黄幼才　刘文宝
李宗华　肖道纲
编　著

中国地质大学出版社

GIS空间数据误差分析和处理

GIS　空间数据误差分析和处理

GIS 空间数据误差分析和处理

黄幼才　刘文宝　李宗华　肖道纲　　编著

中国地质大学出版社

·(鄂)新登字第12号·

内容提要

地理信息系统(GIS)是一种集信息采集、贮存、分析、处理和管理于一体的计算机软件系统，现已广泛应用于国民经济各个领域，并逐渐成为各部门决策者不可缺少的技术手段。由于目前世界上还没有一种地理信息系统具有评估GIS产品质量的功能，因此评价GIS输出结果可靠性的理论和方法已成为GIS界的热门课题。

本书主要介绍GIS空间数据误差来源的确定，误差度量和削弱误差影响的方法以及GIS操作运算中的误差传播规律。它全面地反映了当前GIS空间数据误差处理和精度分析这个领域的研究方向和一些阶段性的研究成果，对于从事GIS应用和理论研究的人员来说是一本有益的参考书。

图书在版编目(CIP)数据

GIS空间数据误差分析和处理/黄幼才、刘文宝、李宗华、肖道剑编著. 武汉：中国地质大学出版社，1995.6
ISBN 7-5625-1015-1

Ⅰ.G…
Ⅱ.①黄…②刘…③李…④肖…
Ⅲ.①地理信息系统-数据-测量误差 地域分析②地理信息系统-数据处理
Ⅳ.P91

出版发行 中国地质大学出版社(武汉市·喻家山·邮政编码 430074)
责任编辑 [illegible] 责任校对 [illegible]
印 刷 武汉测绘学院印刷厂

[illegible]
[illegible]

前　言

目前，GIS 空间数据误差分析与处理已成为国际 GIS 界普遍关心和重视的基础理论研究课题之一。它涉及的领域非常广泛，在研究过程中不仅需要借用相关学科中现有的一些数据处理理论，而且还需要根据 GIS 数据来源不同、种类繁多的特点以及 GIS 运算处理方式对有关理论进行概括和发展。可以说，该理论研究的出现表明 GIS 从单纯的技术开发阶段走向基础理论研究阶段，它将结束 GIS“无自身基础理论”的历史。

本书综合了国外有关 GIS 空间数据误差分析和处理方面的大量最新文献和作者近年来的一些研究成果编著而成，它反映了这个领域当前国际上的研究动态和最新进展。全书共分六章。前三章分别介绍了 GIS 空间数据误差研究的基本指导原则和发展趋势，GIS 空间数据误差的来源、特性和分析，以及 GIS 空间数据误差的度量方法和 GIS 操作运算中的误差传播规律，这是构成 GIS 空间数据误差理论框架的最基本部分。后三章介绍了空间位置表达中的误差问题和 GIS 实际应用中所遇到的一些误差问题，总结了现有的一些空间数据误差处理方法，并展望了这个领域的未来趋势。

需要指出的是，由于目前还没有形成一套完整的 GIS 空间数据误差理论体系，大多数研究还处在“立题”阶段。因此，本书中各章节内容间的衔接不像经典理论著作和一般教科书那样紧密。每章各节基本上反映了这个领域的一个研究方向。此外，书中所介绍的某些成果还处在探索和发展之中，有的还有待完善。

作者期望，对于从事 GIS 实际应用的读者，通过阅读本书能较快地掌握 GIS 数据误差分析和处理的一些基本方法，并将其用于解决本行业的实际问题中；对于侧重基础理论研究的读者，能够较快地了解本领域研究的基本概貌和发展方向，并能逐步接触近代文献和研究工作。因此，本书可供从事 GIS 行业的科技人员和 GIS 用户以及大专院校师生参考。

在本书编写过程中得到了武汉市勘测设计研究院领导的关心和大力支持。书中作者的一些研究成果是在武汉测绘科技大学测绘遥感信息工程国家重点实验室的专项研究基金支持下完成的。武汉测绘科技大学陶本藻教授仔细审阅了本书初稿，并提出了许多宝贵意见。作者谨此一并表示最衷心的感谢！

由于作者水平所限，书中难免有不足甚至错误之处，恳请广大读者批评指正！

作　者

1994 年 7 月于武汉

目　录

第一章 绪 论

地理信息系统(GIS)是60年代初期诞生的一门新技术,它是以采集、存贮、管理、分析和描述空间物体的地理分布数据及与之相关的属性,并回答用户问题等为主要任务的计算机软件系统。1963年,加拿大测量学家Tomlinson首先提出了地理信息系统(GIS)这一术语,并组织建立了世界上第一个GIS——加拿大地理信息系统(CGIS)。由于受当时计算机技术的限制,早期的GIS功能十分简单,多带有机助制图色彩。进入70年代以后,计算机硬件和软件技术的飞速发展给GIS注入了新的活力,大大提高了GIS的功能。80年代,GIS在全世界迅速发展并推广应用,90年代已开始深入普及到各行各业。

由于地理信息系统是基于应用而形成的一门信息学科,因此以往的研究重点集中在系统的建立、功能的改善和应用方面。在GIS初步形成之后,才提出了建立地理信息理论的课题,其中空间数据误差的处理和分析被列为90年代重点研究的问题之一。

