

-13

普通高等教育“十一五”国家级规划教材

空间数据分析教程

王劲峰 廖一兰 刘 鑫 编著

科学出版社

北京

P208-43
W231

第 6 章 点格局识别	(74)
6.1 样方分析	(74)
6.2 最邻近距离统计	(76)
6.3 Ripley's K 函数	(79)
第 7 章 点数据插值	(82)
7.1 趋势面方法	(82)
7.2 反距离加权法 (IDW)	(84)
7.3 Kriging 方法	(87)
7.4 Co-Kriging 方法	(89)
7.5 核心估计函数法	(95)
7.6 “3G”方法	(96)
第 8 章 格数据统计	(101)
8.1 空间自相关	(101)
8.2 可变面元问题	(110)
8.3 空间热点探测	(117)
第 9 章 格数据回归	(126)
9.1 通用模型	(126)
9.2 空间滞后模型	(126)
9.3 空间误差模型	(130)
9.4 地理加权回归	(132)
第 10 章 决策树	(141)
10.1 原理	(141)
10.2 案例	(142)
10.3 算法	(148)
第 11 章 贝叶斯网络	(149)
11.1 原理	(149)
11.2 案例 1: 出生缺陷预测	(150)
11.3 案例 2: 交通流预测	(161)
11.4 数学模型	(164)
第 12 章 人工神经网络	(166)
12.1 原理	(166)
12.2 案例	(167)
12.3 数学模型	(174)
第 13 章 粗糙集	(176)
13.1 原理	(176)

引 论

0.1 举 例

出生缺陷,是婴儿死亡和残疾的主要原因,是指任何功能或结构异常,在出生或其后表现出来的事件。出生缺陷是由出生前的某些因子作用引起的,包括遗传性和获得性的。但是与遗传和(或)环境关联的风险因子很难精确地分离开来。空间统计以其独特的切入点对此实现突破。以下以中国山西和顺县出生缺陷的环境与遗传因子识别为例演示(Wu et al., 2004)。

和顺县地处山西省境东陲,太行之巅,东西长 75km,南北宽 30km,总面积 2250km², 326 个行政村,总人口 14 万(图 0.1(a)),其中农业人口 11.8 万;地势高

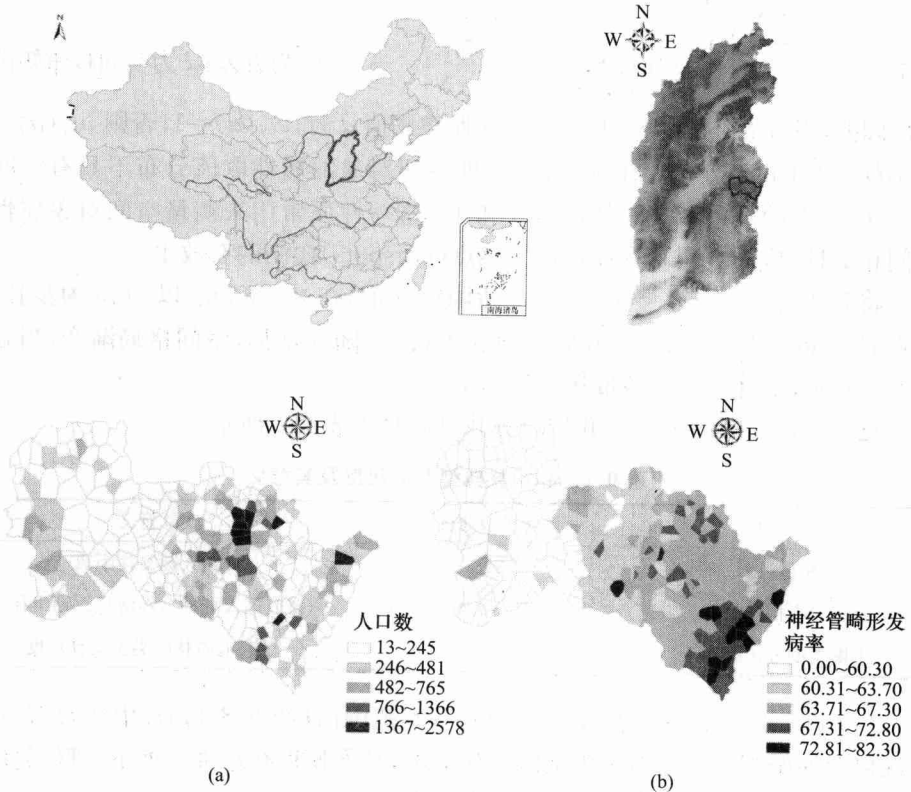


图 0.1 和顺县人口(a)、和顺县神经管畸形发病率(b)

峻,以山地、丘陵居多,一般海拔在 1300m 以上;属温带大陆性气候。春季干燥多风,夏季温暖多雨,秋季凉爽,阴雨较多,冬季漫长而寒冷。年平均气温 6.3℃,1 月平均气温零下 10℃左右,年降水 593mm,霜冻期为 9 月中旬至次年 5 月中旬,无霜期 124 天;为全国重点产煤县之一,全县经济以农业为基础,主要种植玉米、谷子、山药及莜麦、荞麦等杂粮;煤炭工业是主导,煤炭、化工、建材、冶金四大行业是主体,有县属焦化厂等工厂。

