经济与社会科学空间分析

王劲峰　Manfred M. Fischer　刘铁军　著

科学出版社

经济与社会科学空间分析

王劲峰　Manfred M. Fischer　刘铁军　著

科学出版社
北京

内 容 简 介

空间位置不仅仅是地理学和区域科学的核心，在社会科学及生态学等领域，人们也越来越关注研究对象的空间维度，将空间分析方法和技术（如地理信息系统、全球定位系统、遥感等）运用到实际工作中。本书分别介绍了社会和环境科学中主要的三类数据分析方法：点数据分析、面数据分析和流数据分析。

本书可供空间数据分析领域，特别是社会与环境科学领域相关人员参考使用。

图书在版编目(CIP)数据

经济与社会科学空间分析/王劲峰等著. —北京：科学出版社，2012

ISBN 978-7-03-035329-0

Ⅰ.①经… Ⅱ.①王… Ⅲ.①经济学-地域研究②社会科学-地域研究 Ⅳ.①F0②C0

中国版本图书馆 CIP 数据核字(2012)第 192163 号

责任编辑：杨 红 / 责任校对：张怡君
责任印制：闫 磊 / 封面设计：迷底书装

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

骏 走 印 刷 厂 印刷
科学出版社发行 各地新华书店经销

*

2012 年 9 月第 一 版 开本：720×1000 1/16
2012 年 9 月第一次印刷 印张：9 1/4
字数：180 000

定价：30.00 元

（如有印装质量问题，我社负责调换）

前　言

人们通常理所当然地认为空间位置仅仅是地理学和区域科学的核心，而现在无论是在社会科学，还是生态学等自然科学领域，人们也越来越关注研究对象的空间维度。越来越多的社会科学研究者将新的方法和技术（如地理信息系统、全球定位系统、遥感等）运用到实际工作中。此外，人们在理论框架中也开始更加关注不同空间位置之间事件的互动。

从广义上讲，空间分析可以定义为对地理空间中的事物的定量化分析（Bailey and Gatrell，1995）。将如此广阔的领域囊括到一本书中是不现实的[可参考 Fischer and Getis（2010）对该领域多样性的一个统计]，因此，我们决定将本书的内容限制到空间分析领域一个重要的部分——空间数据分析。这样做的目的是更多地关注地理空间过程中的数据，并据此描述或者解释这个过程的模型、方法和技术。通过这种方式定义空间数据分析，我们将本书定位于统计学的描述和空间数据建模领域，并将其内容限定于特定的一组方法上。这样我们剔除了一些重要的计量方法，如多种模式的网络分析和空间配置分析等本应被包含在更广泛的空间分析主题中的方法。

无论空间数据分析是否可以视为一个单独的学术领域，在过去的 20 年里，它显然已经成为理解空间现象的一个重要的副产品。空间数据是指那些与空间位置有关的数据，空间位置可以是绝对的，如一个邮件地址或者网格参考，也可以是相对的，如遥感影像中的一个像素。

过去的几十年间，研究人员在该领域已经发表、出版了很多出色的成果（Cliff and Ord，1981；Upton and Fingleton，1985；Anselin，1988b；Griffith，1988；Ripley，1988；Cressie，1993；Haining，1990，2003；Bailey and Gatrell，1995；LeSage and Pace，2009）。其中大部分是给研究人员使用的，本书则以“数据驱动”为出发点，而不是“以理论为主导”的方式，将空间数据分析介绍给研究生。本着这个目标，我们并不打算详尽地讨论整个空间数据分析领域，而是着重介绍三种主要的空间数据类型的分析：①点数据——其定义为一组点位，如华北平原城镇分布，或者一组点位及其属性值，如空间随机抽样的社会经济调查样本；②区域数据——也称多边形数据，如中国各省（自治区、直辖市）组成的多边形及其社会经济属性值；③空间互动数据——定义为记录地理空间中的各点或者区域，或者两两之间的测量数据，如资金、人员或者信息等事物的流动量。

