普通高等教育“十一五”国家级规划教材

中国科学院大学研究生教材系列

空间数据分析教程

（第二版）

王劲峰　廖一兰　刘　鑫　编著

科学出版社

普通高等教育“十一五”国家级规划教材
中国科学院大学研究生教材系列

空间数据分析教程

（第二版）

王劲峰　廖一兰　刘　鑫　编著

科学出版社
北京

内 容 简 介

环境与社会科学的数据多存在于地理空间之中，空间数据分析方法是分析挖掘地理空间数据信息和知识的有效手段。本书包括了空间探索性分析、空间统计学、机器学习和时空分析，以及空间分析软件包和案例数据等内容。本书介绍的各种方法和模型均附有真实案例和数据，以及软件和数据下载地址和操作步骤，读者可以按照书中描述重复这一过程，然后输入自己的数据迅速得到自己的计算结果。阅读本书只需要概率统计的基本知识即可。

本书可供地学和社会科学领域的本科生、研究生使用，以及地理信息科学的学者参考。

审图号：GS(2019)974 号

图书在版编目(CIP)数据

空间数据分析教程/王劲峰，廖一兰，刘鑫编著. —2 版. —北京：科学出版社，2019.4

普通高等教育“十一五”国家级规划教材　中国科学院大学研究生教材系列

ISBN 978-7-03-060789-8

Ⅰ. ①空…　Ⅱ. ①王…　②廖…　③刘…　Ⅲ. ①空间信息系统-数据处理-研究生-教材　Ⅳ. ①P208

中国版本图书馆 CIP 数据核字(2019)第 043905 号

责任编辑：杨　红　程雷星/责任校对：何艳萍

责任印制：师艳茹/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

北京市密东印刷有限公司 印刷

科学出版社发行　各地新华书店经销

*

2010 年 3 月第　一　版　开本：787×1092　1/16

2019 年 4 月第　二　版　印张：17 1/2

2019 年 5 月第五次印刷　字数：448 000

定价：59.00 元

（如有印装质量问题，我社负责调换）

第二版前言

拙著《空间分析》(王劲峰等，2006)侧重理论，本书侧重应用，两部书互相补充，形成姊妹篇。《空间数据分析教程》(王劲峰等，2010)第一版一经出版，很快售罄，曾多次加印，并荣获第二届全国优秀地理图书奖。网络评论和读者来信给予了很多鼓励和反馈意见。同时，在基于空间自相关性(spatial autocorrelation)的经典空间统计学基础上，近年关于空间分层异质性(spatial stratified heterogeneity)的统计理论取得突破，方法和应用取得长足进展，时空统计学方法进一步发展，这些都为空间数据分析提供了新的研究工具。现在市场上流行本书第一版的 pdf 版说明读者对本书还有需求。因此，作者对第一版进行了全面修订，形成第二版。新版主要变化如下。

(1)结构做了大幅调整。第一版以数据类型划分各章，第二版以问题导向划分，读者可以根据研究问题“单刀直入”选择所需要的方法。

(2)增加了数据集，更新了书中方法软件的免费下载网址。每章均有案例图解，数据可从 http://www.sssampling.cn/201sdabook/main.html 免费下载。

(3)增加了空间分层异质性的分析方法，这是本书与国内外同类书籍比较的一大特色。随着研究范围的增加和空间分辨率的提高，特别是空间大数据涌现，空间分层异质性现象凸显，是一个蕴含丰富信息的“金矿”。相应的统计理论方法取得突破，将在本版中介绍。

(4)增加了时空建模方法，编为第 16～20 章。随着空间数据的累积，形成了大量的时空数据集以供分析使用。

(5)删去了第一版中读者可以方便地从其他图书中找到的统计学一般和次要内容。

本书的读者对象是 GIS 和空间统计学的零基础者，也可供相关专业的研究人员参考。书中的案例数据和使用的软件均有免费下载地址。读者可以重复书中描述的模型计算步骤，然后输入自己的数据，得到自己的计算结果。读者需要针对研究问题选择变量，对输出结果给出专业的解释，从而形成一篇好的研究论文或报告。