获得各村($i=1,2,\dots,N;N=326$)4 年的神经管畸形累计发病人数,并计算发病率(图 0.1(b)),记为 y_i ,使用局域 Getis G^* 统计 $G_i^*(d)$ 探测发病热点并与怀疑可能的致病因子空间格局比较,推断研究区的神经管畸形发病原因,提出防控措施。

$$G_i^*(d) = \frac{\sum_{j=1}^N w_{ij}(d)y_j - \bar{y} \sum_{j=1}^N w_{ij}(d)}{S(i) \sqrt{\{N \sum_{j=1}^N w_{ij}^2(d) - [\sum_{j=1}^N w_{ij}(d)]^2\} / (N-1)}} \quad (0.1)$$

式中, $\bar{y}_i = \frac{\sum_{j=1, j \neq i}^N y_j}{N}$ 为均值; $S(i) = \sqrt{\frac{\sum_{j=1, j \neq i}^N y_j^2}{N} - (\bar{y}_i)^2}$ 为方差; d 为空间权重矩阵的距离阈值,当村落 j 与村落 i 的距离小于距离阈值 d 时, $w_{ij}(d)=1$, 否则 $w_{ij}(d)=0$ 。 $G_i^*(d)$ 近似于正态分布。在零假设下,即空间对象的属性取值分布不具有空间相关性, $G_i^*(d)$ 的期望和方差分别为 0 和 1。这一性质常用来衡量空间对象属性的空间相关性,成为空间事物和现象的热点(hotspots)探测的有效手段。

将和顺县 y_i, N 输入 $G_i^*(d)$, 不断调整 d 值,在 0~30km,以 1km 为步长,发现在 $d < 7\text{km}$ 时, $G_i^*(d)$ 为空间聚集状(图 0.2),随 d 增加,空间格局渐变,当 d 达到 22km 时,出现明显的条带状(图 0.3)。

这种空间尺度现象提示我们应寻找其解释如表 0.1 所示。

表 0.1 和顺县典型距离尺度及其意义

统计项	距离值 d/km	实际意义
偏僻村落距最近村落距离	5.848	日常人际交往距离
乡镇中心相距距离	6.165~9.309	研究区人群社会经济日活动半径
土壤类型分辨距离	19.5~30	土壤、地质状况类型变异尺度

(1) 在该区的人群社会经济活动的基本范围内(约 6.84km),生活习俗、经济状况以及通婚圈范围等对出生缺陷产生影响,从而使得在这种尺度下,神经管畸形出生缺陷的空间分布热点呈现聚团分布状态。

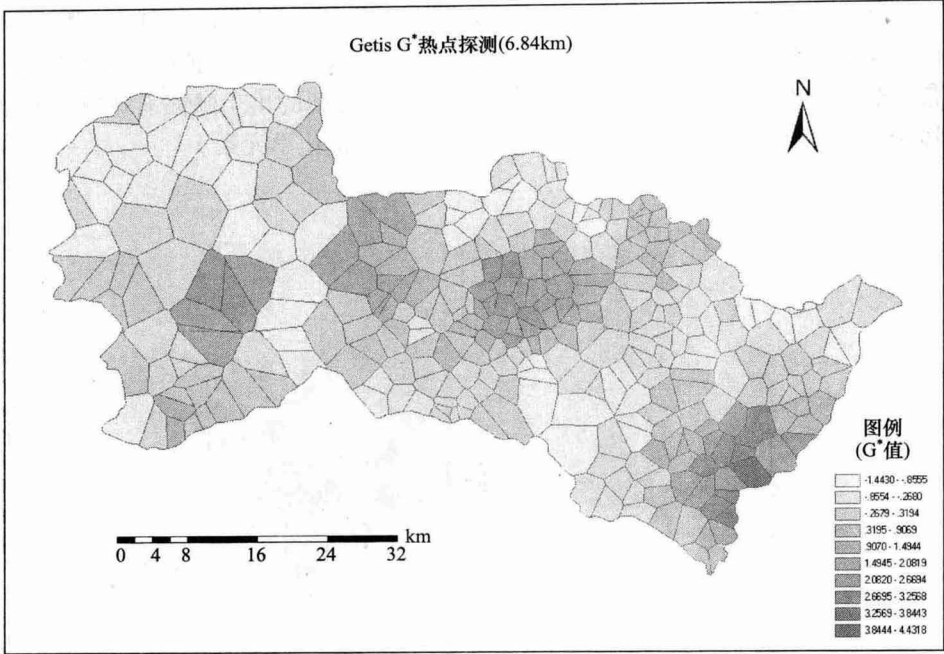


图 0.2 社会活动半径距离尺度下聚团形热点区域分布(6.84km)