顾名思义,空间数据误差处理和分析就是对采集的空间数据确定误差的来源、性质和类型,提出度量误差的指标,分析误差在GIS空间操作中的传播机制,研究削弱误差对GIS产品质量影响的方法。也就是说,对影响GIS产品质量的录用数据误差和处理过程误差进行全面仔细地分析,并提出"治疗"方案。

本章共分3节,§1.1简要介绍GIS数据误差研究的内容、意义及其在地理信息系统中的地位和作用;§1.2介绍研究GIS数据误差的指导原则;§1.3介绍GIS数据误差研究的重要进展与若干发展趋向。

§1.1 GIS数据误差研究的内容、意义和作用

地理信息系统(GIS)是由在规划、管理、决策和预报等方面的巨大应用动力而迅速发展起来的,而支持各种应用的是经查询、分析、处理等提供给用户的图表等不同类型的GIS产品。由于生产GIS产品的"原料"——GIS的原始录用数据本身包含着不可避免的误差,描述数据的模型也只能是客观实体的一种近似,并且GIS产品的"生产"过程中——各种空间操作、处理等又会引入新的误差或不确定性。因此,人们自然有理由要问:GIS产品的质量如何?GIS所输出的图表精度和可靠性是多少?GIS综合分析、推理所得结论的精确度和可信度是多少?GIS原始录用数据中的误差和错误会不会严重干扰GIS对问题所作的结论?等等。用户在使用GIS解决具体问题的过程中,必须首先谨慎地弄清上述一系列问题,才能作出正确的决策。这一点,在以往的GIS设计中常常被忽视,使得由GIS生成的各种漂亮精美图件与其内在质量不相符合而导致决策失误。

GIS空间数据误差处理和分析就是针对上述背景而提出的研究课题,其核心是建立一套误差分析和处理理论体系。根据GIS数据误差研究的成果,未来的GIS应当在提供产品的同时,附带提供产品的质量指标,就像测量工作者在提供大地坐标时,同时提供坐标精度一样。

从应用角度看,GIS空间数据误差分析和处理的研究内容可概括为正演和反演两大问题。当GIS录入数据的误差和各种操作中引入的误差已知时,计算GIS最终生成产品的误差大小和数值的过程是误差的正演问题。反之,根据用户对GIS产品所提出的误差限值要求,确定GIS录入数据

误差和质量，则是误差的反演问题。显然，误差传播机制是解决正、反演问题的关键。需要指出的是，由于反演问题的解一般不唯一或不确定，因此研究反演问题应当特别小心。为了得到反演问题的唯一解，通常需要根据实际问题的特点，顾及先验信息，在给定的约束条件下解求，具体算法可采用“模拟法”或“解析法”。

GIS 数据误差的研究，对评价 GIS 产品的质量，确定 GIS 录用数据的标准，改善 GIS 的算法，减少 GIS 设计与开发的盲目性以及 GIS 的其他研究领域都有深远影响。另外，测量工作者利用研究 GIS 数据质量方面的优势可以率先进入空间信息交易市场，因为数据质量问题有可能会阻止其他机构的介入。

下面列举一些著名 GIS 学者的观点进一步说明研究空间数据误差问题的重要性。Goodchild 和 Dubuc(1987)将一个没有以准确数据为基础的 GIS 系统比喻为“一位体魄壮如运动员，却只有幼儿智力的人”。而 Abler(1987)则说这样的系统“能以相当快的速度生产各种垃圾，而这些垃圾看起来似乎是精美无比的”。Goodchild 和 Gopal(1990)指出：“由于目前的大多数矢量库 GIS 都能执行求交、覆盖或生成缓冲区等操作，但不考虑其精度。结果当用户发现 GIS 产品提供部门的建议与地理实况相差大得令人吃惊时，GIS 公司会在用户中立刻失去信誉”。因此，GIS 若要生存发展下去，必须花大力气从理论上研究 GIS 空间数据的误差问题。

§1.2 GIS 数据误差研究的指导原则

人们逐渐认识到：GIS 产品的有效性和 GIS 本身的生命力与空间数据质量研究的成效是密切相关的。因此，可以预料，GIS 的数据误差问题会受到更多 GIS 工作者的重视。

1988 年 12 月，由美国地理信息和分析中心(NCGIA)主持召开的，由来自大学、研究所和 GIS 公司等部门的 30 多位 GIS 专家学者参加的专题讨论会，曾针对 GIS 数据误差研究问题提出了一些指导原则，现归纳并扩展如下：

(1) GIS 的数据处理精度几乎是无限的

如果不考虑计算机字长的表示精度，忽略原始数据的误差，则 GIS 处理过程中不会产生新的误差。因此，GIS 处理的精度可以认为是无限的。即使考虑计算机字长的表示精度(一般而言，单精度为 8 位数，双精度为 16 位数)，由于字长表示的局限所导致的精度损失和数据本身的误差相比很小，可忽略不计。故可以认为 GIS 数据处理的精度几乎是无限的。

(2) 所有空间数据的精度是有限的

空间物体的位置坐标都是通过测量与计算处理得到的。这些坐标的精度取决于所用的测量仪器类型、测量员的技术，及各种外界条件的影响，故其精度是有限的。此外，表示一个物体在地球表面上的位置有许多方法，假设的地球形状与真实地面的拟合程度，即参考椭球体的合适性也会影响物体空间位置的表示精度，这一问题将在第四章中详细讨论。更重要的是，表示空间物体所处的空间位置的数据往往只是这个物体外形的抽象综合。例如，标明为“A 类土壤”的地区并不说明该地区的土壤百分之百都是 A 类土壤，而有可能是 95%。用光滑曲面模拟人口密度分布并不能说明每个地区都严格地服从这个曲面模型。除了空间几何位置外，物体的属性也存在着不确定度。例如，卫星像片中某像元的属性误差很可能与该像元的位置误差不同。