在第二版修订中，得到了很多学者和同学的帮助。徐成东组织研制了用于空间分层异质性研究的 Geodetector 软件，参与了第 8 章和第 6 章的写作；姜成晟组织研制了空间抽样与统计推断软件；任周鹏撰写了第 9 章；李俊明撰写了第 11 章和第 18 章；李东岳撰写了第 16 章；孔令才撰写了第 17 章和第 20 章。张宁旭、李春林、符晨曦、彭超和刘小驰撰写了部分章节案例。本书是在广泛深入的科研和长期教学的基础上完成的，一直以来得到陈述彭、丁德文、程国栋、Robert Haining、Manfred M. Fischer、George Christakos、王志峰、宋长青、冷疏影、周成虎、刘高焕、陆峰、苏奋振、裴韬、葛全胜、庄大方、李新、黎夏、陈军、闾国年、汤国安、李满春、童小华、秦昆、徐冰、董玉祥、刘彦随、秦大河、傅伯杰、林珲、史文中、隋殿志、关美宝、吴嘉平、Zhang Tonglin、Meng Xiaoli 教授；201 空间分析研究组及历任班长

孙英君、李新虎、曹志冬、胡茂桂、王妍、任周鹏、殷倩、张杰昊、徐冰、汪洋；教学过程中邵雪梅、陈东、张学珍、宋现锋，以及马志鹏、路小娟、姚一鸣等各位老师和同学的指导、支持和帮助；指导本书写作的领导、朋友和家人没有一一列出，在此一并表示衷心的感谢。

本书得到中国科学院大学教材出版中心资助，在此表示感谢！

王劲峰　廖一兰　刘　鑫

2019 年 2 月 5 日春节

第一版前言

有空间坐标或相对位置的数据通称为空间数据，如发病率在各社区、乡村的分布，气象台站监测的温度、降雨、辐射，大气污染分布，土壤重金属含量在区域各抽样点的数值，全国各主要城市的GDP，区域社会经济调查(抽查或普查)数据，城市各路段的瞬时交通流量，遥感影像各像元的光谱值，等等。

统计学是数据描述、总结、推断、预测分析的基本方法，大多数情况下要求样本互相独立、大样本、多次重复。空间数据通常具有互相不独立性，空间异质性、不可重复性。将经典统计学理论直接运用于空间数据其结论将是有偏和非最优的。因此，经过地理学家和数学家近50年的研究发展，形成了独特的空间数据特有的分析理论。

拙著《空间分析》(王劲峰等，科学出版社，2006)一经出版，各书店和网络售书很快售罄；国内外的几位地理信息科学著名学者给予很好的评价；作者还被告知该书被剑桥大学地理系推荐为参考书；从中国大陆去美国求学的一些学子在其航空行李重量严格受限的宝贵空间里携带了此书；被同行朋友作为枕边书；作者的欣慰还特别来自于该书读者的评价，鞭策作者放下案头繁重的科研工作，撰写一本适合地理信息科学更加普及的空间分析读本。

一部成功的著作，不仅被初学者视为深入浅出的入门教材，而且也被该领域的著名学者的研究论文经常引用。其成功的秘诀可能在于用简单的语言描述深刻复杂的问题本质，而不是用较多的数学公式为主要语言。实际上，文字和数学是描述一个对象的两种工具，对于复杂的问题，纯粹用语言描述经常难以表达复杂的关系，显得力不从心，读者不知所云；而纯粹用数学描述，亦复杂，不易被读者理解其本质。真实世界的终极本质可能是简单和相互联系的，时间C、质量M和能量E分别处于三个互相垂直维度上的核心变量，竟然能够被$E=MC^2$如此简单的数学方程联系起来，反映了发现者深刻的洞察力、也提示“越本质、越简单”这一真理，在某种意义上，“越复杂、越肤浅”。科学家的任务应当是将复杂留给自己，将简单奉献给人类。是否反映了问题的本质、读者是否容易理解和可重复，是作者在写每一句话、每一个公式选取最佳表达方式的唯一标准。这是本书写作过程中始终铭记在心的。