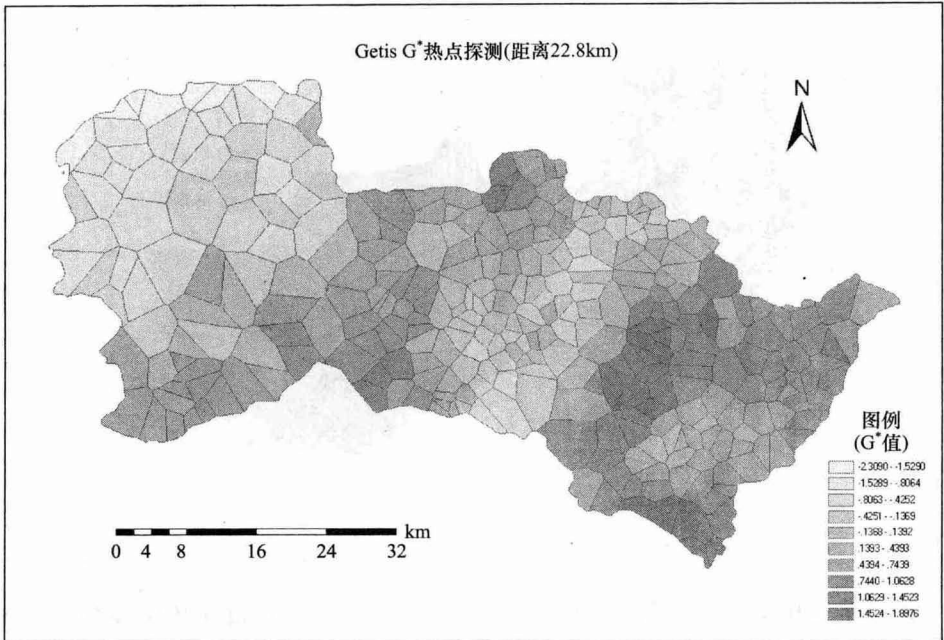


图 0.3 地质、土壤变异距离尺度下条带形热点区域分布(22.8km)

(2) 该区的地质、土壤等自然环境状况具有条带状分布的特点(图 0.4、图 0.5),故当热点探测采取土壤变异尺度作为空间权重距离阈值时,其结果呈现条带状热点分布,这种结果表明了地质环境对神经类型的出生缺陷有影响。自然环境中可能伴随岩石类型、土壤类型的分布,存在异常化学元素。

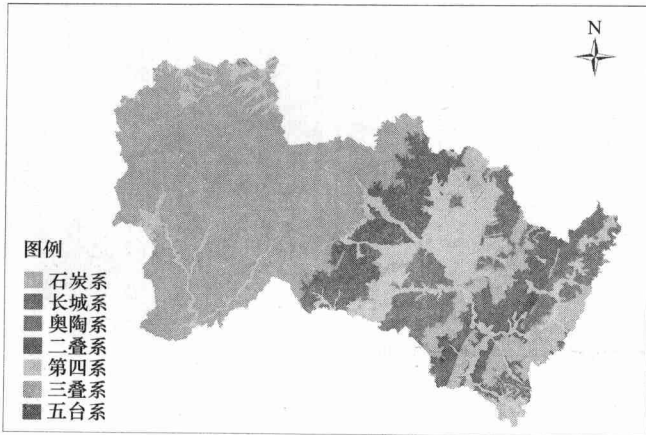


图 0.4 和顺县岩性类型分布示意图

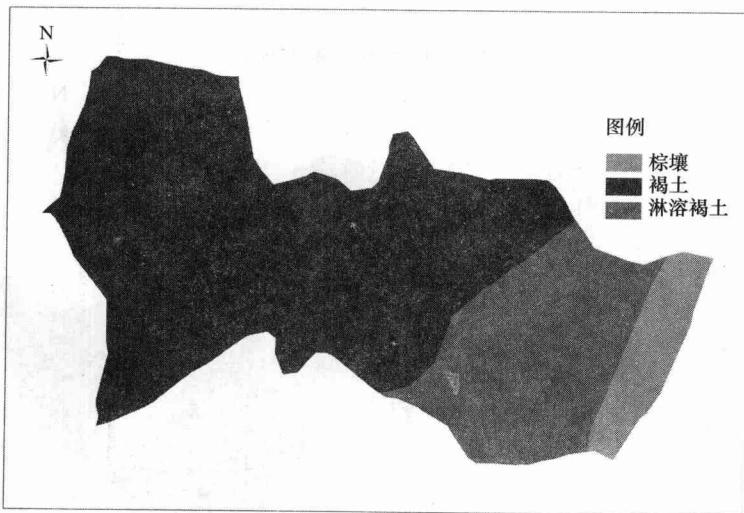


图 0.5 和顺县土壤类型分布示意图

进一步运用地理探测器(Wang et al., 2009a),可以发现健康风险多种影响因素的交互作用:方式和程度。生理代谢组学测试(Zhang et al., 2008)验证了地理探测器的发现。