(3) GIS 的处理精度超过数据本身的精度

从理论上而言，每个 GIS 数据都应附上一个精度指标。例如，用分划值单位为度的温度计来测量一个房间的温度，设结果为华氏 70 度，精度为 1 度(华氏)。如果把这个温度转换成相同精度(1

度)的摄氏温度,则GIS转换运算的结果为21℃。如果用数学运算,其转换结果为21.111…℃。GIS的处理是按最高名义精度来传递数据的,并按高于数据的实际精度输出结果。

(4) 在传统的地图分析、处理中,计算处理精度和地图数据精度是一致的

地图上地物的表达精度受到线宽最小分辨率的限制,一般定为0.5mm。由于湿度变化引起的纸张伸缩变形使实际精度大大低于这个指标,但在目前的数据采集和地图编辑中还是以0.5mm为目标精度。而传统的制图方法不能像GIS那样很容易地改变地图的比例尺,因此一般是在统一尺度下进行数据采集、编辑和综合处理,所以这些处理的精度都以图上最小分辨精度为基础,即0.5mm。

传统的地图处理和分析方法有透明格网纸叠置处理、求积仪量算面积和数点法等。这些方法比较简单,它们处理的精度都近似地与地图的表示精度一致。例如,图上多边形边界线的精度为0.5mm,而用求积仪所得面积的精度近似为1mm^2,两者的精度是相当的。故传统制图处理的精度估算比较容易。

(5) GIS处理精度与数据本身的精度不匹配

GIS有两个重要功能:改变比例尺和图形叠置分析。如果把不同比例尺或不同图形精度的数据叠置在一起生成新的地图,这个过程就会产生不匹配问题。例如,把1∶1千和1∶5千的地形图叠置在一起(因为GIS很容易进行比例尺缩放),然后按1∶1千的比例尺输出结果,则输出地形图的精度是多少?能否按1∶1千的常规地形图来定精度?目前还没有一种GIS能评估不同精度数据叠置处理后的精度,也没有一种GIS具有评价上述图形叠置处理的可行性和合理性的功能。大多数GIS的矢量操作运算(例如直线相交等)、叠置、缓冲区的生成均按坐标的最高名义精度进行,与地图的原有比例尺无关,结果造成GIS分析的结论与实际情况有很大出入。

(6) 没有适当的方法来描述复杂空间目标的精度

传统的测量方法可以很精确地确定地面点的三维坐标。但是,对于一个复杂的空间实体(包括线和面),用传统的方法来描述它的形状和各部分的变化是很困难的。目前还没有很好的方法来模拟抽象化了的空间物体之间的复杂关系及其空间变化。因此,需要研究一些能描述数据内在精度和跟踪GIS复杂处理过程中误差传播的方法。

(7) GIS信息误差研究的最终目的是建立GIS产品的合格证制度

通过研究建立的GIS空间数据误差理论来提供原始录用数据和操作处理过程中的各种误差信息,以及GIS产品的精度或可靠性的质量报告,使每个GIS产品附有质量指标。

在某些情况下,用置信区间来评价GIS产品的可靠性是可行的。这个置信区间依赖于原始数据中的误差和这些误差在GIS处理过程中的传播方式。例如,根据多边形的边界误差模型可以建立多边形面积估值的置信区间。但是,如果在GIS处理过程中需要考虑一些复杂的规则和限制条件,用这种方法评价GIS产品的质量就很困难。对于某些特殊问题,根本无法确定置信区间,只能用报警系统或防御系统来提醒用户注意。

上述7点是研究GIS空间数据误差理论的基本出发点。GIS处理不同类型数据的功能是很强的,因而受到人们的青睐。但是,来源不同的数据含有不同的误差。用GIS对精度、类型不一致的数据进行处理时,就产生了一个对GIS产品精度如何评价的问题。这些误差传播模型比常规的测量平差和精度评定要复杂得多。一方面,GIS的数据源复杂,一般是多种不同分辨率、不同精度和不同时间的源数据组合。此外,还存在着精确数据和非精确数据、定性数据和定量数据之间权的确定。另一方面,测量平差数据之间存在着严格的几何条件,它们的误差传播规律比较容易跟踪,而GIS系统中操作运算比较复杂,所研究的对象除了空间位置外还有属性,是一个抽象化的复杂物体。因此,

GIS 中的误差传播模型相当复杂。

§1.3 GIS 数据误差研究的发展概况与若干趋向

1.3.1 GIS 数据误差研究的发展概况

前已述及，GIS 数据误差研究的主要对象是 GIS 数据中的固有误差和操作处理中产生的误差，研究内容为这些误差的性质、度量和传播。固有误差的来源和度量依赖于数据采集的直接法（指从野外直接进行数据采集）或间接法（指从地图等图件上进行数据采集）。因此，这方面的研究历史可追溯到 GIS 建立之前的大地测量、工程测量和摄影测量以及制图学中的经典误差理论。在 GIS 空间操作运算产生的误差方面，1969 年，Frolov 建立了一个估计拓扑匹配误差的公式。1975 年，Switzer 提出了一种估计从矢量到栅格数据转换精度的方法。1978 年，Goodchild 给出了检验多边形叠置过程中产生的无意义多边形的统计量。1982 年，Chrisman 引入了著名的“ε—误差带”。1986 年，Burrough 对空间数据误差这一领域内的重要研究成果进行了总结。此外，Openshaw 也是从事该方面研究的著名学者。还应当特别提到的是，早在 1975 年，MacDougall 就用令人信服的例子说明了不考虑空间数据误差所带来的严重后果。

