，系统弹出一个新的对话框，如图 9.1 所示。

在第一个标签 Input 下，使用 Import Wizard 来定义事件文件（Case File）：点击对应 Case File 右边的按钮  > 选择 qztai.dbf 作为输入文件 > 在 Import Wizard 对话框中，在 Source File Variable 下选 qztai-id 作为 Location ID，选 TAI 作为 Number of Cases。利用类似的方法，定义非事件文件（Control File）和坐标文件（Coordinates File）。

在第二个标签 Analysis 下进行选择操作。在 Type of Analysis 中点击 Purely Spatial 选项，在 Probability Model 中点击 Bernoulli 选项，在 Scan for Areas with 中点击 High Rates 选项。

在第三个标签 Output 下，输入 Taicluster 作为结果输出文件，在 dBase 下点击所有四个选项按钮。

最后，在主菜单 Session 下选择 Execute Ctl+E 来执行空间聚类分析。分析结果保存在文件名含 Taicluster 的几个 dBase 文件中，文件中变量 CLUSTER 用于标志每个区位是否在集聚圈中(= 1 为一级集聚圈, = 2 为次级集聚圈, = <null> 不在任何集聚圈内)。

3. 分析结果的制图

171

在 ArcGIS 下，基于关联码（Taicluster.gis.dbf 中的变量 LOC_ID 对应 qztai 属性表中的变量 qztai-id）将 dBase 文件 Taicluster.gis.dbf 连到图层 qztai 上。图 9.2 中使用高亮符号来突出那些在一级和次级集聚圈内的地名，手工绘制的两个图大致显示了集聚圈的范围。

空间聚类分析确认了泰语地名主要集中分布在钦州的西边，在中部还有一小块为次一级的集中地。

9.3 基于面的空间聚类分析

172

在这一节中，首先讨论定义空间权重的不同方法，然后介绍两种已经可用 ArcGIS 9.0 计算的统计指数。与基于点的空间聚类分析相似，基于面的空间聚类分析方法也有全局检验和局部检验，前者的发现早于后者。基于面的方法还有罗杰申（Rogerson, 1999）的 R 统计值（参见 Wang, 2004）和其他方法，这里就不再讨论了。

9.3.1 空间权重定义方法

基于面的空间聚类分析利用空间权重来定义各个观察对象之间的空间关系。

基于距离来定义空间权重，方法有：

1. 以距离倒数为权重 ($1/d$)
2. 以距离平方的倒数为权重 ($1/d^2$)
3. 以距离阈值定义权重(如在阈值范围内定义为 1，在阈值范围外定义为 0)
4. 定义权重为距离的一个连续函数如

$$w_{ij} = \exp(-d_{ij}^2 / h^2)$$

其中， d_{ij} 是地区 i 和 地区 j 之间的距离， h 是距离阈值范围(Fotheringham et al., 2000, 第 111 页)，阈值范围的选取决定于距离影响程度，一个高的 h 值表示地区之间的相互影响距离远，影响范围大。除了上述基于距离定义的空间权重外，还可以根据多边形的邻接关系定义空间权重 (见第 1.4.2 节)，比如，如果地区 j 与地区 i 邻接，则定义 $w_{ij} = 1$ ，否则为 0。

上述基于距离的空间权重，在 ArcGIS 软件中都有相应的定义工具，具体是在“Conceptualization of Spatial Relationships”（定义空间关系）的时候选定的。ArcGIS 中可选的前三项为“Inverse Distance”、“Inverse Distance Squared”和“Fixed Distance Band”分别对应着上述的“距离倒数”、“距离平方倒数”和“距离阈值”这三种空间权重定义方法。第四项为“Zone of Indifference”，指给定一个距离阈值，在阈值范围内由距离倒数定义空间权重，在范围之外定

义为 0。上述几种方法都使用地区的几何中心来代表，所谓“以点代面”，距离既可以是欧氏距离，也可以是曼哈顿距离。最后面的“Get Spatial Weights From File”是指调用事先已经定义好的空间权重文件，空间权重文件应包含三项，目标区 ID、相关区 ID 和权重值(可以根据行进距离、时间、成本来定义，也可以根据多边形的邻接关系来定义)。

当前版本 ArcGIS 的 Spatial Statistics Tools（空间统计工具包）还只限于基于距离来定义空间权重，不能用多边形的邻接关系来定义空间权重（除非是调用事先已经定义好的权重文件）。GeoDa 软件可以直接用多边形的邻接（包括 R 邻接和 Q 邻接）来定义空间权重，并可计算相关的空间聚类指数。