0.2 空间分析理论体系

从分析方法的角度,空间过程或数据分为三类:连续数据,如温度分布等,钻孔或土壤采样可通过空间插值生成连续数据;多边形数据,无论是规则还是不规则的遥感图像像元,如统计单元内的社会经济数据;点数据,其空间位置是重要的,不涉及属性值,如禽流感暴发点的空间分布等。每类数据可以用同样的空间数据分析方法。空间过程的一般形式(Cressie, 1991)为

$$\{Z(s); s \in D\} \quad (0.2)$$

式中, D 为域或“研究区”, $D \subset \mathbb{R}^d$ (实数空间), d 为维数。假设 $Z(s)$ 在某一点 s 上是一个随机变量, $z(s)$ 是其观察值。下面具体地介绍空间数据。

连续数据(continuous data), 亦称地统计数据(geostatistical data), 这里 D 为 \mathbb{R}^d 的一个连续固定子集, 如一个国家内在一些地点上抽取的臭氧样品、一个野外场地内抽取的雪深样本、一系列气象台站的温度值、不同点测量的高度值、土壤样品中的氮浓度、湖水样品中的污染浓度等。

多边形数据(polygon data), 也称为面数据(areal data)、格数据(lattice data), 这里 D 是 \mathbb{R}^d 的一个可数但是固定的子集, 如用节点表示的格网。举例, 一个县里的疾病患者数目、一个大的连续地点的动物计数、果园内每颗树上的蔬果数目、一个道路系统上每个道路段的机动车事故数目、每段河流的鱼数等。

点数据(point data), 这里 D 是 \mathbb{R}^d 的一个随机子集。假如 $Z(s)$ 是点 $s \in D$ 上的随机向量, 则它就是标注点过程, 如果 $Z(s) \equiv 1$, 即一个退化随机变量, 那么, 仅 D 是随机的, 称为空间点过程。举例, 森林中树木的位置、天空中恒星的位置、一个区域内闪电攻击的位置、肺癌患者的居住位置、动物的出生位置等。

1854年, John Snow 通过对伦敦霍乱暴发病例的空间分析发现了传染源, 从而控制了疫情的继续传播, 成为空间数据分析和流行病学两个学科领域的共同起源。空间连续数据分析的理论研究起源于地质学的钻孔数据空间插值(Matheron, 1963; Issaks and Srivastava, 1989; Christakos, 2005); 空间多边形数据(或称格数据)分析方法起源于社会经济统计单元数据回归(Cliff and Ord, 1981; Anselin, 1988; Haining, 1990, 2003; 应龙根和宁越敏, 2005)和计量地理学(Fotheringham et al., 2000; 张超, 1984; 秦耀辰, 1994; 徐建华, 2002; 朱长青, 2006); 点数据分析起源于生态学样方分析(Diggle, 1983)。另外, 空间点状或连续数据之间的空间关系是通过点间距离或半变异函数来表达的; 格数据的空间关系通过多边形之间的连接矩阵来实现和表达。因此两种类型的数学模型不同。

实际上,空间数据类型可以互相转换,反映不同的问题。例如,上节神经管畸形发病率在 326 个行政村的空间分布,属多边形数据;若以发生和未发生神经管畸形制图,则形成点数据;若将 326 个行政村神经管畸形发病率用等值线表达,则生成连续数据;连续数据栅格化生成(规则)多边形数据,等等。Fotheringham 等(2000)将连续数据分析的核心内容 Kriging 模型和多边形数据分析的核心内容 SAR/MA/CAR 回归模型统一到一个建模体系内。不确定性始终贯穿于空间数据及其转换之中(柏延臣和王劲峰,2003;葛咏和王劲峰,2003;史文中,2005)。

近年,随着数据从单纯的空间数据到时空数据的积累,科学研究从以发现规律和科学预报为目的,发展到科学调控的理念升华、学科交叉与方法互鉴,时空数据分析(Christakos,2000)、时空运筹(Wang et al.,2002a,2008;Wang and Li,2008;郑新奇,2004),数据智能计算(Fischer and Leung,2001;Li et al.,2008;黎夏等,2007),得到了发展,成为新的增长点。遥感是地学计算有广阔前景的一个应用领域(周成虎等,2009);随着空间数据获取的方便性大为增加,在拥有共同起源 150 年之后,近年,健康领域对空间分析理论技术需求大量增加(McMichael,2001;Wang et al.,2006;Lai et al.,2009;谭见安,2004),而健康领域丰富的时空病例数据、明确的研究对象、可验证的研究结果,为空间分析理论研究提供了理想的实践领域。

0.3 本书结构

依据上节讨论的空间数据分析的理论体系,本书内容包括当今主流的空间数据可视化与探索分析、空间统计学、空间智能计算、空间运筹和时空分析以及空间分析软件包,共 22 章。每章大体遵循问题的提出、原理、案例、软件操作和数学模型的体例,达到学以致用目的。

表 0.2 是空间分析的理论体系框架以及各章在该体系中的位置和具体内容。空间分析的研究目标包括:空间数据的可视化和探索分析、参数获取、格局识别、空间预报、空间运筹、时空分析等内容;空间分析的研究对象包括点数据和格数据,这些数据属性包括位置和数值;分析方法包括统计方法和智能计算类方法。

表 0.2 本书结构

研究内容	输入	统计方法		智能计算
		点数据	格数据	点数据、格数据
可视化和探索分析		第 1 章 GIS 简介;第 2 章 地图分析;第 3 章 探索性空间分析		