GIS 数据误差问题真正受到重视还是从 80 年代末开始的。如 §1.2 所述，1988 年 12 月由 NCGIA 主持召开的专题讨论会，其宗旨就是为 GIS 空间数据误差研究拟定方向和立题。这是 GIS 误差理论研究史上的一个里程碑，标志着人们对 GIS 误差问题进行系统研究的开始。

1990 年以前，GIS 数据误差研究的重点集中在误差的来源分析、空间和非空间误差度量指标的建立以及由数据变换处理函数所引入误差的模拟等。这一时期的特点是没有在 GIS 环境下将误差传播模拟的众多内容联系起来，甚至有些研究是独立于 GIS 环境之外进行的，这就是至今还没有能够进行误差处理分析的实用 GIS 的原因之一。但可以深信，随着 GIS 数据误差问题各项研究的深入，预计不远将来的 GIS 将具备这一功能。

1.3.2 GIS 数据误差研究的若干发展趋向

尽管 GIS 数据误差理论的研究内容繁多，但就目前来看，最有前途的发展方向可概括为下列 7 个：

（1）建立误差分析体系

这个体系包括误差源的确定、误差的鉴别和度量方法、误差传播模型的建立以及控制和削弱误差对 GIS 产品影响的方法。传统的概率统计仍是建立误差分析体系的理论基础。但是，必须根据 GIS 操作运算的特点对经典的概率统计理论进行扩展和补充。

（2）用敏感度分析法确定评价 GIS 产品质量的置信域

一般而言，精确确定 GIS 输入数据的实际误差非常困难。为了从理论上了解输出结果如何随输入数据误差的变化而变化，我们可以人为地在输入数据中加上扰动值来检验输出结果对这些扰动值的敏感程度。根据适合度分析，置信区间是衡量由输入数据误差引起输出结果变化的指标。目前应用得最广泛的两种适合度分析是加权叠置和加权多维尺度变换。为了确定置信域，需要对适合度分析进行地理敏感度测试，以便发现由输入数据的变化引起输出数据变化的程度，即敏感度。从这种研究中得到的并不是输出结果的真实误差，而是输出结果的变化范围。对于某些难以确定的误差，这种方法是行之有效的。在 GIS 中，敏感度检验一般有下面几种：地理敏感度、属性敏感度、面

积敏感度、多边形敏感度和增删图层敏感度。敏感度分析是一种间接测定GIS产品可靠性的方法。

(3) 尺度不变空间分析法

地理数据的分析结果应与采用的空间坐标系统无关,即尺度不变空间分析,它包括比例不变和平移不变。在集合分析和建模过程中,当把面元作为空间数据采集单元时,为了保证在改变面元集合方式的情况下不影响分析结果,需要满足尺度不变条件。此外,若把空间集合看成空间滤波器时,用尺度不变空间分析法就可以严格地测定空间集合的影响程度。尺度不变是数理统计中常用的一个准则:一方面能保证用不同方法得到的结果一致;另一方面又可在同一尺度下合理地衡量估值的精度。

(4) 空间集合与分区法

在GIS分析中,常常把小区域看成面元,而一个大区域又由若干面元组成。这在城市规划和社会经济分析中是常见的。这种面元可以是正规的方格形,也可以是不规则的三角形。每个面元的大小是空间精度的一个函数,由此引入了一个用于处理空间数据误差或不确定性的基本方法。由于将面元看成是建立GIS空间数据误差模型的随机抽样点。因此,需要首先划分研究区域,然后对每个子区或面元所包含的信息进行集合或综合抽象,而面元的大小和信息的综合方法又直接影响结果的精度。

(5) 空间数据误差的概念模式

我们可以把地理要素定义在空间(几何位置)、专题(属性)和时间三个维度中,每个维度的精度可由相应的误差大小来描述,例如,空间位置误差是由三维坐标精度来描述的,专题数据精度取决于数据的类型,它们常常与位置精度有关;在空间数据精度分析中常常被忽视的是时间精度,数据的可靠程度通常是时间的反函数,因为数据的空间属性和专题属性是随时间的变化而变化的。

空间数据误差的特点之一是多样性。数据质量包括6个主要部分:位置精度、属性精度、数据情况说明、逻辑一致性以及完整性和时间精度。位置精度和属性精度分别指精度的空间因素和专题因素。数据情况说明系指数据的来源、数据处理和编码方法以及对数据所进行的变换。逻辑一致性指数据编码关系的可靠性,包括拓扑、空间属性(例如同类多边形的边长和面积)以及专题属性的一致性。完整性是指描述数据库中目标以及目标的抽象概括之间的关系。总之,空间数据误差可以认为是由空间、专题和时间三个误差分量组成的。