本书将介绍空间数据分析领域中的一些我们认为相对容易并且更具实用价值

的模型、方法和技术。其主题既包括非正式的数据探索技术和方法，又包括正式的统计建模、参数估计和假设检验。本书分为三个部分，每个部分尽可能地独立。第一部分讲述点数据分析，包括仅考虑点位的空间格局识别和同时考虑点位及其属性值的场数据分析。第二部分着重讲述区域数据分析，这里讲的区域可以是规则网格，如遥感影像，也可以是一组不规则的面(polygon)对象。第三部分讲述空间互动数据，这种数据也被称为是“源-汇流”(origin-destination flow)，其内容与交通规划、移民、旅游、购物行为、货运物流甚至信息和知识的转移等方面研究相关。

本书中不涉及时空数据，而是假设数据是纯粹的空间数据。同时，虽然空间数据的计量、存储和检索都很重要，但也没有被包含入本书。地理信息系统提供了软件工具，通过建立地理参考(georeference)将空间与非空间、定量和非定量数据以非常方便的方式整合进入数据库。为了使本文的篇幅不至于过长，我们假定本书的读者具备中等水平的统计学和数学基础和初级的地理信息系统能力。

本书第一部分是基于已有成果编写的；第二部分和第三部分中的理论和方法部分是基于 M. M. Fischer 和 J. F. Wang 的*Spatial Data Analysis: Models, Methods and Techniques*(2011)翻译而来的，个别语句有所调整，特别增加了中国的真实案例。

王劲峰，北京
M. M. Fischer，维也纳
刘铁军，北京
2012-6-24

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第二部分 区域数据分析

第三部分　空间互动数据分析

第1章 引 言

本章我们主要介绍空间数据分析，并将其与其他的数据分析相区别。空间数据的含义是包含空间位置的数据，同时还可能包括属性信息。我们主要关注三类空间数据：点状数据、区域数据和源-汇流数据。点状数据是指分布于空间上的点状对象，包括空间非连续分布的变量，如疾病爆发点和产业聚集点；空间离散分布的监测点数据，并且通过对其属性进行插值可以生成空间连续分布图，如气象台站、哨点医院、土壤污染采样点、生态样方等。区域数据是指在空间上非连续变化的变量，由一组覆盖研究区的子区域组成。这些子区可以是规则网格，如遥感图像的一组像元，也可以是不规则的区域单元，如一组县界、一组流域、一组格网等。源-汇流（也称空间互动）数据则是与地理空间中的点对或者区域对有关。这种数据代表人口、商品、资金、信息或者知识从一组源地到一组汇地之间的流动，已经被广泛地运用于交通规划、人口迁移、旅游、消费行为、物流以及信息和知识在空间上的传输等方面的研究。

我们认为数据的空间自相关性使得传统的统计分析不再安全，需要空间分析工具作为支持。空间自相关性意味着观测值在空间上不再是独立的。最后，我们对一些空间分析所面临的实际问题进行简要的讨论。表1.1讨论地图、GIS和空间分析三种工具的适用问题。

表1.1 空间数据的研究目标和分析工具

目标	工具		
	地图	GIS	空间分析
显示	* * *	* * *	
查询	*	* * *	
格局识别	*	* *	* * *
因子分析			* * *
模拟、预报			* * *
影响分析、调控			* * *

* * * 很适合；* * 适合；* 一般

1.1　空间数据及其分析

数据包括数字或符号，在一定意义上是中性的，与上下文无关。信息则与之相反。原始的地理现象，如在特定时间和地点的温度，就是数据的一个例子。按照Longley等的定义(2001)，我们可以把空间数据视为由与地理世界相关的基本元素或者事实所构成。在其最原始的方面，空间数据的一个基本单元(严格地说，是一个数据)通常是与一个地理位置以及某些属性相关联的。例如，对于这样一个描述，“2011年9月20日下午3时，在北纬39度54分，东经116度24分26秒的温度为21摄氏度”，它将位置和时间关联至大气温度属性上。因此，我们可以说，空间数据连接了位置、时间和属性(在这里是温度)。

属性有多种形式。一些是自然或环境的，而另一些则是社会或经济方面的。一些属性是简单的标记空间位置，如邮政地址或土地所有权的记录等。另一些属性是计量在一个位置的事物(如大气温度和收入)，还有一些是分类信息，如农业、住宅和工业在土地利用上的分区。