续表

研究目标	属性	统计方法		智能计算
		点数据	格数据	点数据、格数据
参数获取	数值	第 4 章 空间相关性和异质性;第 5 章 空间抽样		第 10 章 决策树;第 11 章 贝叶斯网络;第 12 章 人工神经网络;第 13 章 粗糙集;第 14 章 支持向量机;第 15 章 粒子群优化算法;第 16 章 期望最大化算法
格局识别	位置	第 6 章 点格局识别		
	数值		第 8 章 格数据统计	
空间预报	数值	第 7 章 点数据插值	第 9 章 格数据回归	
空间运筹	数值	第 17 章 空间运筹		
时空分析	数值	第 18 章 BME 模型;第 19 章 演化树预报模型		第 20 章 Meta 建模
软件介绍	数值	第 21 章 空间统计学软件包		第 22 章 空间智能计算软件包

第 1 章 GIS 简介

生活中,我们常常会面临这样的问题:采用哪条路线会使到达目的地的距离最短?如何在综合考虑到达超市、学校、公司、游乐场等设施的方便程度后,挑选一处合适的住宅?在过去的一年里,某县的土地利用情况发生了怎样的变化?城市规划中如何才能合理地布置地下管线?对于某一类疾病的患病人口,在空间上呈现怎样的分布?在一次伴随有大风的森林火灾中,火势将如何发展?

对于上述问题以及其他更多的类似问题,都与地理环境及其地理过程密切相关。地理信息系统领域的人都很清楚,要回答上述问题,就需要访问具有多维(x 、 y 、 z 空间坐标, t 时间坐标, 属性)的、大容量的地理信息(Longley et al., 1999)。

地理信息系统(geographical information system, GIS)是一种信息查询、分析和决策支持系统,其特点是存储和处理的信息是经过地理编码的,地理位置及与该位置有关的地物属性信息成为信息检索的重要部分。在地理信息系统中,现实世界被表达成一系列的地理要素和地理现象,这些地理特征至少由空间位置参考信息和非位置信息两个部分组成(邬伦等, 2001)。

从 20 世纪 60 年代至今, GIS 已迅速发展成为一个独特的研究领域,并应用于区域规划、土地管理、水利水资源管理、旅游管理、城市管理、交通、卫生、农业、军事等领域,形成一个全球性的重要行业。

1.1 举 例

GIS 是用来管理、分析空间数据的信息系统,几乎所有使用空间数据的部门都可以应用 GIS,以提高管理水平。本节简要介绍地理信息系统在一些具体领域的应用。

1. 环境保护

随着经济的发展,环境污染直接影响了人们的生活质量,环境质量问题也得到了越来越多的重视。在环境保护建设中, GIS 作为信息工具平台和信息服务平台,能够把各种环境信息同地理位置和有关视图结合起来提供给环保工作者。其最大的特点在于把环境中的各种信息与反映地理位置的图形信息有机结合在一起(图 1.1),并根据需要对这些信息进行相关和综合分析。GIS 技术被充分利用到环境领域中,在提高环境保护工作效率的同时,也影响着环境保护工作方式的转变。



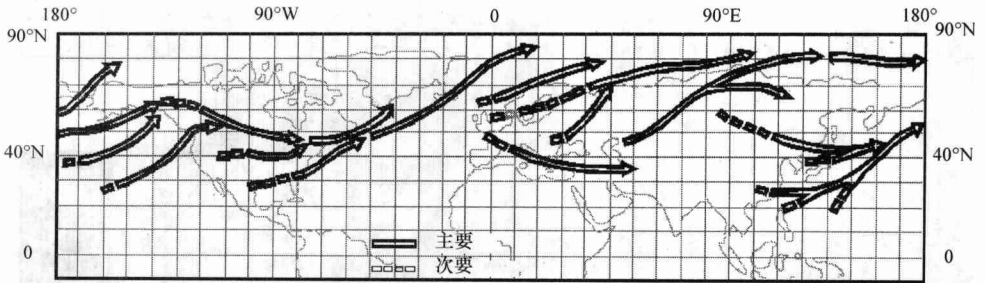
图 1.1 GIS 在环境中的应用示例

2. 应急减灾

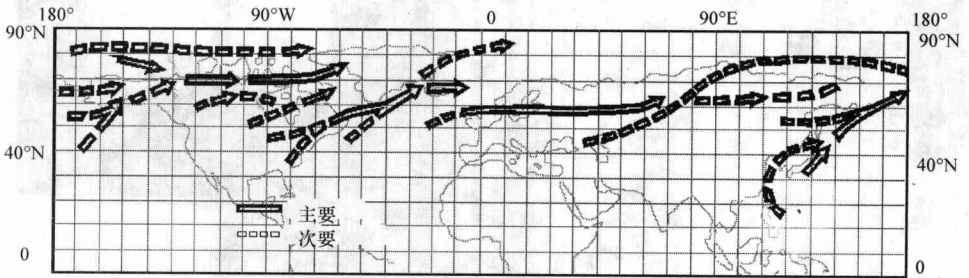
GIS 是评估潜在危险的强大工具,评估灾害可能在哪里发生,它们可能造成什么样的影响、伤害和损失等。GIS 把事发位置信息、追踪路径、传感器、视频以及其他与 GIS 数据相关的动态数据(影像、高程、街道、重点基础设施等)与交通、医院、气象结合起来,能够为决策者提供有力的支持。当危机出现时,GIS 会为应急行动计划的制订、毁坏情况的评估以及灾害信息的共享提供相关信息和帮助(图 1.2)。GIS 支持应急管理的所有阶段,包括灾情缓解、预防和准备、快速反应以及恢复重建。