9.3.2 基于面的全局聚类检验

莫兰 I (Moran's I) 指数 (Moran, 1950) 是最早应用于全局聚类检验的方法(Cliff and Ord, 1973)。它检验整个研究区中邻近地区间是相似、相异（空间正相关、负相关），还是相互独立的。莫兰 I 指数计算公式如下：

$$I = \frac{N \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \quad (9.2)$$

这里， N 是研究区内地区总数， w_{ij} 是空间权重， x_i 和 x_j 分别是区域 i 和 j 的属性， \bar{x} 是属性的平均值。

莫兰 I 指数可以看作是观察值与它的空间滞迟 (Spatial Lag) 之间的相关系数。变量 x 的空间滞迟是 x 在邻域 j 的平均值，定义为：

$$x_{i,-1} = \sum_j w_{ij} x_j / \sum_j w_{ij} \quad (9.3)$$

因此，莫兰 I 指数值处于 -1 到 1 之间，值接近 1 时表明具有相似的属性集聚在一起(即高值与高值相邻、低值与低值相邻)；值接近 -1 时表明具有相异的属性集聚在一起(即高值与低值相邻、低值与高值相邻)。如果莫兰 I 指数接近于 0，则表示属性是随机分布的，或者不存在空间自相关性。

与莫兰 I 指数相似, 吉瑞 C 指数(Geary's C) (Geary, 1954) 也是全局聚类检验的一个指数。计算莫兰 I 指数时, 用的是中值离差的叉乘, 但是, 吉瑞 C 强调的是观察值之间的离差, 其公式为:

$$C = \frac{(N-1) \sum_i \sum_j w_{ij} (x_i - x_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \quad (9.4)$$

吉瑞 C 指数值通常在 0 到 2 之间, 虽然 2 不是一个严格的上界。其值为 1 时, 表示属性的观察值在空间上是相互独立的, 值在 0 到 1 之间时表示空间正相关, 值在 1 到 2 之间为空间负相关。因此, 吉瑞 C 指数与莫兰 I 指数刚好相反。吉瑞 C 指数有时也称为 G 系数 (Getis-Ord general G), 例如在 ArcGIS 中就用这个名, 用于区分在局部聚类分析中使用的指数 G_i statistic。

可以通过随机化的方法来检验莫兰 I 指数和吉瑞 C 指数的统计显著性。

在 ArcGIS 9.0 软件新增的空间统计工具包中, 提供了莫兰 I 指数和吉瑞 C 指数的计算功能, 具体步骤是: ArcToolbox > Spatial Statistics Tools > Analyzing Patterns > 选 Spatial Autocorrelation (Moran's I) 计算莫兰 I, 选 High-Low Clustering (Getis-Ord General G) 计算吉瑞 C。GeoDa 和 CrimeStat 软件包中也有莫兰 I 指数和吉瑞 C 指数的计算工具。

9.3.3 基于面的局部聚类检验

安索林 (Anselin, 1995) 提出了一个局部莫兰指数 (Local Moran Index), 或称 LISA (Local Indicator of Spatial Association), 用来检验局部地区是否存在相似或相异的观察值聚集在一起。区域 i 的局部莫兰指数用来度量区域 i 和它邻域之间的关联程度, 定义为:

174

$$I_i = \frac{(x_i - \bar{x})}{s_x^2} \sum_j [w_{ij} (x_j - \bar{x})] \quad (9.5)$$

其中, $s_x^2 = \sum_j (x_j - \bar{x})^2 / n$ 是方差, 其它符号与式(9.2)中的相同。注意式中对 j 的累加不包括区域 i 本身, 即 $j \neq i$ 。正的 I_i 表示一个高值被高值所包围(高-高), 或者是一个低值被低

值所包围(低-低)。负的 I_i 表示一个低值被高值所包围(低-高)，或者是一个高值被低值所包围(高-低)。

类似地，格迪思和欧德 (Getis and Ord, 1992)开发了一个吉瑞 C 指数的局部聚类检验版本，称之为 G_i 指数(G_i statistic)，用来检验局部地区是否存在统计显著的高值或低值。 G_i 指数的定义如下：

$$G_i^* = \frac{\sum_j (w_{ij} x_j)}{\sum_j x_j} \quad (9.6)$$

公式中的符号与 (9.5) 式相同，同样地，式中对 j 的累加不包括区域 i 本身，即 $j \neq i$ 。这个指数用来检验局部地区是否有高值或低值在空间上趋于集聚。高的 G_i 值表示高值的样本集中在一起，而低的 G_i 值表示低值的样本集中在一起。 G_i 指数还可用于回归分析中的空间滤波处理，解决空间自相关问题(Getis and Griffith, 2002)，详见附录 9。

