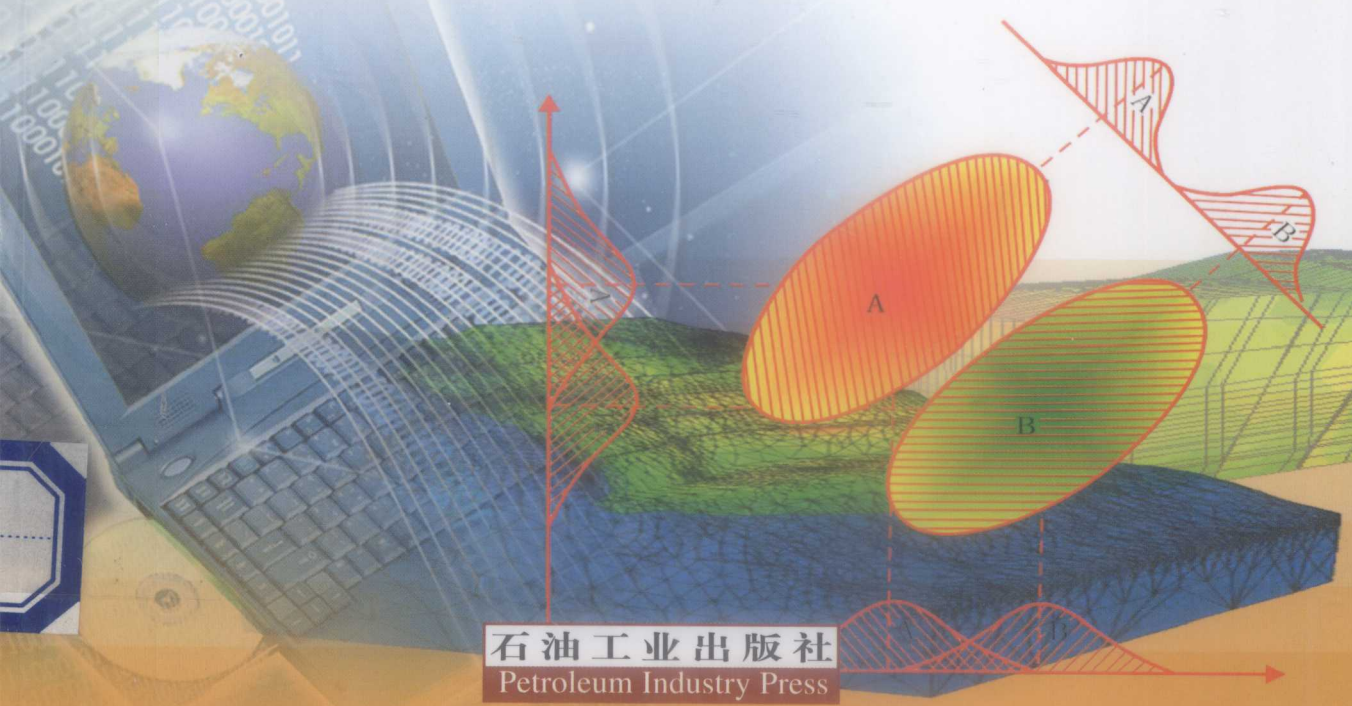
普通高等教育“十一五”国家级规划教材

高等院校石油天然气类规划教材

数学地质方法及应用

刘绍平　汤　军　许晓宏　主编



石油工业出版社
Petroleum Industry Press

普通高等教育“十一五”国家级规划教材
高等院校石油天然气类规划教材

The Methods and Applications on Mathematical Geology

数学地质方法及应用

刘绍平　汤　军　许晓宏　主编

石油工业出版社

内 容 提 要

本教材是针对地质勘探与开发人才的培养而设计的，内容以多元统计为主，涵盖了单变量分析和多变量分析，还包括油气资源定量评价和数学模型等。书中由易到难对多元统计分析和数学模型的理论、方法在油气地质及矿产地质中实际应用等作了较全面的讲解，构成一个地质勘探领域从基础到应用的数学地质方法体系。各章有实例和习题，以帮助读者领会各章知识要点，掌握基本内容。