3. 交通运输

GIS 在交通方面的应用得到了广泛的重视,并形成了专门的交通地理信息系统 GIS-T(GIS-transportation)。它是 GIS 在勘测设计、规划、管理等交通领域中的具体应用。GIS-T 通过地理信息系统与多种交通信息分析和处理技术的集成,可以为交通规划、交通控制、交通基础设施管理、物流管理、货物运输管理提供操作平台,如运输企业可以借助路径选择功能,对营运线路进行优化选择,并根据专用地图的统计分析功能,分析客货流量变化情况(图 1.3),制订行车计划。运输管理



(a)



(b)

图 1.2 飓风 Ivan 轨迹跟踪及波浪高度预测



图 1.3 北京市某时刻道路流量状况

部门还可以利用它对危险品等特种货物运输进行路线选择和实时监控。

4. 国土资源管理

国土资源包括土地资源、矿产资源及海洋资源,这些资源都分布在一定地理空间环境中,和地理位置密切相关。在国土资源管理中,经常需要对这些资源进行空间定位、面积测算、类型调查以及权属确认等。国土资源的这些特点,决定了国土资源信息天然就是一种地理信息(图 1.4)。地理信息系统最早应用在资源环境管理中,目前已经广泛应用于资源环境(如森林、矿产、水利、农业、牧业等)管理,自然资源(如林业、地质矿藏、水资源等)调查,自然灾害(如水灾、旱灾、虫灾、震灾等)监测、预报、评估,环境保护(如水土流失、荒漠化等的治理)等方面(陆守一等,2001)。

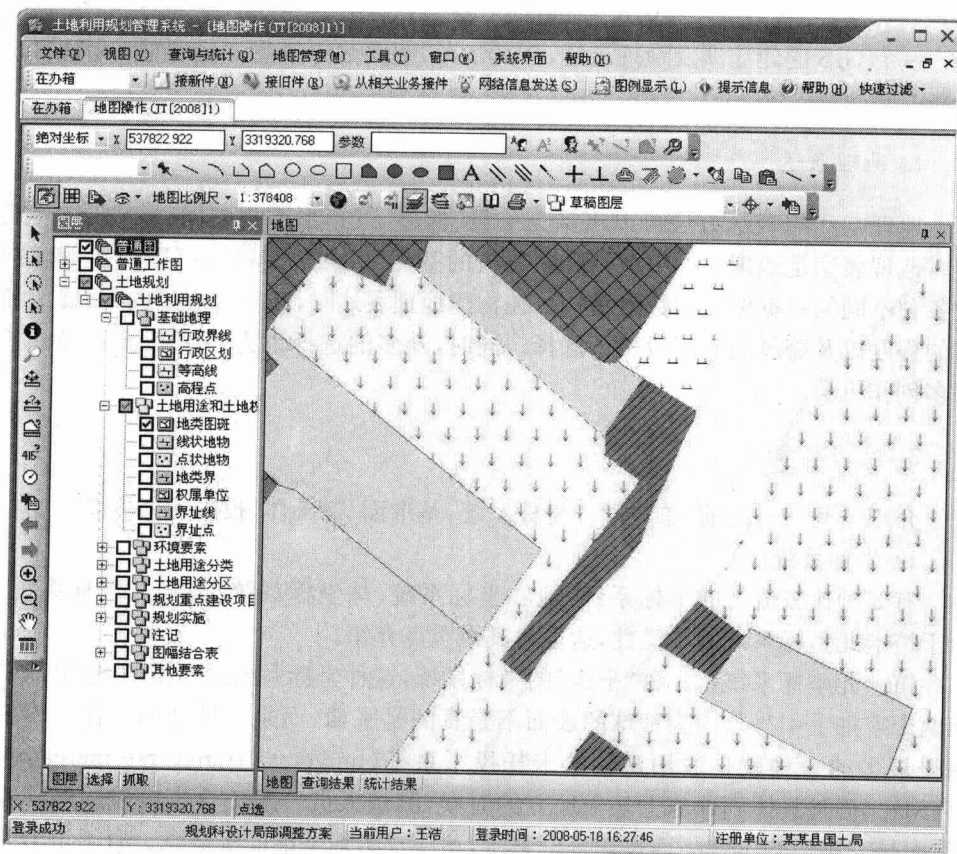


图 1.4 土地利用规划管理信息系统

5. 公共卫生

公共卫生是一个涉及微观结构和宏观系统的多分支科学,且大量数据具有空间分布特点。传染病的发生与流行、地方病的分布及病因、许多疾病的地方高发性特点以及医药卫生机构的分布等都与空间信息密切相关。同时,健康和疾病受到各种生活方式和环境因素的影响,这些具有定位特点的影响因素,为健康与环境的流行病学研究提供了有价值的线索。医学数据资料的这种空间相关特点成为 GIS 应用的前提。GIS 在公共卫生中的应用包括疾病监测及流行病学研究、环境健康研究、卫生服务利用与决策、公共卫生突发事件的应急处理等。