本书是在2006年已经出版的《空间分析》专著的基础上重写，简化、添加了空间数据分析中被证明是强有力的最新成果，删略了一些过泛的内容。每个理论和模型均配有公开免费下载的软件的操作案例，运用真实典型案例，step by step的软件操作步骤截图。本书的各章的体例大体为：引言，说明该模型的用途；原理，用文字和关系图说明该模型的基本思路；案例；数学模型。据此，读者在初步了解模型的用途和基本思想后，就可以迅速地重复这些算例：输入自己的数据得到计算结果；如果读者有进一步兴趣了解具体数学模型的细节，可参考各章最后的数学模型部分。作者以为这种体例对读者学习和迅速使用空间数据分析理论是十分方便的。该书被遴选为国家级教材，供地学、环境和社会科学领域的本科生、研究生自学，并供授课老师和研究人员参考。2006年版的《空间分析》侧重理论性，而本书更侧重实用性。

我们在空间数据分析领域的研究和实践得到了OAD Scholarship、Marie Curie Fellowship、

国家留学基金、中国科学院高访基金、中国科学院、国家自然科学基金、973 计划、863 计划、国家科技支撑、科技部国际重大合作项目、国家重大科技专项的支持。感谢陈述彭、丁德文、程国栋、何建邦、周成虎、闾国年、刘高焕、黎夏、史文中、隋殿志、梁怡、宋长青、冷疏影、刘纪远、陈军、刘昌明、陆大道、郑度、李小文、孙九林、毛汉英、高晓路、应龙根、赵作权、王志峰、王道辰等许多先生对我们的长期指导和支持。感谢我们的长期指导、支持与合作者：Robert Haining(空间统计学)、Manfred M. Fischer(空间计量经济学)、George Christakos(空间随机场)、Tony McMichael(空间流行病学)、Niels Becker(生物统计学)、Katie Glass(生物数学)、Ben Reis(计算流行病学)、郑晓瑛(人口学)、杨维中(流行病学)、曾光(流行病学)、李新(遥感)、庄大方(地理信息科学)、钟耳顺(地理信息科学)、葛咏(不确定性)、关元秀(生态建模)、李连发(抽样)、柏延臣(不确定性)、王智勇(技术扩散)、朱彩英(遥感反演)、武继磊(空间统计)、孙英君(随机模拟)、何绍福(生态经济)、韩卫国(地学计算)、刘旭华(土地动力学，参与撰写本书第 19 章)、孟斌(空间统计，参与撰写本书第 8 章和第 21 章)、李新虎(空间统计)、王海起(交通优化)、李三平(不确定性)、赵艳荣(流行病学)、王磊(流行病学)、孙腾达(交通模拟)、赵永(CGE 模型)、迟文学(空间统计)、林华亮(流行病学)、冯小磊(空间抽样)、高一鸽(时空数据可视化)、曹志冬(空间统计建模)、郭瑶琴(弹性网络)、申思(空间认知地图)、徐一土(软件系统)、姜成晟(空间抽样)、常超一(空间流行病)、王娇娇(城市交通预报，参与撰写本书第 15 章)、胡茂桂(超分辨率模型)、白鹤翔(粗糙集，参与撰写本书第 15 章)、姜新利(软件系统)、吴凡(登革热评估)、马爱华(空间抽样)、李小洲(参与写作本书第 25 章)、郭燕莎(空间抽样)、胡艺(健康与地质)等。

支持和指导我们的领导、朋友和家人没有一一列出，在此表示衷心的感谢。

王劲峰

2009 年 4 月 20 日

目　录

第三篇 机 器 学 习

第四篇　时 空 分 析

附　录

引　论

0.1　举　　例

出生缺陷是指婴幼儿任何功能或结构异常，在出生或其后表现出来。出生缺陷是由出生前的遗传和环境交互作用引起的，但是与遗传和环境关联的风险因子很难分离开来。空间统计以其独特的视角突破了这一难题。下面以某县出生缺陷的环境与遗传因子识别为例演示(Wu et al., 2004)。