局部莫兰指数和 G_i 指数也可以通过随机化的方法来检验其统计显著性。

在 ArcGIS 中，计算局部莫兰指数和 G_i 指数的具体步骤是，ArcToolbox > Spatial Statistics Tools > Mapping Clusters > 选 Cluster and Outlier Analysis (Anselin Local Morans I) 计算局部莫兰指数，选 Hot Spot Analysis (Getis-Ord G_i^*) 计算 G_i 指数。计算结果可以分别通过“Cluster and Outlier Analysis with Rendering”和“Hot Spot Analysis with Rendering”的工具来绘图显示。GeoDa 和 CrimeStat 软件包也能计算局部莫兰指数，但不能计算 G_i 指数。

在应用中，空间聚类分析的各种指数值和相应的统计检验都具有重要意义。例如，Shen(1994, 177 页)在分析旧金山地区各社区控制发展政策的影响时，利用莫兰 I 指数检验了两种理论。第一种理论是，那些制订并实施控制发展政策（以防止人口大量迁入导致的交通堵塞、学校拥挤、环境恶化等问题）的社区往往是很吸引人的地方，很多人不能迁入这些社区，只好在其邻近的社区（条件也不错但没有控制发展）找地方住下来，这样一来，实施控制发展政策的人口低增长地区就邻近于无控制发展政策的、次优的高增长地区，在空间分布上表现为负自相关。第二种理论与所谓的 NIMBY (Not In My Backyard，即“不在我后院”)现象有关，控

制发展的社区也不让其相邻的社区发展太快，这样低增长社区会聚集在一起，而一些鼓励发展的社区也聚集在一起，在空间分布上表现为正自相关。究竟哪一种理论更有说服力，哪一种现象更明显，就得靠空间聚类分析和严谨的统计检验来判断。

9.4 案例研究 9B: 空间聚类分析在伊利诺伊州癌症分布研究中的应用

案例中的资料来自于伊利诺伊州公共卫生厅的癌症登记中心 (Illinois State Cancer Registry 或 ISCR)，网址为 <http://www.idph.state.il.us/about/epi/cancer.htm>，我们具体要用的只是其中以县为单元的数据。ISCR 每年公布一次全州的癌症病例数据，但为保护病人家庭隐私权，数据中病例的诊断时间只告诉其所在的 5 年段，如 1986-90, 1987-91 等，而不是具体的年度。本案例使用的是 1996-2000 年的资料。为了方便，我们只是简单地计算各县的癌症病例总数和发病率，并没有细分年龄、性别、种族和其他因子等具体情况。本研究将分析 4 种高发病率的癌症，即乳腺癌、肺癌、肠癌和前列腺癌。ISCR 提供癌症资料的同时，还发布了每县每年的人口资料，我们取 1996~2000 年 5 年间的平均人口值。

资料经整理后生成图层 `ilcnty`，包含有 6 个变量，其中县标识项不参加分析，其他 5 项为：`POPU9600` (1996-2000 人口平均值)、`COLONC` (5 年直肠癌病例总数)、`LUNGC` (5 年内肺癌病例总数)、`BREASTC` (5 年内乳腺癌病例总数)和 `PROSTC` (5 年内前列腺癌病例总数)。

1. 癌症发病率的计算与制图

在 ArcGIS 中打开图层 `ilcnty` 的属性表，增加变量 `COLONRAT`、`LUNGRAT`、`BREASTRAT` 和 `PROSTRAT`，分别代表肠癌、肺癌、乳腺癌和前列腺癌的发病率。癌症发病率一般以每 10 万人为单位计算，比如肠癌 5 年发病率的计算公式为 $COLONRAT = 100000 * COLONC / POPU9600$ 。表 9.1 总结了 1996-2000 年间各县癌症的基本统计数据。注意，全州发病率是由州内总病例数除以州人口总数得来的，不同于各县发病率的平均值。

下面仍以肠癌为例说明各个分析步骤。如图 9.3 显示了伊利诺伊州各县在 1996-2000 年间的肠癌发病率分布图。图中第一类图例显示的是发病率比州发病率 (288.3) 低的县，主要在东北角上的芝加哥大都市区。第二类图例显示的是发病率在州发病率 (288.3) 与各县平均值