(6) Monte Carlo实验仿真

GIS处理过程中的空间数据误差传播模型是很复杂的。由于GIS数据来源繁多,种类复杂,既有描述空间拓扑关系的几何数据,也有描述空间物体内涵的属性数据。对于属性数据的精度常常只能用打分或不确定度来表示。对于不同的用户,由于专业领域的限制和需要,数据可靠性的评价标准并不相同。因此,想用一个简单的、固定不变的统计模型描述GIS的误差传播规律似乎是不可能的。在对所研究问题的背景不十分了解的情况下,Monte Carlo模拟仿真是一种有效方法,它首先依据经验对数据误差的种类和分布模式进行假设,然后利用计算机进行模拟实验,将所得结果与实际结果进行比较,找出与实际结果最接近的模型。对于某些无法用数学表达式描述的过程,用这种方法既可得到实用公式,也可检验理论研究的正确性。

(7) 空间滤波

获取空间数据的方法可能是不同的,既可以采用连续方式采集,也可以采用离散方式。这些数据的采集过程又可以看成是随机采样,其中包含倾向性部分和随机性部分。前者代表所采集物体的形状信息,它可以是确定性参数,也可以是带有先验性质的信号;后者是由观测噪声引起的。

空间滤波分高通滤波和低通滤波。本书中,前者指从含有噪声的数据中分离提取噪声信息的过

程；而后者指从数据中提取信号的过程。经高通滤波后可得到一个点(或线、面)的随机噪声场，然后按随机过程理论或方差-协方差分量估计理论求得数据采集误差。

作者在建立地图数字化误差估计模型时，曾用到差分算子和样条函数把数字化过程中的倾向部分和随机噪声分离开。在没有粗差的情况下，这些去掉倾向的随机误差可以看作为相互独立的白噪声，利用最小二乘法可以求得数字化误差的方差和协方差。在有粗差存在的情况下，用于表示倾向的数学模型受到歪曲，倾向部分与随机部分无法有效地分开，因而随机部分的随机性和独立性不能得到保证。此外，即使这两部分能完全区分开，数据中也不允许存在粗差。为了防止粗差的干扰，可以利用抗差估计理论对数字化误差估计模型稳健化或者利用统计检验剔除粗差。

应当指出，以上 7 点在本书的各章节中均有反映。此外，由于 GIS 空间数据误差从不同角度提出了千差万别的研究课题。因此，除了传统的概率论、数理统计仍是研究该问题的理论基础外，还需要寻找信息论、地图传输论、模糊逻辑、人工智能，以及数学规划等基础理论来支持，而随机几何学和分形几何学也早有应用。

第二章　GIS 数据误差的来源、特性和分析

GIS 的主要功能之一是综合不同来源、不同分辨率和不同时间的数据，利用不同比例尺和数据模型进行操作分析。这种不同来源数据的综合和比例尺的改变使 GIS 数据误差问题变得极为复杂。为了对误差问题向用户提供一种解释，并寻找能有效抵抗和削弱误差影响的方法，需要比较清楚地了解 GIS 数据及其产品中所含误差的来源和特性等。本章主要讨论这一问题，内容涉及误差的来源、特性、模拟、分析和估计以及各种各样的处理方法。

本章共分 9 节。§2.1 概述有关 GIS 误差的来源、特性和分类等基本问题。§2.2 从数据误差角度考察地图叠置，其中叙述了地图叠置误差问题，并回顾了叠置操作运算中有关误差问题的重要研究成果。§2.3 通过综合土地统计部门的遥感图像、DEM 高程数据和来自地图的土壤信息，举例说明 GIS 中不同来源数据的复杂性，论证产生误差的原因以及与 GIS 的联系和用户的考虑。§2.4 介绍在 GIS 空间数据库中，由于缺少制图过程中使用的专家知识（例如地图图例、说明书等）所引起的误差和不可靠性，并指出通过建立专家系统可以探测和修正土壤图中的误差。

总之，前 4 节主要讨论 GIS 误差的基本问题、GIS 叠置分析操作中误差的类型和特性。在 GIS 的数据误差中，目前研究最多、最充分的是地图数字化过程中引入的误差，特别是点方式数字化过程中由数字化仪十字线的不精确放置引起的对点误差。本章后 5 节重点讨论这一问题。§2.5 概述采用传统的统计检验分析地图数字化误差的来源、性质、大小及相互影响的方法，着重介绍 Bolstad 等人的研究成果。§2.6 介绍通过一个精心设计的数字化试验方案来考察地图比例尺和数字化速度的变化对地图数字化误差影响的方法。此外，还简要介绍这种误差如何通过 GIS 操作进行传播，为下一章中全面讨论误差传播问题作准备。§2.7 介绍基于时间序列分析的地图曲线流方式跟踪数字化误差估计。在可以忽略随机性运动的情况下，得到了一组简明公式。§2.8 讨论数字化误差的抗差估计方法，以及在估计数字化误差时探测、发现和剔除粗差的方法。§2.9 讨论数字化误差统计模型中趋势项分离的非差分方法——样条函数，并论述曲线复杂性的度量指标。

§2.1　概　述

GIS 空间数据误差通常认为是数据与真值的偏离（概念或数量）。由于数据来源众多，因而产生误差的原因也很多，概括起来有以下三个方面：

①自然界的固有属性。变化和模糊是自然界的两个固有属性，它们影响着 GIS 信息的准确表达。例如，草原的范围并不总是确定的，而是向森林或沙漠区域逐渐移动；土壤单元的边界、植被类型的划分也常常是模糊的。不同的地理信息有不同的数据结构，地图或其他空间数据产品并不是现实世界的准确描述，空间特征分布的不明确造成了空间信息的不确定性。