该县地处山区(图 0.1)，东西长 75km，南北宽 30km，总面积 2250km^2，326 个行政村，总人口 14 万[图 0.1(b)]，其中，农业人口 11.8 万人；地势高峻，以山地、丘陵居多，一般海拔在 1300m 以上，交通不便，历史以来与外界交往相对封闭；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；年平均气温为 6.3℃，1 月为–10℃左右，年降水量 593mm，霜冻期为 9 月中旬至次年 5 月中旬，无霜期 124 天；全县经济以农业为基础，主要种植玉米、谷子、山药及莜荞麦等杂粮；是全国重点产煤县之一，以煤炭工业为主导，煤炭、化工、建材、冶金四大行业是其主体。

调查获得该县 i 村($i = 1, 2, \cdots, N$; $N = 326$)4 年的神经管畸形累计发病率 y_i[图 0.1(a)]，使用局域 Getis G*(Getis and Ord, 1992)统计探测发病率热点，并将探测出来的热点区域分布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，为制定防控策略提供线索。

$$G_i^*(d)=\frac{\sum_{j=1}^{N}\boldsymbol{w}_{ij}(d)y_j}{\sum_{j=1}^{N}y_j} \tag{0.1}$$

$$E(G_i^*(d))=w_{i.}^*(d)/n \tag{0.2}$$

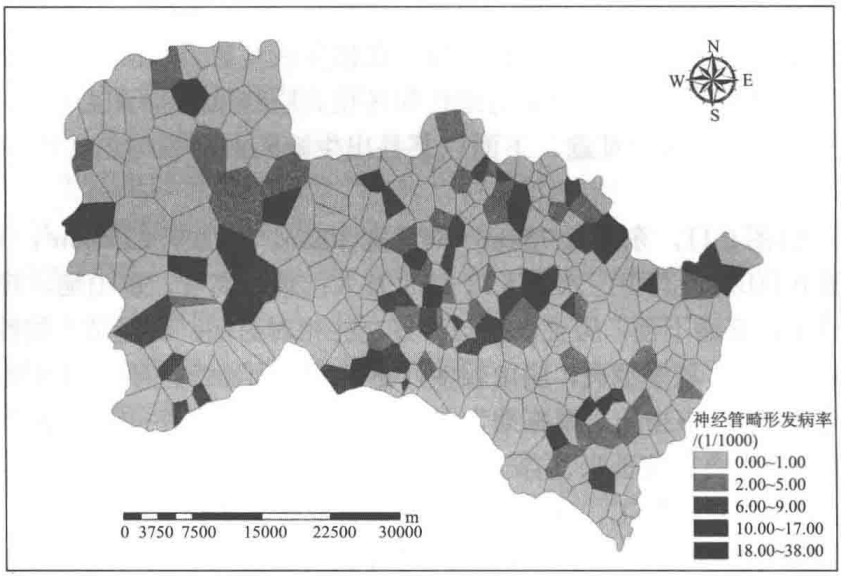
$$v(G_{i.}^*(d))= w_{i.}^*(n-w_{i.}^*)s^2/n^2(n-1)\,\bar{y}^2 \tag{0.3}$$

$$Z_{G_i^*(d)}=\frac{G_i^*(d)-E\left(G_i^*(d)\right)}{\sqrt{v\left(G_i^*(d)\right)}} \tag{0.4}$$

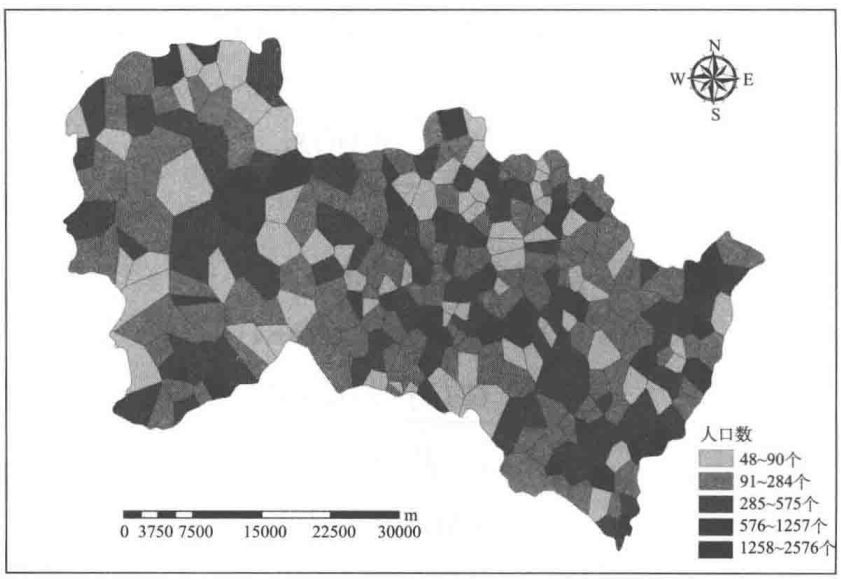
$$Z_{G_i^*(d)} \sim N(0, v(G_{i.}^*(d))) \tag{0.5}$$