(374.6) 之间的县。另两类图例显示的是更高的肠癌发病率，分布在东南角，西部也有一部分。

2. 计算全局聚类检验指数

在 ArcToolbox 工具包中选择 Spatial Statistics Tools > Analyzing Patterns > 选择 High-Low Clustering (Getis-Ord General G) 工具，弹出对话框如图 9.4 所示，选择面图层 ilcnty 作为输入要素 (Input Feature Class)，定义其中的 COLONRAT 为输入变量 (Input Field)，选中 “Display Output Graphically”，其他项使用缺省值即可 (如在 “Conceptualization of Spatial Relationships” 列表中使用 “Inverse Distance”)。分析结果在图形窗口中显示信息 “全局空间聚类为随机分布的可能性小于 5%”，也就是说，肠癌发病率在空间上存在集聚性，其统计显著性高于 5% (百分比越低越显著)。相关的统计指数见表 9.2。

选择 “Spatial Autocorrelation (Moran’s I)” 工具，重复上面的步骤，可以计算全部莫兰 I 指数，其结果同样表示肠癌发病率空间上存在集聚性，统计显著性更高 (在 1% 的水平上)。

吉瑞 C 指数 (Getis-Ord general G) 和莫兰 I 指数都采用正态分布的 z 统计检验，即 $z = (\text{统计指数} - \text{期望值}) / \sqrt{\text{方差}}$ 。如果 z 值大于关键值 1.960 时，统计显著性在 5% 水平；如果 z 值大于关键值 2.576 时，统计显著性在 1% 水平。例如，针对肠癌发病率计算的莫兰 I 指数为 0.09317，期望值为 -0.0099，方差为 0.0001327，可以计算出 $z = (0.09317 - (-0.0099)) / \sqrt{0.0001327} = 8.9489$ ，这时 z 值大于 2.576，表明统计显著性水平在 1% 之上。

根据上面的步骤计算其他癌症的发病率和相关的空间统计指数，表 9.2 为分析结果。从表 178 中可以看出，两种全局聚类指数都表明肺癌的 z 值最高，具有最强的空间聚类特征，其次为肠癌、前列腺癌和乳腺癌。同时还可以看出，基于吉瑞 C 指数 (general G) 的统计显著性比基于莫兰 I 指数的弱一些。

3. 计算局部莫兰指数和局部 Gi 指数

在 ArcToolbox 中，选 Spatial Statistics Tools > Mapping Clusters > Cluster and Outlier Analysis (Anselin Local Morans I)，弹出对话框。输入类似于第 2 步的图层和属性变量，定义输出图层为 Colon_Lisa。计算结果在输出图层中增加了几项参数，其中 LmiInvDst 为局部莫兰指数（用距离倒数定义的空间权重），LmzInvDst 为对应的 z 值。可以直接利用局部莫兰指数（LmiInvDst）来绘图，也可以利用另一个工具“Cluster and Outlier Analysis with Rendering”来制图。局部莫兰指数只是表示属性相似（莫兰指数值为正）或相异（莫兰指数值为负）的观察值集聚在一起，并不表示属性值（癌症发病率）究竟是高还是低。由于我们一般感兴趣的是那些发病率高的地区，因此，可以首先将发病率低于全州发病率(288.3)的县排除在外，然后对那些高发病率的县分类显示。图 9.5 显示主要的高发病率集聚区分布在东南角上。

利用 “Hot Spot Analysis (Getis-Ord G_i^*)”工具重复上面的步骤，可以计算 G_i^* 指数。在输出结果中，增加了一个新变量 GiInvDst，存放 G_i^* 指数。高的 G_i^* 指数表示高发病率的集聚区（热点），而低的 G_i^* 指数表示低发病率的集聚区（冷点）。图 9.6 显示了用 G_i^* 指数分析肠癌空间分布的结果。

9.5 空间回归分析方法

181

空间聚类分析用来探测空间自相关性，而空间自相关就是变量的观察值与地理位置相关（比如高值与高值在一起，低值与低值在一起）。在变量没有空间自相关的情况下，即空间上相互独立时，可以使用通常的 OLS 回归模型（参见第六章附录 6B）来分析问题。如果变量的观察值存在空间自相关时，则需要使用空间回归模型。

一般回归模型用矩阵形式表示为：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (9.7)$$

式中 \mathbf{y} 是因变量，写成 n 个观察值的向量形式； \mathbf{X} 代表 m 个自变量，每个自变量有 n 个观察值，所以是一个 $n \times m$ 的矩阵； $\boldsymbol{\beta}$ 为对应于 m 个自变量的回归系数，也写作向量形式； $\boldsymbol{\varepsilon}$ 是随机误差向量，或称残差向量，残差向量的分布要求是相互独立的且中值为 0。