对于空间数据分析而言，时间只是一个可选项，而地理位置才是其最根本的，也是其区别于其他数据分析的标志。如果我们仅处理属性，而忽略样本位置之间的空间关系，那么我们就不是在做空间数据分析。即使属性可能是非常重要的数据，但是只要与它们的空间信息分离，它们就失去了价值和意义(Bailey and Gatrell,1995)。为了实施空间数据分析，我们要求必须同时具备空间信息和属性信息，而不管属性信息是如何衡量的。属性值可能随空间位置而变化，如温度和收入；也可能不随空间位置而变化，如仅表示事件的发生地点。

空间数据分析需要一个基本的空间框架，在其上放置研究的空间现象。Longley等(2001)区分了地理表达的两个基本方式：空间现象的离散观和连续观。区别在于，是将地理空间看做是由离散对象所填充的，还是将其看做是由本质上连续的表面所覆盖。前者被称做空间的对象观(object view)，后者被称做场观(field view)。

对象观认为，被分析的空间现象的种类是依据其维度来识别的。占据区域的对象被称为二维，通常也被称为地区。那些更像是一维的线对象，包括河流、铁路或者道路等，被表示为一维对象，通常也称为线。那些零维的对象，如植物个体、人、建筑物、地震的震中等被称为点(Longley et al.,2001；Haining,2003)。表面或者体积对象具有长、宽、高，因此是三维的，但这类物体不在本书的讨论之列。当然，如何才能恰当地识别地物取决于我们所要研究的空间尺度(我们用以代表现实的细节程度)。如果我们是在探寻全国尺度上城市住区的分布状况，那么最好是将住区按照点的分布来处理。道路可以当做线对象来处理，但是它也取决于比例尺。在大比例尺的城市地图上道路就有宽度了，这对于汽车导航尤其重要。线也被用

于标记区域的边界，区域一般是指如国家、省、县、普查区等行政或法律上定义的实体，还包括诸如在地图上的植被或者土壤等“自然区”。

场观的重点在于空间现象的连续性，地理世界可以用一组有限的变量来描述，每个变量都是可以在地球表面任意点上测量到的，并且测量值沿着地表变化(Haining，2003)。如果我们考虑自然环境中的现象，如温度、土壤类型等，那么这些变量能够在地球表面的任何地方观测到(Longley et al.，2001)。当然，在实践中，这些变量是离散的。例如，温度是由一组站点监测到的，并且用一组线(即“等温线”)来表达。土壤类型也可以离散的采样并以连续的变动区域来表达。在所有这些案例中，都需要从离散样本估计出空间事物潜在的连续性(Bailey and Gatrell，1995)。

1.2　空间数据类型

在描述空间数据性质的时候，很重要的一点在于，要判断变量是在离散空间还是在连续空间上被监测到的，以及其变量值本身是离散的还是连续的。如果空间是连续的，如温度，那么变量值一定是连续的值，因为离散的变量不可能保持“场”的连续性。如果空间是离散的或者一个连续的空间被离散化，则变量值可以是连续变量也可以是离散变量(Haining，2003)。

对空间数据的空间和度量水平的概念型分类，是选择适当的统计技术来解决问题的第一步。但是这个分类并不是很充分，因为同样的空间对象可能代表截然不同的地理空间。例如，点(或者叫“重心”)也可以用于代表区域。表 1.2 提供了一个将空间数据分为四个类型的分类方法。

表 1.2　空间数据的类型：概念方案和示例

空间数据类型	概念方案		示例	
	变量	空间索引	变量	空间
点位数据	随机变量(离散或者连续)	空间位置固定	树木：健康与否 城堡：按照类型分类	二维离散 二维离散
场(地统计)数据	变量是位置的一个连续值函数	在二维空间中到处都定义了变量	温度 大气污染	二维连续 二维连续
区域数据	变量(离散或连续)是随机的	变量所依附的区域对象是固定的	区域生产总值 犯罪率	二维离散 二维离散
空间互动(流)数据	变量代表互动频率的均值，是随机的	与流变量所依附的点位置对	国际贸易 人口移民	二维离散 二维离散