式中，$\boldsymbol{w}_{ij}(d)$ 为 i 和 j 两村之间的连接矩阵，如果在指定距离 d 之内取值 1，否则为 0；也可灵活地定义为距离衰减函数。$w_{i.}^*(d)=\Sigma_j w_{ij}(d)$，$\bar{y}$ 和 s^2 分别为观测值 y 的均值和方差。$G_i^*(d)$ 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。局域 Getis G*统计量的统计检验值($Z_{G_i^*}$ 值)常用来衡量空间对象属性的空间相关性的显著性。如果 $Z_{G_i^*}$ 值为正且通过显著性检验，则表明 i 村周边村落的神经管畸形发病率与 i 村的发病率相近，存在空间聚集。

在该县范围内，0～30km，以 1km 为步长调整 d 值，发现在 d<7km 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人们寻找其解释。通过现场调查和分析数据发现典型距离尺度，含义如表 0.1 所示。



(a)



(b)

图 0.1　某县神经管畸形发病率(a)、人口数(b)

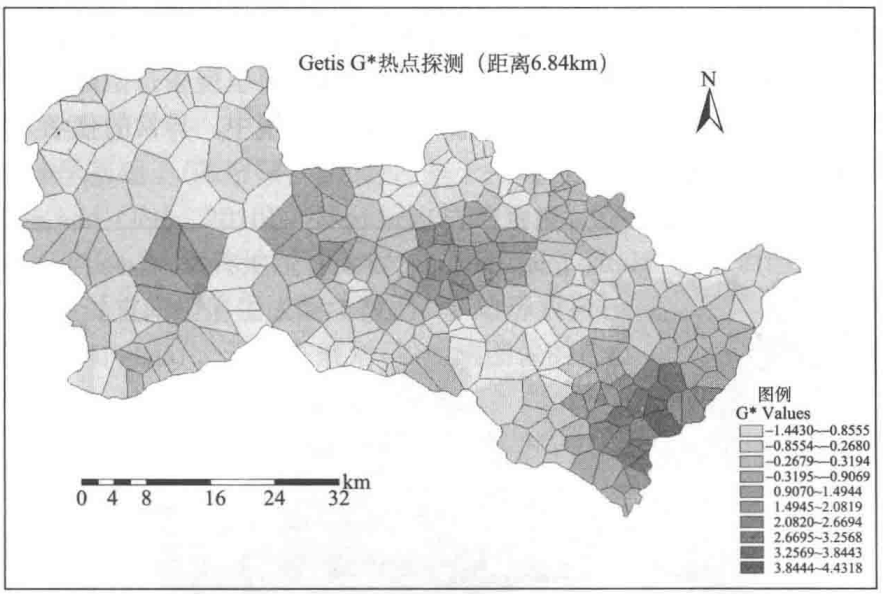


图 0.2　聚团形热点区域分布(6.84km)