当空间自相关存在时，残差就不再相互独立，所以 OLS 回归模型不再适用。在这一节主要讨论两个常用的最大似然估计法，来解决这种空间自相关的情况。第一个为空间滞迟模型 (*Spatial Lag Model*, 参见 Baller et al., 2001), 或称空间自回归模型 (*spatially autoregressive model*, 参见 Fotheringham et al., 2000, 167 页)。模型的右边加了一个“自变量”，是因变量 y 邻域的平均值，即 y 的空间滞迟，许多文献中写成 y_{-1} 。用矩阵 \mathbf{W} 表示空间权重，空间滞迟可以表示为 $\mathbf{W}\mathbf{y}$ ，其定义与前面的 9.3 式相同，矩阵 \mathbf{W} 的第 i 行和第 j 列的值为 $w_{ij} / \sum_j w_{ij}$ 。空间回归的滞迟模型可以写成以下公式。

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (9.8)$$

其中， ρ 为空间滞迟变量的回归系数，其他变量和参数与 9.7 式定义相同。

重新组织 9.8 式可以写成

$$(\mathbf{I} - \rho \mathbf{W})\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

假如矩阵 $(\mathbf{I} - \rho \mathbf{W})$ 是可逆的，则上式可变为

$$\mathbf{y} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon} \quad (9.9)$$

上式是空间滞迟模型的终结式 (reduced form)。可以看出，每个在 i 地区的值 y_i 不仅仅与这个地区的 x_i 有关(就象一般的回归分析那样)，同时，通过乘上一个空间因子 $(\mathbf{I} - \rho \mathbf{W})^{-1}$ ，也受其他地区的 x_j 值影响。这个模型与时间序列分析中的自回归模型 (autoregressive model) 不同，它不能用 SAS 中的时间序列分析程序如 AR 或 AMAR 来计算。

第二个考虑到空间自相关的回归模型为空间残差模型 (spatial error model, 参见 Baller et al., 2001), 或称为空间移动平均模型 (spatial moving average model, 参见 Fotheringham et al., 2000, 169 页), 或 SAR 模型 (simultaneous autoregressive model, 参见 Griffith and Amrhein, 1997, 276 页)。前面述及的空间滞迟模型强调因变量在空间上是自相关的，而空间残差模型把残差看作 182 是空间上自相关的。模型表述为：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (9.10)$$

其中，残差 \mathbf{u} 又可用它的空间滞迟来表示，也就是：

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \quad (9.11)$$

这里， λ 是空间残差自回归系数，剩余的第二个残差项 $\boldsymbol{\varepsilon}$ 是相互独立的随机误差。

解方程 (9.11) 得到 \mathbf{u} ，代入 (9.10)，得到终极模型：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \lambda \mathbf{W})^{-1} \boldsymbol{\varepsilon} \quad (9.12)$$

上式说明每个地区 i 的值 y_i 是受所有其他地区 j 上的随机误差 ε_j 影响，影响系数为 $(\mathbf{I} - \lambda \mathbf{W})^{-1}$ 。

无论是 (9.9) 式的空间滞迟模型，还是 (9.2) 式的空间残差模型，都可用最大似然法来测算 (Anselin and Bera, 1998)，下一节的案例研究将演示这两种模型在 GeoDa 软件上的实现 (参见 Anselin, 2001)。安索林 (Anselin, 1988) 讨论了用几种统计指数来衡量到底什么情况下更适合用哪种模型，可是实践中很难分出来 (Griffith and Amrhein, 1997, 277 页)，其实二者差别不大。

9.6 案例研究 9C：芝加哥谋杀犯罪研究中的空间回归分析

本节继续分析第八章介绍过的芝加哥市谋杀犯罪案例。第八章中的案例 8 使用了常用的 OLS 回归方法，而这一章将使用空间回归方法来控制空间自相关问题。

除了在案例 8 研究中使用的多边形图层 citytrt 以外，随书光盘还提供了本案例要用到的一个多边形图层 CityCom，它包含芝加哥市的 77 个社区 (community areas)，不包括 O'Hare 机场。

关于图层 citytrt 的属性信息，请参考第八章第 8.4 节中的详细描述，另外说明一下，属性表中的变量 comm 是用来标识每个人口普查区到底是属于哪个社区的，就是图层 citycom 中的“社区”。我们将用这两个地理单元 (也就是人口普查区和社区) 来继续分析芝加哥市的就业便捷度和谋杀犯罪率之间的关系。有兴趣的读者，还可以利用第八章尺度空间聚