1.2 GIS 原理

一个 GIS 的建立涉及地理表达、空间参考、空间数据模型三个概念,这里简介其基本概念及内容。

1. 地理表达

地理空间的表达方法可以概括为矢量、栅格、三角形不规则网、Voronoi 等几类。地理表达是地理数据组织、存储、分析的基础。以此为基础,可以构造地理空间各种不同的数据模型和数据结构。在构建地理表达时,必须对表达内容、展示细节的程度以及跨越的时间段进行选择。同时,众多的选择也为 GIS 工作者提供了许多创作机会。

2. 空间参考

介绍空间参考之前,首先简介坐标系统、基准面、椭圆柱、投影 4 个概念。

1) 坐标系统

有 3 种比较流行的坐标系统:地心坐标系统、球坐标系统、笛卡儿坐标系统。由于笛卡儿坐标系统的广泛性,这里对其做重点介绍。

笛卡儿坐标系统是一种“平面”的坐标系统,这种坐标系统是二维的,这里的平面两个字加上引号是因为地球的表面不是真的是平面,而是一种球面。在实践中用得最多的一种就是通用横轴墨卡托投影系统(universal transverse mercator, UTM)。但是具体到地球上某个地方的时候,测量人员一般不会直接采用这种投影,而是一种成为本地平面投影坐标系统,这涉及本地基准面等概念。有了笛卡儿坐标系统,人们可以非常方便地在地图上进行长度、角度和面积等各种量算。

2) 基准面和椭圆柱

借助现在的卫星监测技术,我们已经知道地球其实是一个不规则的球状体。

为了应用的方便,我们通常的做法是采用椭圆柱体去逼近实际的地球的形状。椭圆柱体主要通过它的长半轴和扁率来描述。

有了这个椭圆柱体,我们就可以引出一系列的概念来帮助描述地球的形状。椭圆柱体的中心和方位构成了所谓的基准面,即利用特定椭圆柱体对特定地区地球表面的逼近而形成所谓的基准面。通过在椭圆柱体上的一系列点,我们可以定义地球的中心。如果在地球的表面建立一系列的控制点,但是由于大陆漂移的存在,这些一开始定义的控制点每年都会变化。已经流行的基准面种类非常多,有些是用来进行全球范围内的测量,有些是用来进行地球上局部地区的测量。

比较常见的基准面有:World Geodetic System 1984(WGS84),主要用于全球范围内的测量和定位;European Datum1953(ED50),主要用于欧洲地区;North American Datum 1983(NAD83),主要用于北美地区;而中国主要有北京 54 和西安 80 两种基准面。其中最为有名的就是 WGS84,GPS 系统就是采用了这种基准面,它比较好地逼近了整个地球范围。

3) 投影

需要投影的理由很简单,我们看到的地图或者在计算机屏幕看到的地图都是平面的或者说是二维的,但是地球却不是平的。所以必须想出一种办法让地球表面上的点跟平面上的点一一对应起来,而这种变换的结果就是把地球表面的点对应到笛卡儿坐标系统中。投影的方式主要有 3 种,如图 1.5 所示。每一种投影都会有不同程度的变形,要么是长度变形,要么是角度变形,要么是面积变形。