②测量的固有属性。利用测量仪器设备进行的任何测量都不可避免地引入误差。这种误差主要受观测条件，即仪器、观测者和外界环境等三方面因素的影响。此外，对于外业调查数据，例如环境调查数据或地块区域单元界线的划分，则还取决于调查员的主观判断。

③测量结果的表达模型。表达地理信息的数据结构也是一个误差源。在矢量数据结构情况下，线或边界的表达具有较高的可靠性或精度，尽管在自然界中可能无法精确地辨别它们；在栅格数据

结构情况下，用像元表达集合的位置或均值精度较高。而通常，现实世界中的一些特征是抽象描述的。例如，为了使所感兴趣的特征在图上易读，需要进行概括和综合，这显然会引入误差和不确定性。此外，不但用于记录测量数据的介质材料本身会引入误差，而且在数据的传输过程中也会引入误差。例如，如果将卫星遥感图像数据转换为像片或透明胶片，则这种产品的误差特性将与原始数字形式的同种信息中所含有的误差不同。图 2.1.1 是 GIS 原始空间数据误差来源的一个略图，表达了可能的误差。

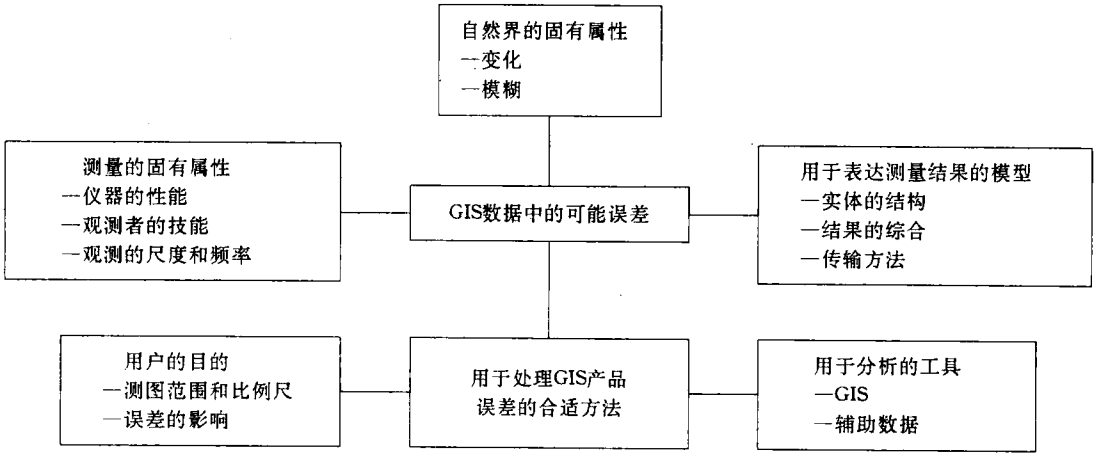

图2.1.1　误差源框图

除了 GIS 原始数据本身带有误差外，在空间数据库中进行的各种操作、转换和处理也将引入误差。由一组测量结果通过转换处理产生另一种产品时，通常转换的次数越多，则产品中引入的新误差和不确定性也越多。

一般将 GIS 数据误差分为两类：位置误差和属性误差(Chrisman，1987)。位置误差主要表达地理空间数据几何位置的精度，但不同数据类型的精度常常差别较大。例如，用现代测量技术确定的地形或地理数据常常是高精度的，而由测量员判断的土壤、植被单元等边界的位置精度常常较低，因为边界本身的确认是模糊的。属性误差用于表达地理数据库中点、线和面属性数据的正确与否，它又可进一步分为质量和数量两种，其中质量误差是指名义变量和标签是否正确，例如土地利用图上的某块“菜地”可能被编码为“豆地”；数量误差是指估计分配值时的偏差，例如不同 pH 值的几块地分配有相同的 pH 值，不同污染度的水域分配有相同污染度值。

根据误差对结果的影响性质可以将 GIS 数据误差分为随机误差、系统误差、粗差和变差。其中前三种是众所熟知的，而变差是由自然界随时间的变化或随空间的变迁引起的。例如，在陆地覆盖问题中森林边界随时间发生的变化引起变差，或在土壤、植被等土地景观自然特征的专题图上由于没有考虑局部变化源而引起变差，像表示为粘土的制图单元内可能有些部分为其他土类，表示为草地的制图单元内可能有小片的其他用地等等。

由于地理要素可以用空间、专题和时间三个广义坐标唯一表达，因而又可以将空间数据误差相应地定义为空间误差、专题误差和时间误差三种广义误差。而每一种误差又是多维的。例如，空间误差就是三维的(水平方向和竖直方向)。广义误差的重要性取决于所研究问题的侧重点。例如，在测量及其相关领域内，空间误差是最重要的，而在着重研究专题信息的领域(像土地覆盖、土壤和植被的分类)中，专题误差又是十分重要的。此外，在误差分析中，时间误差总是很重要的，但尚未引起重视。