类分析生成的新地理单元，练习更多的空间回归分析。正如 Wang（2005）所做的，分析所用的空间单元不断扩大，从人口普查区增大到第一级聚类区，再增大到第二级聚类区，最后到社区，形成一个研究芝加哥市谋杀犯罪的多级空间单元系列。

9.6.1 第一部分：用 GeoDa 软件进行基于人口普查区的空间回归分析

1. 空间回归分析之前的准备工作

如果没有做完第八章中的案例分析，那么按第 8.4 节中的第一、第二步，生成一个 shape 格式的图层文件 citytract，其属性表有 845 个人口普查区，含有变量 Lhomirat，是谋杀犯罪率的对数值。

2. 用 GeoDa 定义空间权重

启动 GeoDa 软件，根据以下操作路线， Tools > Weights > Create ，激活定义空间权重的对话框，如图 9.7 所示。在对话框中选择文件 citytract.shp 作为输入文件，键入文件名 tract 作为输出文件，选择 Queen Contiguity（多边形邻域关系）来定义空间权重，最后按 Create 执行。生成的空间权重文件为 tract.GAL。

3. 在 GeoDa 中执行 OLS 回归分析

在 GeoDa 软件中，选择 Methods > Regress。在 Get DBF File 对话框中选择文件 citytract.dbf 作为输入文件名。在下一个对话框 Regression Title & Output 中输入 OLS Regression for Census Tracts 作为报告名称，然后输入 Trt_OLS 作为输出文件名，点击 OK 激活模型构建的对话框，如图 9.8 所示。(1) 使用按钮“>”、“>>”、“<”、“<<”将变量 Lhomirate 从左边的 Select Variable 框中移到右边的 Dependent Variable 框中，从列表框中将变量 factor1、factor2、factor3 和 JA 移到 Independent Variables 框中；(2) 在模型 Models 选择栏中选择 Classic；(3) 点 Run 执行分析。结果存在文件 Trt_OLS.OLS 中，见表 9.3，这个分析结果与利用 SAS 软件生成的结果(在表 8.3 中)是相同的。

184

4. 在 GeoDa 中执行空间时滞回归分析

操作步骤与上面的基本相同，注意定义一个不同的输出结果文件名。另外不同的还有两点：
：（1）在 **Weight Files**（权重文件）这一项下，选 `tract.GAL` 定义空间权重文件；（2）在 **Models**（模型）下，选 **Spatial Lag**，分析结果见表 9.3。

5. 在 GeoDa 中执行空间残差回归分析

操作步骤与上面的基本相同，注意在 **Models**（模型）下，选 **Spatial Error**，分析结果见表 9.3。

由于上述两个空间回归模型都是基于最大似然法来计算的，所以结果中没有 OLS 回归中的 R^2 ，相应的是拟合系数（Sq. Corr.）。

9.6.2 第二部分：用 GeoDa 软件进行基于社区的空间回归分析

185

1. 准备社区的图层文件

在 ArcMap 中打开图层 `citycom`，选择 `popu > 0` 的社区，共有 77 个社区，将结果输出为 `shape` 文件 `citycomm`。

2. 将属性数据从人口普查区合并到社区

相关变量(`homirate`、`factor1`、`factor2`、`factor3` 和 `JA`)，可以从人口普查区图层合并到与之对应的社区图层，具体方法参见第八章第 8.4 节的第 7 步。另外，在社区图层 `citycomm` 的属性表中增加变量 `Lhomirat`，并计算 `Lhomirat = log(homirate+1)`。

3. 定义空间权重并执行回归分析

按照第一部分的第 2 到 5 步，基于社区图层可以定义一个新的空间权重文件 `comm.GAL`，然后运行 OLS 回归分析、空间迟滞回归和空间残差回归三种分析模型，结果见表 9.4。

9.6.3 讨论

可以从表 9.3 和表 9.4 的回归分析结果中，总结几点重要的发现。

1. 在基于人口普查区的分析中，对应于空间迟滞参数 ρ 和空间残差参数 λ 的 t -统计值都很显著，表示用空间回归取代 OLS 回归是十分必要的。在基于社区的分析中，空间迟滞参数

186

具有 0.05 的统计显著性水平，而空间残差参数则不显著，这表明社区的空间自相关性没有人口普查区的强。也就是说，如果我们用一般的 OLS 方法分析社区的数据，未尝不可。