(1) 点位数据，即在某个研究区内一系列的点位所组成的一个数据集，研究者所关注的事件(如病例或者犯罪等)在这些点位上发生。

此为试读，需要完整PDF请访问：www.ertongbook.com

(2) 场数据(也被称为地统计数据或空间连续数据),与在概念上连续的变量相关,并且其观测值来自一组预先确定的固定点位。

(3) 区域数据,是与固定数目的分布区单元(区域对象)相关的观测数据值,它可以是规则网格,也可以是一组不规则的区域或者分区。

(4) 空间互动数据(也被称做"源-汇流"或者"连接数据"),由两两空间点位或者两两区域之间联系观测值构成。

本书将着重介绍点数据分析(见第一部分)、空间区域数据分析(见第二部分)和空间互动数据分析(见第三部分)。在人类行为的研究方面,空间互动数据分析拥有悠久而杰出的历史,如交通流、移民、信息和知识传播。区域数据是空间数据分析的重要方面,特别是在社会科学领域。点数据则更多地存在于环境科学领域。

1.3　空间数据矩阵

本书中所有的分析技术都是用一个矩阵来记录空间数据的。空间数据按空间对象(点对象、区域对象、流对象)类型进行分类,变量及其观测水平依附于空间对象。

令 Z_1, Z_2, …, Z_k 代表 K 个随机变量,S 代表点或者区域的位置,那么空间数据矩阵(Haining,2003)一般可以表示为

$$\begin{array}{cccccc} & K\text{个变量的数据} & & & \text{位置} & \\ Z_1 & Z_2 & & Z_K & S & \\ \begin{bmatrix} z_1(1) & z_2(1) & \cdots & z_K(1) & s(1) \\ z_1(2) & z_2(2) & \cdots & z_K(2) & s(2) \\ \vdots & \vdots & & \vdots & \vdots \\ z_1(n) & z_2(n) & \cdots & z_K(n) & s(n) \end{bmatrix} & & & & & \begin{matrix} \text{案例} & 1 \\ \text{案例} & 2 \\ \vdots & \\ \text{案例} & n \end{matrix} \end{array}$$

上式可简化为

$$\{z_1(i),\ z_2(i),\ \cdots,\ z_k(i) \mid s(i)\}_{i=1,\cdots,n} \tag{1.1}$$

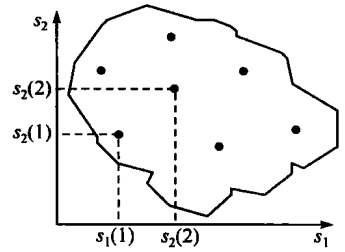
式中,z_k代表变量 $Z_k(k = 1, \cdots, K)$ 的一个实现,括号中的 i 是案例编号。附加在每个案例之后的 $s(i)$ 是空间对象的坐标位置。我们仅关心二维空间的数据,因此,$s(i)$包含两个地理坐标 s_1 和 s_2。这样,$s(i) = [s_1(i),\ s_2(i)]'$,其中 $[s_1(i),\ s_2(i)]'$是 $[s_1(i), s_2(i)]$ 的转置矢量。需要注意的是,本书主要讲述将空间位置视为固定不变的分析方法,而不涉及随机位置的问题。

如果数据是二维空间中的点对象,那么第 i 个点可以用一对笛卡儿直角坐标来表示,如图 1.1 所示。其坐标轴的建立必须要有坐标体系作为参考。如果数据是不规则的多边形对象,我们可以用点(如"重心")来代表这些多边形,然后就可以用与点对象相同的方法去标识 $s(i)=[s_1(i),\ s_2(i)]'$。如果是规则的区域,如遥感

影像，则可以用如图 1.1(c)所示的方法标识。

(a)指定点对象的位置

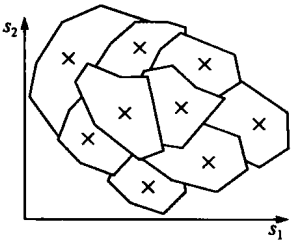
案例 i	$s(i)$		变量			
	s_1	s_2	Z_1	Z_2	…	Z_K
1	$s_1(1)$	$s_2(1)$	$z_1(1)$	$z_2(1)$	…	$z_K(1)$
2	$s_1(2)$	$s_2(2)$	$z_1(2)$	$z_2(2)$	…	$z_K(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$s_1(n)$	$s_2(n)$	$z_1(n)$	$z_2(n)$	…	$z_K(n)$