如§1.3中所述，衡量空间数据质量的标准有6个，即数据情况说明、位置精度、属性精度、逻辑一致性、完整性和时间精度。

§2.2 分类图叠置中的误差问题

虽然GIS诞生的历史不长，但近年来却发展非常迅速。目前已推出了许多商品化GIS软件，并且在全世界范围内建立了许多地区性标准数据库。然而GIS中的分析功能却没有以同样的速度发展。在GIS技术中应用最多的至今仍是使传统的地图叠置自动化，而且GIS输出的图表产品和其他成果都没有附上与数据误差有关的精度或可靠性指标。

在支持现今GIS应用技术蓬勃发展的众多理论研究课题中，空间数据的误差问题几乎是最重要的(Chrisman，1989)。由于空间信息的表达形式多种多样，因此，需要平行地研究空间信息误差问题。本节首先叙述叠置问题和误差概念，然后回顾有关叠置分类图中误差问题的重要研究成果。

2.2.1 多边形叠置

由GIS存贮和管理的大量数据都是多边形形式，这是一种带有属性的二维封闭图形。GIS软件设计中的一个关键问题是如何表达这些属性的关系结构，寻找由Berry(1964)提出的"地理矩阵"。但GIS涉及到一些超出由程序数据(像SPSS，SAS或关系数据库)提供的标准属性管理操作，这些特殊操作需要空间结构和明确的空间误差公式。

许多学者研究过在空间分析中使用的各种各样的操作(Chrisman，1982a；Tomlin，1983；Guevarra，1983)。每一项研究都发现，由地图叠置所揭示的位置重合对于设计优秀算法是很重要的。叠置中使用位置信息构成新区，它共享了各独立数据层的"血缘"关系。地图叠置作为一种直观的非定量方法曾使用过多年(Sauer，1918；Steinitz等，1976)。从根本上看，叠置就是对由Warntz(1968)描述的地理问题使用Venn图(集合逻辑)。70年代末研究的重点是估计地图叠置的有效性，而80年代末已从不精确的手工处理转向可靠的自动化处理，并且有许多商品化的软件程序能够完成叠置操作。

然而，这种表面上的迅速发展是不健全的，因为现有的软件系统还没有一个能削弱空间数据误差对操作结果的影响。虽然早就有介绍叠置分析误差与错误的警告性文章(MacDougall，1975；Hopkins，1977)，但没有研究出现问题的原因。为了保证正确地应用GIS，必须研究叠置图中的误差，并建立有关的误差模型。

2.2.2 误差概念与分类层定义

为了改善GIS的空间分析算法，需要建立通用的空间数据误差模型，以便指导算法的改进方向。例如，地理数据统计分析的改进依赖于在其他学科中所建立的误差模型。但是，对叠置分析问题，单纯地靠借用策略是无效的。

长期从事传统人工制图的制图员很难理解并接受误差的统计概念。在他们看来，误差是不容许的，必须完全消除掉，这一感性认识与误差的统计概念是不同的。他们认为的误差与摄影测量学中的粗差在概念上非常相近，但制图线中的误差并非如此，它们还包括一类更细微的随机误差，这类误差与统计分类处理中的误差相似。各种各样来源的误差混入地图中，每种误差所产生的影响不同。尽管可以通过较高费用的技术来削弱这些误差，但这种高花费是不值得的。因为在了解误差的来源和特性后，比如随机误差，可以用概率统计理论来处理。许多严密科学的分析不是基于无误差

的确定性模型，而是随机模型。

在讨论误差模型时，涉及的空间信息是各种各样的。为研究方便，以下只考虑处理分类层的一类二维分布情况。

在GIS应用中，分类层是使用相当多的一种特殊类型的多边形图。空间采集单元普查信息的有效性已促进了空间统计学的发展。特别是，通常称为“可修改面元”的统计相似域的处理推动了空间自相关的发展。空间自相关是一种强有力的工具，它提供了一个有用的开端，并基于随机配置单元(图 2.2.1)。

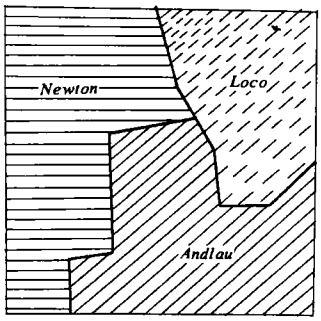

图 2.2.1 随机配值单元(一类分类图)

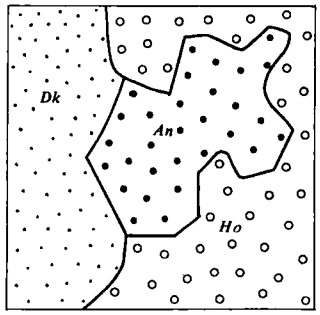

图 2.2.2 分类层(与随机配置不同的另一类分类图)

空间图形或属性，何者取逻辑优先权是一种重要考虑。在纯理论情况下(像城市市区这样的行政单元)，目标的位置图形优先于任何属性值，这些图形从纯理论角度看就是等值线图。由于配置区域是随机的，因此，重要的是发展能够消除由区域边界的偶然变动所引起偏差的统计方法，像空间自相关。这种方法假定一些基本的空间分布通过使用配置区而变得模糊不清。方法的关键是权矩阵，它通常基于配置单元间的一些拓扑关系或模糊关系获得。

空间分析的实质是应用随机配置区，尤其在社会地理学和经济地理学领域中。分析的单元常是个人、家庭或公司，但空间集合统计仅是可利用的资料。对其他形式的数据，需要另外的空间模型。