本书可作为高等院校资源勘查工程、地质工程等相关专业本科生和研究生的专业课教材，也可作为数学地质工作者和相关研究人员的参考用书。

图书在版编目（CIP）数据

数学地质方法及应用/刘绍平，汤军，许晓宏主编．
北京：石油工业出版社，2011.4
（普通高等教育“十一五”国家级规划教材·高等院校石油天然气类规划教材）
ISBN 978-7-5021-8361-5

Ⅰ.数…
Ⅱ.①刘…②汤…③许…
Ⅲ.数学地质-高等学校-教材
Ⅳ.P628

中国版本图书馆 CIP 数据核字（2011）第 052220 号

出版发行：石油工业出版社
（北京安定门外安华里 2 区 1 号　100011）
网　址：www.petropub.com.cn
编辑部：（010）64523612　发行部：（010）64523620
经　销：全国新华书店
印　刷：北京中石油报印刷厂

2011 年 4 月第 1 版　2011 年 4 月第 1 次印刷
787×1092 毫米　开本：1/16　印张：13.75
字数：350 千字

定价：21.00 元
（如出现印装质量问题，我社发行部负责调换）

前　言

为适应石油工业和石油高等教育发展的形势，中国石油教育学会和石油工业出版社于2004年10月在北京怀柔召开了石油地质与勘探专业教学与教材规划研讨会。与会的石油高校有关领导和教师就当前石油高等教育和教材问题进行了研讨，并制定了一系列教材编写计划。本教材是其中的一本，主要是针对资源勘查工程、地质工程等本科专业的教学要求而编写的，约40学时。2006年8月，本教材被教育部列为普通高等教育“十一五”国家级规划教材。

根据专业教学的特点，本教材主要介绍多元统计分析的基本概念、特点、计算方法和模型及其在地质勘探中的应用，重点是单变量分析、多变量分析及其数学模型的实际应用方法，同时简要介绍了20世纪60年代以来的数学地质进展。教材的编写以主编教师的授课讲义为基础，并参考了国内外代表性的数学地质教材和最新国内外有关数学地质的研究成果。书中由易到难地对多元统计分析、油气资源定量评价和数学模型的理论、方法在油气地质及矿产地质中的实际应用等作了较系统的讲解。

本教材的编写工作由长江大学刘绍平、汤军和许晓宏共同完成。其中第一章、第二章由刘绍平编写，第三章、第四章和第七章由汤军编写，第五章、第六章由许晓宏编写。全书由刘绍平负责统稿。

在教材的编写过程中，长江大学教务处和相关院系领导给予了大力支持和帮助，云南大学王学仁教授、长江大学田时芸教授对本教材的编写提纲和部分章节提出了有益的建议，研究生周晓阳、何建红、李哲、杨争光、吴雪超、安丹和周静平等同学参与了文字、附表的录入和绘图工作，在此一起表示衷心感谢！

由于编者水平有限，书中不足之处敬请读者批评指正。

编　者

2010年12月

目　录

第一章 绪 论

[本章学习目的和要求] 了解数学地质的基本概念、性质、背景知识，数学地质研究的内容和方法，以及数学地质学科发展的历史。

第一节 数学地质的产生

一、数学地质的概念

数学地质是由地质学、数学和计算机科学相结合而形成的一门边缘学科。它是运用数学的理论和方法研究地质学基础理论，解决地质学实际问题的一门科学。计算机技术是数学地质研究的主要技术手段，为定量地研究地质学问题提供了计算工具。数学地质的产生把地质学从定性的描述逐步提高到向定量研究方向发展，对地质学的定量化进程起到了重要的推动作用。