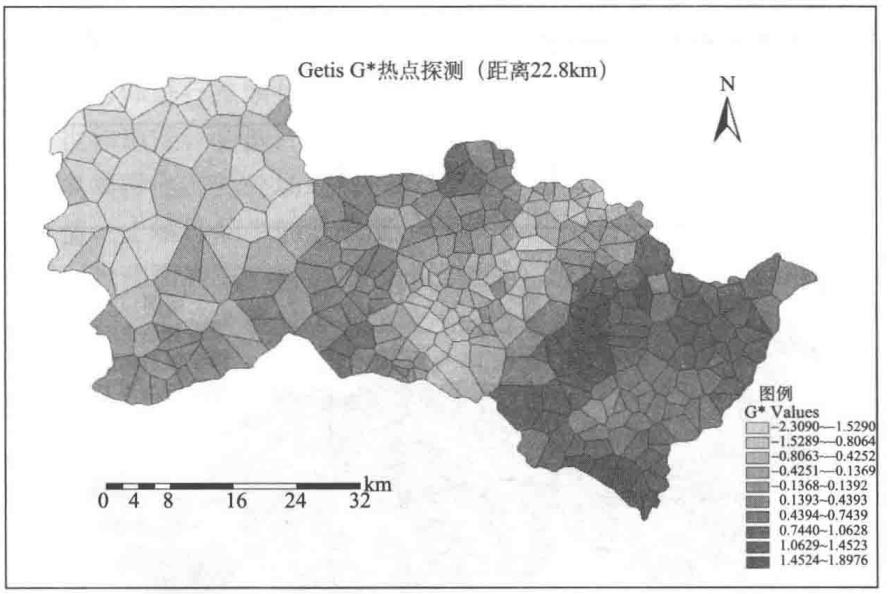


图 0.3　条带形热点区域分布(22.8km)

表 0.1　某县典型距离及其意义

统计项	距离值	实际意义
偏僻村落距最近村落距离	5.848km	日常人际交往距离
乡镇中心相距距离	6.165～9.309km	研究区人群社会经济活动半径
土壤类型距离	19.5～30km	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(6.84km 左右)，生活习俗、经济状况及通婚圈范围等对出生缺陷的发生产生影响，从而使得在这种尺度下，神经管畸形发生率呈现空间聚团分布状态。

(2)该区的地质、土壤等自然环境要素具有条带状分布的特点(图 0.4 和图 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，其结果呈现条带型热点分布，这种结果表明了地质环境可能对神经管畸形发生产生影响。自然环境中，异常的化学元素可能存在于某些特定类型的岩石和土壤中。