(b)指定不规则面对象的位置

案例 i	$s(i)$	变量			
		Z_1	Z_2	…	Z_K
1	1	$z_1(1)$	$z_2(1)$	…	$z_K(1)$
2	2	$z_1(2)$	$z_2(2)$	…	$z_K(2)$
⋮	⋮	⋮	⋮	⋮	⋮
n	n	$z_1(n)$	$z_2(n)$	…	$z_K(n)$

1,2,…,n
查找表



(c)指定规则区域对象的位置

案例 i	$s(i)$		变量			
	p	q	z_1	z_2	…	z_K
1	$s_1(1)$	$s_2(1)$	$z_1(1)$	$z_2(1)$	…	$z_K(1)$
2	$s_1(2)$	$s_2(2)$	$z_1(2)$	$z_2(2)$	…	$z_K(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$s_1(n)$	$s_2(n)$	$z_1(n)$	$z_2(n)$	…	$z_K(n)$

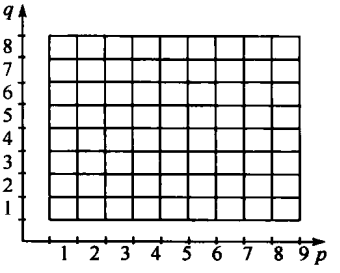


图 1.1 指定空间对象的位置(Haining,2003)

在有些情况下，式(1.1)的 $\{s(i)\}$ 所提供的地理参考信息必须要有与之相邻的对象信息作为补充，这种信息不仅要描述哪两个区域是相邻的，而且要量化这种相邻关系的紧密程度。很多空间统计模型，如空间回归模型，都需要用到这种邻接信息。

值得注意的是，在本书中的很多地方，变量 Z_1，…，Z_k 将被划分为若干个组并给予不同的标识。在数据建模过程中，被模拟的变量被标记为 Y，其若干解释变量则被标记为 X_1，…，X_Q。

空间互动数据记录网络中不同位置之间的流量，观测值 $y_{ij}(i, j = 1, \cdots, n)$ 中的每一个元素对应空间位置 i 和 j 之间的人口、货物、资金、信息和知识等的移动量。这些空间位置可以是点，也可以是区域或者分区。这些数据以“源地-汇

地”或者空间互动矩阵的形式存储。

$$\begin{array}{c} \\ \text{源地位置} \end{array} \begin{array}{c} \text{汇地位置} \\ \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{n'1} & y_{n'2} & \cdots & y_{n'n} \end{bmatrix} \end{array} \tag{1.2}$$

式中，行列数分别对应源地和汇地的数量，第 i 行第 j 列的元素 y_{ij} 记录了源地 i 到汇地 j 之间观测到的流动量。在一些特别的情况下，每个空间位置既是源地又是汇地，即 $n'=n$。要建立每个源地和终点的空间位置的地理参照，可按照之前所述的对空间对象的处理方法进行。

1.4　空间自相关

空间数据分析的基本原则是空间相关性和空间异质性，空间相关性是指不同事物在空间上距离越近，其相似程度就越高。这种数值与距离之间成反比的关系首先是托普勒第一定律所总结的---“任何事物之间都是有联系的，但是邻近事物之间的联系程度要高于相隔较远的事物”。空间相关性是各种空间数据分析的核心参数。空间异质性指均值和方差在不同子区域之间存在差异，是空间数据的另一个基本特征，尚未得到较好的挖掘利用(Wang et al.，2010)。在现有的空间分析体系中，空间指的是地理空间，用地理距离或者几何距离来度量。实际上，还存在其他各种空间，如光谱空间(Lees，2006)、属性空间(Wang et al.，2011)、社会空间(Watts，2002)、基因距离(Fraser et al.，2009)等，其数据与地理空间存在对应关系，是传统空间分析可以扩展的研究和应用领域。