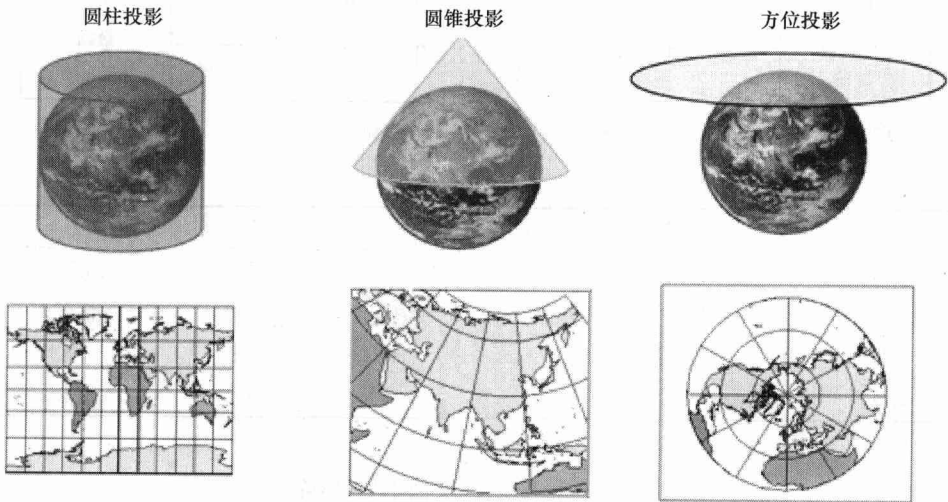


图 1.5 3 种地图投影方式

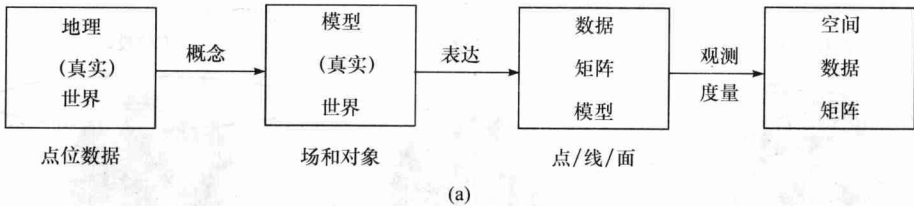
空间参考总的来说就是上面几个概念的综合,就是从比较概括的角度来说明如何把地球上的点最终转换到平面上去。空间参考首先需要一个椭圆柱,由这个椭圆柱派生出一个基准面,在基准面的基础上选择不同的坐标系统,把球面上的点转换到平面上去。

3. 空间数据模型

在计算机中,现实世界是以各种符号形式来表达和记录的,必须经过对现实世界的描述。

在地理信息系统中,有关空间目标实体的描述数据可分为 3 种类型:空间特征数据、时间属性数据和专题属性数据。对于绝大部分地理信息系统的应用来说,时间和专题属性数据结合在一起共同作为属性特征数据,而空间特征数据和属性特征数据统称为空间数据。空间数据通过观察或量测获得,或是通过进一步的计算获取。

空间数据可根据它们的收集方式、存储方式、说明内容、使用目标等,用不同的数据模型进行组织。地理信息系统中最常用的数据组织方式为矢量模型和栅格模型。在矢量模型中,用点、线、面表达世界,在栅格模型中用空间单元(cell)或像元(pixel)来表达。图 1.6(a)表达了这一从真实世界到计算机存储的“空间表达”过程。而 GIS 中所存储的属性表(图 1.6(b))是空间数据分析的具体操作对象。



k 个变量的观测数据				位置	
$z_1(1)$	$z_2(1)$...	$z_k(1)$	$s(1)$	情形 1
$z_1(2)$	$z_2(2)$...	$z_k(2)$	$s(2)$	情形 2
\vdots	\vdots		\vdots	\vdots	\vdots
$z_1(n)$	$z_2(n)$...	$z_k(n)$	$s(n)$	情形 n

(b)

图 1.6 空间表达(a)和空间数据矩阵(b)

空间数据是对现实世界中空间特征和过程的抽象表达。由于现实世界的复杂性和模糊性,以及人类认识和表达能力的局限性,这种抽象表达只能是一定程度的

接近真值,因此,数据质量发生问题是不可避免的。同时,对空间数据的处理也会导致一定的质量问题。

1.3 ArcGIS 软件使用步骤

1. ArcGIS 简介

ArcGIS是目前世界上使用最广泛的GIS软件,是由美国ESRI公司(Environmental Systems Research Institute Inc.)研发的。该公司1969年成立于美国加利福尼亚州的Redlands市,从事GIS工具软件的开发和GIS数据生产,其创始人是原哈佛大学空间分析实验室的Jack Dangermond。ArcGIS系列是ESRI公司一个全面的、完善的、可伸缩的GIS软件平台,针对不同用途,可分为如下几部分(图1.7)。

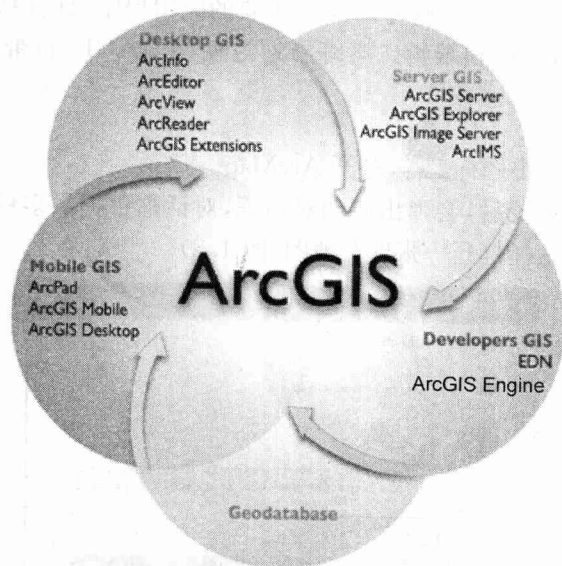


图 1.7 ArcGIS 框架

(1) 桌面 GIS。桌面 GIS(Desktop GIS)软件产品是用来编辑、设计、共享、管理和发布地理信息的。ArcGIS 桌面可伸缩的产品结构,从 ArcReader,向上扩展到 ArcView、ArcEditor 和 ArcInfo。目前 ArcInfo 被公认为是功能最强大的 GIS 产品。通过一系列的可选软件扩展模块,ArcGIS Desktop 产品的能力还可以进一步扩展。

(2) 服务器 GIS。服务器 GIS(Server GIS)包括 ArcGIS Server、ArcGIS Explorer、ArcGIS Image Server 和 ArcIMS,其用于创建和管理基于服务的 GIS 应