GIS软件的许多用户不能重新更换配置单元资料。输入到GIS中的层多半是土壤图、植被图和主权领土等等。尽管差异不是绝对的，且来源于不同途经的这些图(图 2.2.2)可以复制成等值线图，但图形显示的相似性使基本的差别模糊不清。一些分类系统(土壤分类学、植被分类学、可征税区域的列表)比地图逻辑优先。地图来自将面积的每一部分分配给一类或另一类，位置精度、比例尺和其他制图关系问题比它们在配置区域情况下变得更为显著。空间自相关中的误差模型基于连续分布模式集合而成离散点和空间单元。分类层误差模型则与之相反。分类层的误差模型必须考虑制图生产过程的制约。

正如社会经济调查的分类数据分析需要采用与标准回归模型不同的误差模型一样，分类层也需要采用与连续空间分布不同的误差模型。

2.2.3 空间数据误差模型的回顾

在空间信息处理的不同学科中，已建立了大量不同的空间数据误差模型。按研究问题的方式不同，有归纳和演绎两种建立误差模型的方法。基于随机模型的演绎法比基于确定性模型的归纳法适用范围广，但目前采用的大多数模型都不能直接用于分类层问题，新的模型可以通过使不同的方法

和误差及数字模型相一致来建立。

对大地测量学、工程测量学和摄影测量学中应用的不同定位系统，已经建立了许多模型。这些相关学科依赖于一个用相当大的联立方程组表示的多余观测模型。最小二乘平差可以计算每个点的误差椭圆。在大地测量的应用中，这种模型是必要的。在另一些应用中，像地图边界角的位置，也可以使用这种指标。但在制图分析中，每个点的误差椭圆不是最终产品，因为许多分析应用中的结果是面积。因此，需要建立面积估值的方差模型。Neumyvakin 和 Panfilovch(1982)对由某些特征点描述的多边形给出了面积的方差估算公式。Chrisman 和 Yandell(1988)补充了遗漏的部分，并改正了数学缺陷，但结果没有在实际中应用。

测量模型的数据压缩量是很大的，但信息模型，尤其是对于大多数的分类层却不能压缩。测量误差模型将误差归结为点，而分类图却不是一系列点图。如果误差仅仅是点的函数，则一条线的位置精度至少处于两个端点之间。从数学上看，这种结果是正确的，但它不符合制图情况。因为制图过程中所选择的碎部测点必须比点间假定的直线更为可靠。

与常规测量一样，遥感也重视制图误差，但把重点放在分类问题上。至今关于属性精度问题的研究成果还很少。尽管遥感的产品通常是分类层，但常规的统计处理方法一般仍将它们作为传统的样本处理。

在空间误差处理方面，最富有理论特色的工作是基于数理统计著作《随机几何学》(Harding 和 Kendall，1974)进行的。《随机几何学》的应用范围很广，除了地理学外，还出现在定量生态学和地质学中。尽管点图最易建立模型(Rodgers，1974)，但处理面积问题模型的可行性受两种情况限制(Getis 和 Boots，1978)：由随机点建立的或由无数随机线建立的 Thiessen 多边形。对于结点区域的一些情况，已建立了第一种模型，但很明显它不能解决分类层的均匀区域问题。随机几何学的主要任务是建立带有完全随机源的空间分布模型。

另一个数学概念——"分形"已在理论地理学和其他学科中推广(Mandelbrot，1977)。从 Mandelbrot 开始现已扩展到 Lucas Films。为了形成分数形式的随机表面，已经做了大量实验。所模拟的景观特性，逼真结果相当好，说明分形模型是成功的。但是，分形仅仅提供了一种结构特性的度量，一种局部变化和整体变化之间的相互关系。

Poiker(Peucker 以前，1976)提出了信息量的一种特殊制图理论。这种基于道格拉斯线的简化算法——"线的宽度分解"对许多软件系统是重要的。制图学中的一些研究(例如，McMaster，1986)证实 Poiker 的理论是探测重要细部的的一种方法。但是，Poiker 的制图线理论描述的是点的几何特征，并非位置误差。

Chrisman(1982a；1982b)利用"ε—距离"术语提出了一种描述多种制图误差的方法。"ε—误差模型"用于表达线的精度，每线条的观测值与线的真值间存在着差异。从几何学角度看，线加粗后形状变为柱状，它反映了线真值的概率密度函数。若把柱的半径 ε 适当调整，使其与表达线精度的标准差相等，则此时的 ε 就表示了区域的平均误差。Goodchild 和 Dubuc(1987)针对清晰边界线评价了空间误差的 ε—误差模型。ε—带用来描述线的平均可能位置范围，以及度量面积观测中的可能误差，但不能用于定义 GIS 应用中"缓冲区"通道明确意义下的误差。ε—误差模型不仅仅是几何上的，ε—区还可以概括成制图分类中的一个方阵。

Goodchild 和 Dubuc(1987)对随机分类覆盖层(或称自然资源图)在假定存在一个连续相位空间且其中的分类是可分辨的情况下提出了一个新模型。这种模型不能提供显示正确拓扑结构的随机分类覆盖层，因为并非所有的分类都是从这种连续的相位空间中派生出来的。许多地物(河流、湖泊、冰川、城市等等)都有边界，因此该模型只能在一定范围内使用。