如果邻近事物(位置相近)的观测值相似，则其整体而言表现的是正的空间自相关。与之相反，负空间自相关指的是空间上越邻近的事物表现出越不相似的趋势(与 Tobler 的定律相反)。当变量值与空间位置无关的时候，空间自相关为 0。值得注意的是，空间自相关与传统统计学通常要求的样本独立性相左，是空间数据分析区别于其他形式的数据分析的标识。

定义空间相关性的关键环节在于确定空间位置的邻接状况。可以这样认为，那些围绕着一个给定点的事物会影响该点的观测值。只是，我们还必须承认，决定邻接关系的过程多少有一些武断。

正规的表达空间相邻关系的方式是用一个 $n\times n$ 的空间邻接或者权重矩阵 $\boldsymbol{W}$ 表示。

$$\boldsymbol{W}=\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{bmatrix} \tag{1.3}$$

式中，n 为空间位置的数量。矩阵中的各个元素标记为 W_{ij}，对应对象 i 和对象 j 之间的相邻关系。矩阵对角线上的元素设为 0，而非对角线行的元素 $W_{ij}(i\neq j)$，如果 i 和 j 在空间上相邻，则被赋予非零值(对于二进制矩阵就是 1)，否则就是 0。

对于多边形对象，如图 1.2(a)所示的简单的 9 个区域，我们通常用是否共边为判断条件建立一阶空间邻接关系，在此基础上，图 1.2(a)可以重新表示为图 1.2(b)所示的图形。如果分区 i 和分区 j 是相邻的，则 $W_{ij}=1$，否则 $W_{ij}=0$。我们可以派生出一个如表 1.3 所示的权重矩阵 $\boldsymbol{W}$，该矩阵示范了一个最简单的构建权重矩阵 $\boldsymbol{W}$ 的方法。

对于一个规则的正方形网格，有三种确定邻近关系的方式可供选择：①rook 相邻，即以两个区域的边界是否有公共边作为相邻与否的判断条件；②bishop 相邻，即以两个区域的边界是否有公共顶点作为判断条件；③queen 相邻，即以两个区域的边界是否同时具有公共边和公共顶点作为判断条件。

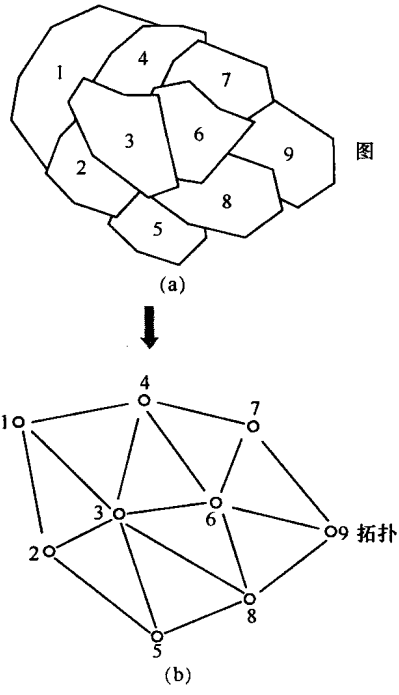


图 1.2　分区体系
(a)一个简单的离散分区拼接图；
(b)以图的方式表达拓扑

表 1.3　来自图 1.2 所列分区图的一阶二进制空间权重矩阵 $\boldsymbol{W}$

	1	2	3	4	5	6	7	8	9
1	0	1	1	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	1	1	0	1	1	1	0	1	0
4	1	0	1	0	0	1	1	0	0
5	0	1	1	0	0	0	0	1	0
6	0	0	1	1	0	0	1	1	1
7	0	0	0	1	0	1	0	0	1
8	0	0	1	0	1	1	0	0	1
9	0	0	0	0	0	1	1	1	0

依据所选择的判断标准不同，平均一个区域将有 4 个(rook，bishop)或者 8 个(queen)相邻区域。这意味着非常不同的相邻关系的结构，甚至在不规则多边形区域的情况下，我们还需要确定到底能否将仅有公共顶点的多边形视为相邻。

空间邻接性通常定义为不同位置之间的一个距离函数。在这个意义上来看，如果两个对象之间的距离在一个选定的范围之内，则可认为它们是相邻的。从本质上讲，空间权重矩阵记录了“图”的计算理论的拓扑数据集(节点和连接)。