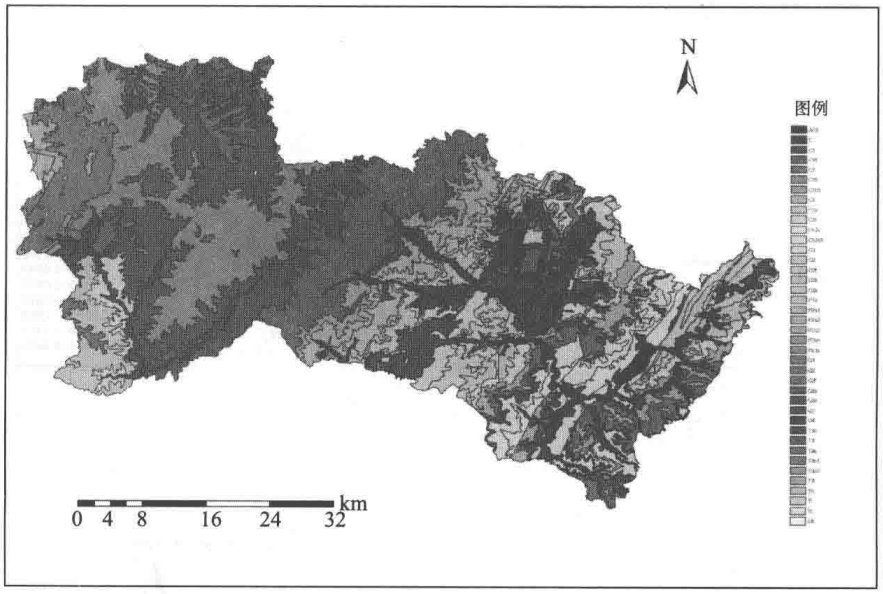


图 0.4　某县岩性分布

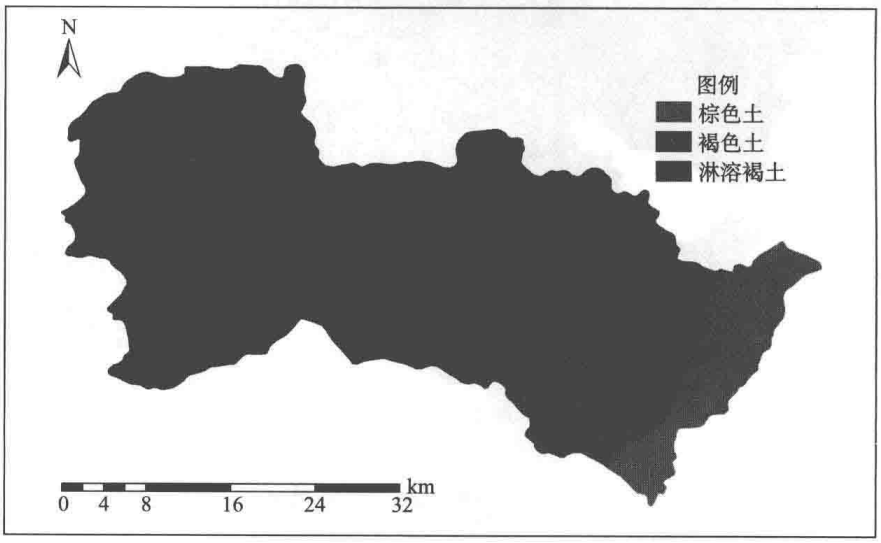


图 0.5　某县土壤分布

之后，运用地理探测器(Wang et al., 2010b, 2016a 及本书 8.3 节)量化这些定性发现，并且发现健康风险多种影响要素的交互作用的方式和程度。同时期同地区进行的生理代谢组学实验结果验证了地理探测器的发现。

0.2　空间数据分析理论体系

1. 空间数据类型

在经典统计学基础上(陈希孺，2002)，基于空间自相关性的空间数据统计学已经形成许多方法(Fischer and Getis, 2010)。从分析方法的角度，空间数据分为三类：空间连续数据(spatial continuous data)，也称地统计数据(geostatistical data)，如地表温度分布、钻孔或土壤采样数据等，可通过空间插值生成连续数据；多边形数据(polygons)，也称面数据(areal data)或区域数据(regional data)，无论是规则(如遥感图像的像元)还是不规则多边形(如记录社会经济数据的行政单元)；点数据(point data)，其空间位置是重要的，不涉及属性值，如居民点的空间分布、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。空间过程或者空间随机场的一般形式(Cressie, 1991)为

$$\{\boldsymbol{Z}(\boldsymbol{s}):\boldsymbol{s}\in D\}$$

式中，D 为域，或“研究区”，$D\subset \mathcal{R}^d$，$\mathcal{R}$为实数，d 为维数。假设在每个点 s 上的观察值 $Z(s)$ 是一个随机变量。所有点 $\boldsymbol{s}$ 上的随机变量集合$\{Z(s), s\in\boldsymbol{s}\}$形成随机变量向量 $\boldsymbol{Z}(\boldsymbol{s})$，称作空间过程；对所有点 $\boldsymbol{s}$ 上的随机变量$\{Z(s), s\in\boldsymbol{s}\}$的一次观察值全体 $z(\boldsymbol{s})$，称作空间总体(population)；对空间总体抽样(sampling)，形成一个样本(a sample)。空间统计的主要目的就是通过样本对空间总体进行推断(Wang et al., 2012a)。

连续数据，也称地统计数据，D 是$\mathcal{R}^d$的一个连续固定子集，如在一个国家内的一些地点上抽取的空气臭氧样品、野外场地内抽取的雪深样本、一系列气象台站的温度值、不同点测量的空气高度值、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，也称面数据、格数据(lattice data)，D 是$\mathcal{R}^d$的一个可数但是固定的子集，如用节点表示的格网，如一个县里各乡镇的疾病患者数目、果园内每棵树上的果实数目、一个道路系统上每个道路段的机动车事故数目、每段河流里鱼的数量、图像各像元值等。如果 $D=\{D_h, h=1, 2, \cdots, L\}$,并且 $\boldsymbol{s}_h\in D_h$, $Z(\boldsymbol{s}_h)$退化为一个常数或标注不同类型，那么，这个多边形数据集就是状态(如发达和欠发达地区)或类型量(如土地利用类型图)。