2. 在基于人口普查区和基于社区的空间回归结果中，从自变量回归系数的正负和显著性水平来看，与 OLS 回归基本是一致的。

3. 在基于人口普查区的回归模型中，就业便捷度差的地区与高谋杀犯罪率是相关的，而且统计上有显著性。作者和合作者（Wang and Minor, 2002）在研究克里夫兰市的犯罪现象时，采用了简单的二元回归法，结果表明就业便捷度和各种（包括谋杀）犯罪率有很强的负相关。本案例的结果更是如此。由于这里考虑了空间自相关的问题，结果更加可信。

4. 在基于社区的回归分析中，虽然自变量就业便捷度的回归系数为负，但统计上并不显著。可能是由于芝加哥市社区基本上是根据地理特征(河流、道路、高速路等)来划分的，各社区不一定是由社会经济特征相似的人口普查区组成。这样一来，将属性数据从人口普查区合并到社区时，相当于对数据进行了平滑处理，大部分差异性被平滑掉了一些，也就是说，一些信息在从人口普查区到社区的合并过程中丢失了。由于研究中地理单元的变化，分析的结论也变了，这就是所谓的 MAUP 问题。 187

5. 在另外三个自变量中，因子一和因子二无论在基于人口普查区还是在基于社区的分析中，都是正相关的，并具有较好的显著性水平，说明社会经济条件差的地区的确遭遇较高的谋杀犯罪率；因子三在基于人口普查区的分析中不显著，在基于社区的分析中具有显著性，表明它的影响不稳定。

9.7 总结

空间聚类分析可以探测空间分布的非随机性或空间自相关。在实践中，基于点和基于面的空间聚类分析方法是明显不同的。基于点的方法分析在某个距离半径内点的分布是否比随机模式下更集聚；而基于面的方法则用来检验邻近目标之间属性是否相似或相异。在犯罪分析和健康分析的应用研究中会经常用到空间聚类分析。本章中案例 9A 中国南部泰语地名的分析演示了基于点的空间聚类分析方法。选择这个案例的目的是想说明，GIS 和空间分析方法应用领域

可以拓展。象传统的文科领域比如历史、语言、文化等学科，表面上看起来，能用到的定量分析方法不多，实际上运用 GIS 和空间分析方法的潜力很大。案例 9B 讲解了基于面的空间聚类分析方法在癌症分布方面的应用。

样本中空间自相关性的存在要求我们在回归分析中使用空间回归分析方法。空间回归分析中常用的是空间迟滞模型和空间残差模型，二者都是用最大似然法来计算的。案例 9C 利用空间回归分析模型检验了芝加哥市就业便捷度和谋杀犯罪率之间的关系。

当前版本的 ArcGIS 软件新加了一些基于面的空间统计分析方法，但是还没有基于点的空间聚类法和空间回归分析。读者可以选用一些免费软件来完成这些任务，如用 SaTScan 来分析点的空间集聚，用 GeoDa 做空间回归分析。这些软件包简单易学，使用方便。

附录 9 回归分析中的空间滤值法

格迪思(1995)和桂菲斯(2000)提出的空间滤值法，采用的是另一种思路来解决回归分析中的空间自相关问题。该方法把每个变量分解成空间影响和非空间影响两部分，滤去变量的空间影响部分就可以用传统的回归（如 OLS）方法来分析(Getis and Griffith, 2002)。该方法与基于最大似然估计的空间回归法相比，最大的优点就是把变量的空间与非空间影响分开，区别各自的贡献，结果比较容易解释。桂菲斯(2000)的特征函数分解法（eigenfunction decomposition method）比较复杂，计算量大，步骤也多，这里就不讨论了。这个附录简要介绍格迪思的方法。

188

格迪思方法的核心思想是将原先存在空间自相关的变量划分为二，一部分是过滤后的非空间变量，另一部分是剩余的空间变量。过滤后的非空间变量就可以作为一般的变量，进行 OLS 回归分析。基于 9.6 式对 G_i 的定义，将原变量 x_i 过滤后的非空间变量 x_i^* 可以写成：

$$x_i^* = \frac{W_i / (n-1)}{G_i} x_i$$

其中, $W_i = \sum_j w_{ij}$ 是平均空间权重 ($i \neq j$), n 是观察值的数目, G_i 是局部 G_i 指数。离差 $L_{xi} = x_i - x_i^*$ 代表变量 i 的空间部分。注意, 分子 $W_i/(n-1)$ 就是 G_i 的期望值。当原变量不存在空间自相关时, $x_i^* = x_i$, 离差 $L_{xi} = x_i - x_i^*$ 为 0, 不存在剩余的空间变量。