更高阶的相邻关系以一种递归的方式定义，即如果对象 A 与对象 B 一阶相邻，而对象 B 又是另一个对象 C 的低阶相邻，那么对象 A 可以认为是对象 C 的高阶相邻。例如，在图 1.2(a)中，区域 1 和 2 是区域 3 的一阶相邻，区域 3 是区域 6 的一阶相邻，因此，区域 1 和 2 就是区域 6 的二阶相邻。

显然，一个特定的空间布局能够派生出很多的空间权重矩阵，特别是当空间权重矩阵不是二进制，而是采用能够反映空间单元 i 和 j 的互动程度的任何可能的数值(如距离反比)的时候。

表 1.3 中列出的这类矩阵使我们可以开发空间自相关程度的计量方法。现在有很多的检测和指示空间自相关的方法，应用较多的是 Moran 的方法(Moran，1950；Cliff and Ord，1973；1981)。在局部尺度上，Getis 和 Ord 的统计方法(Getis and Ord，1992；Ord and Getis，1995)和 Anselin 的 LISA 统计方法(Anselin，1995)可以用于估算在特定地点的空间自相关程度。我们在第 2 章将对其进一步讲述。

1.5　空间数据的任意性

空间数据分析依赖于数据的质量，好数据是可靠的，很少或者没有错误，可以很放心地使用。但是，实际上几乎所有的空间数据在一定程度上都是有缺陷的，错误既可能出现在空间对象的位置描述上，也可能出现在属性描述上(葛咏等，2003)。例如，当测量位置的时候，每个坐标点都有可能存在误差。

属性的误差可由收集、存储、编辑或者查询等环节产生，也可来自于测量过程中固有的不确定性因素(Haining，2003)，以及从抽样方式和数据统计中产生(Wang et al.，2010)。解决这个问题的办法是在相应环节采取必要措施：①或者设计有效算法进行纠正(Heckman，1979；Wang et al.，2011)，以避免让错误数据影响研究结果；②或者对误差进行度量，对误差传递进行模拟，最后对结果做出不确定性说明。

空间聚集的特定形式(如尺度、形状或者组合)也会影响到分析的结果(Openshaw and Taylor，1979；Baumann et al.，1983)。这个问题通常被看做是可变面元问题(MAUP)，该术语来自于这样一个事实，即面单元并非是自然形成的，而通常是人为随意构造的。基于聚集数据的研究在流行病学中经常被称为生态学研究

(ecological study),与病例-对照研究(case control study)、队列研究(cohort study)等相区别。

保密或隐私方面的限制通常也导致了最基本的观测单元数据不能发布,而代之以一组武断的面域汇总。例如,在人口调查中,每个调查到的个人或者家庭的数据一般不会发布,而只发布大片区的人口普查数据。任何时候进行面数据分析建模都会遇到这个问题,该问题包括两方面的影响:一个来自于在保持整体规模和地域单元数量不变的条件下,选择不同的面域边界(区域划分的影响);另一个来自于减少面域单元的数量而增加每个单元的大小。MAUP 没有解析解(Openshaw,1981),但是可变面元问题通过模拟大量的替代系统的地域单元(Longley et al.,2001)来调查到。这种系统显然可以采用很多种不同的形式,它既与空间分辨率有关,也与区域形状有关。

近年来,有关数据适用性的问题引起了越来越多的关注。在一些发表了的研究成果中,我们经常可以发现研究者使用的数据是一种空间尺度,而得到结论却与一个更精细尺度的过程相关,出现"以全概偏"的错误,即生态谬误(ecological fallacy)。众所周知,这种生态谬误会导致我们对技术力量的错觉,并且使我们的结论变得没有意义(Getis,1995)。生态谬误和可变面元问题一直以来都被认为是空间数据分析运用中存在的问题,并且,通过空间自相关的概念,它们被理解为是相关联的问题。因此,在空间分布数据进行分析中需要考虑单元选取的合理性,对最终统计结果解释时需要说明生态研究本身的限制性。

此为试读,需要完整PDF请访问: www.ertongbook.com