点数据，D 是$\mathcal{R}^d$的一个随机子集。假如 $\boldsymbol{Z}(s)$是点 $\boldsymbol{s}\in D$ 上的随机向量，则它就是标注点过程，如果 $\boldsymbol{Z}(s)\equiv 1$，即一个退化随机变量，那么，仅 D 是随机的，被称做空间点过程，如森林中树木位置、天空中恒星位置、一个区域内闪电攻击位置、癌症患者的居住位置、动物出生位置等。

1854 年 John Snow 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。空间连续数据分析理论分别起源于采矿的钻孔数据空间插值(Matheron, 1963；Issaks and Srivastava, 1989；Christakos, 1992)和气象要素空间插值(Gandin，1963)；空间多边形数据分析方法起源于社会经济统计单元数据的空间自相关性度量和回归(Moran, 1950, Cliff and Ord, 1981; Anselin, 1988, Haining, 1990；应龙根和宁越敏, 2005)及计量地理学(Fothringham 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；而格数据的空间关系则通过多边形之间的连接矩阵来实现和表达。因此，两种类型数据分析的数学模型

形式不同，但思路相近。

实际上，空间数据类型可以互相转换，反映不同的问题。例如，0.1 节神经管畸形发病率在 326 个行政村的空间分布，是多边形数据；若以一段时间内各行政村发生和未发生神经管畸形事件制图，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。Fothringham 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)。

2. 空间统计信息流

研究区域的所有单元集合，称为总体(population)。空间数据分析一般需要经历这样一个过程：通过观测得到总体的一个子集，也就是一个样本(a sample)，将样本带入一个统计量(estimator)对总体进行推断。这一过程的信息流如图 0.6 所示。因此，空间数据统计分析的误差是由(总体、样本、统计量)“三位一体”空间抽样与统计推断(spatial sampling trinity)所决定的(Wang et al., 2012a)。空间抽样与统计推断总是互相联系的，当总体存在空间分异并且样本量较小时，样本的不同放置位置和不同的统计量的统计推断结果差异较大，则需要使用具有样本纠偏能力的统计量进行统计推断(Wang et al., 2013a; Hu et al., 2013; Xu et al., 2013; Heckman, 1979; Meng, 2018)。

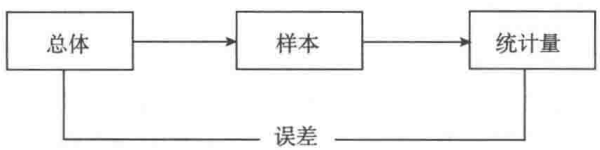


图 0.6 “三位一体”空间统计信息流(Wang et al., 2012a)

0.3 模型选择与效果评估

本书将介绍不同模型，以及其假设条件。如何选择模型？估算精度如何评估？可以根据总体-抽样-统计量“三位一体”空间统计信息流进行考察(Wang et al., 2012a)。

1. 模型选择

地理空间分布对象总体可能具有空间自相关性(用 Moran's I 或半变异函数检验，参见 3.1 节空间自相关性)，也可能具有空间分异性(用地理探测器 q 统计检验，参见 3.2 节空间分层异质性)，以及具有可变面元问题(不同的分层 strata 或等级具有不同的 q 值，按照专业知识进行分层或者按照 q 最大进行分层，参见 3.3 节可变面元问题)。

如果模型假设与研究对象总体性质是一致的，那么，就是恰当模型。因此，在选择模型时，首先应当判断对象总体的性质：

(1)如果是独立同分布的，采用经典统计学(复旦大学，1979)是恰当的。

(2)如果空间相关性强，而空间分异性弱，并且没有明确的解释变量或解释变量不可获取，那么，Kriging 方法(第 5 章 5.4 和 5.5 节)是恰当的。

(3)如果空间分异性强，而空间相关性弱，并且没有明确的解释变量或解释变量不可获取，那么，三明治(Sandwich)模型(第 5 章 5.6 节；免费软件 www.sssampling.cn)是恰当的。