将过滤后的变量(包括自变量和因变量)代入传统的 OLS 回归分析中, 就是空间过滤回归模型:

$$y^* = f(x_1^*, x_2^*, \dots)$$

其中, y^* 是过滤后的因变量, x_1^*, x_2^* 等是过滤后的自变量。

在最终的回归模型中, 因变量和自变量都包括过滤后的非空间和剩余的空间变量两部分:

$$y = f(x_1^*, L_{x_1}, x_2^*, L_{x_2}, \dots)$$

其中, y 是原始的因变量, $L_{x_1}, L_{x_2} \dots$ 是对应自变量 x_1, x_2, \dots 的空间部分。

与 G_i 指数相似, 格迪思的空间滤值模型仅适用于那些零起始的正值变量, 不适用于百分比、比率、有负值之类的变量 (Getis and Griffith, 2002, 132 页)。

表 9.1 1986-2000 年伊利诺伊州各县癌症发病率（每十万人）

癌症类型	全州发病率	各县平均	最小值	最大值	标准差
乳腺癌	351.23	384.43	225.59	596.59	66.28
肺癌	349.09	446.77	228.73	758.82	119.38
肠癌	288.30	374.60	205.93	584.13	80.66
前列腺癌	316.82	369.09	198.74	533.26	83.33

表 9.2 癌症发病率的全局聚类分析结果（n = 102）

空间聚类指标		乳腺癌	肺癌	肠癌	前列腺癌
Moran's <i>I</i>	指标值	0.0426	0.1211	0.0932	0.0696
	期望值	-0.0099	-0.0099	-0.0099	-0.0099
	方差	1.3234E-4	1.330E-4	1.3270E-4	1.3384E-4
	Z 值	4.5619***	11.3630***	8.9489***	6.8706***
General <i>G</i>	指标值	2.0320E-6	2.0508E-6	2.0411E-6	2.0402E-6
	期望值	2.0186E-6	2.0186E-6	2.0186E-6	2.0186E-6
	方差	7.3044E-17	1.7702E-16	1.1436E-16	1.2590E-17
	Z 值	1.5662	2.4209*	2.0993*	1.9257

注：*** 表示 0.001 的显著度，**表示 0.01 的显著度，*表示 0.05 的显著度

表 9.3 芝加哥谋杀犯罪率的 OLS 回归和空间回归分析结果

(n = 845 个人口普查区)

自变量	OLS 回归	空间滞迟模型	空间残差模型
截距	6.1324 (10.87) ***	4.5338 (7.52) ***	5.8304 (8.97) ***
因子一	1.2200 (15.43) ***	0.9654 (10.91) ***	1.1777 (12.89) ***
因子二	0.4989 (7.41) ***	0.4048 (6.01) ***	0.4777 (6.01) ***
因子三	-0.1230 (-1.84)	-0.0993 (-1.53)	-0.0858 (-1.09)
就业便捷度	-2.9143 (-5.41) ***	-2.2056 (-4.13) ***	-2.6321 (-4.26) ***
空间滞迟(ρ)		0.2750 (5.90) ***	
空间残差(λ)			0.2627 (4.82) ***
拟合系数 (Sq. Corr.)	0.395	0.424	0.415

注：括号内为 t 值；*** 表示 0.001 的显著度，**表示 0.01 的显著度，*表示 0.05 的显著度

表 9.4 芝加哥谋杀犯罪率的 OLS 回归和空间回归分析结果

(n= 77 个社区)

自变量	OLS 回归	空间滞迟模型	空间残差模型
截距	5.5679 (5.63) ***	4.2516 (4.07) ***	5.3882 (5.11) ***
因子一	1.2415 (8.92) ***	1.0671 (7.22) ***	1.2185 (8.44) ***
因子二	0.4287 (3.45) ***	0.4095 (3.54) ***	0.4244 (3.37) ***
因子三	-0.3641 (-3.40) **	-0.3055 (-2.95) **	-0.3657 (-3.20) **
就业便捷度	-1.4246 (-1.48)	-1.0768 (-1.20)	-1.2599 (-1.23)
空间滞迟(ρ)		0.2369 (2.45) *	
空间残差(λ)			0.1647 (1.01)
拟合系数 (Sq. Corr.)	0.750	0.769	0.755

注：括号内为 t 值；*** 表示 0.001 的显著度，**表示 0.01 的显著度，*表示 0.05 的显著度