二、数学地质的性质

数学地质是一门新兴的地质学分支学科，对其性质存在着两种不同的观点。一种是广义的观点，以加拿大的学者阿格特伯格(F. P. Agterberg，1974)为代表，认为数学地质应包括地球科学中的全部数学应用。持类似观点的还有戴维斯(J. C. Davis，1973)的地质数据的定量分析方法。另一种是狭义的观点。前苏联学者维斯捷列乌斯(А. Ь. Вистелиус，1977)提出了不同的看法，他认为“数学地质是建立、检验和解释地质过程的概念、随机模型的科学”。对于这一观点，我国学者刘承祚(1981)指出：“虽然随机类数学在地质学中占据着主导地位，但不是唯一地位……在地质学中还应用着大量的确定性数学方法。”

赵鹏大教授等(1983)提出：“数学地质是研究地质运动数量规律性的科学”，对正确认识数学地质学科的性质有重要意义。

三、数学地质产生的背景

地质学是一门以地壳为主要研究对象，有着悠久历史的自然科学。传统的地质学主要靠记录和描述性方法得出地质学的规律和结论，在地质学的研究中很少采用数学的方法，基本上是一门定性的科学。由于现代科学技术的发展，许多新理论、新方法和新技术不断地向地质学中渗入，使地质学在发展的过程中不断和物理、化学、力学、数学相结合而产生了相应的分支学科。如地质学与物理学相结合，产生了地球物理学和地球物理探矿法(电、磁、重、地震)；地质学与化学相结合，产生了地球化学和地球化学探矿法；地质学与力学相结合，产生了地质力学和地质力学探矿法；地质学与数学相结合，产生了数学地质，使地质学的研究逐步定量化。

产生数学地质的背景主要有以下四个方面：

(1)测试仪器的不断改进和更新，地质数据的大量出现，使得地质研究人员无法采用定性的方法来处理、分析和利用这些定量的地质资料，故地质学研究中必须引入定量的方法。

(2)矿产资源的需求量越来越大,矿产的勘探开发由地表或浅层逐步扩展到地壳深部和海洋。由于地质工作的难度加大,为了有效地利用地质信息,也应采用数学的方法。

(3)计算机技术的发展,提供了利用数学方法解决地质问题的计算工具,推动了数学地质的产生和发展。

(4)地质学深入发展的需要。地质学如同其他学科一样,只有成功地应用数学,才能真正达到完善的程度。

第二节　数学地质的内容和方法

一、地质多元统计分析

地质多元统计分析是运用概率统计的思想,研究解决地质学中多指标问题的理论和方法。地质多元统计分析为定量处理地质数据提供了理论和方法,是数学地质最成熟的内容,也是应用最广、效果最为明显的数学地质方法。

地质多元统计分析方法可以简化地质数据结构、分析地质变量的相依性、组合地质变量、对样品或变量进行分类,常用的方法有回归分析、趋势面分析、聚类分析、判别分析、因子分析等。对特殊类型的地质数据,如成分数据、定性数据、定向数据,通过数据变换后,也可使用地质多元统计方法进行计算处理。有关常用多元统计方法的原理、计算方法步骤以及在地质学中的应用将在后面章节中作详细的介绍。

二、地质统计学

地质统计学自20世纪60年代初创立以来,在地质学研究领域中得到了广泛应用,已成为数学地质的重要分支学科。地质统计学是以区域化变量理论为基础、以变异函数为主要工具的一种数学地质新理论和新方法。地质统计学最初用来研究矿石品位的空间结构、矿床储量计算和误差估计。地质统计学在油气勘探中的储集层建模、油藏描述、油气资源量预测等领域也得到了应用。地质统计学除成功地应用于地质学领域之外,还不断地应用于环境科学、农林科学、水文及工程科学、海洋科学等领域。

地质统计学在应用的过程中,不断地产生一些新的地质统计学方法,如指示克立格法、协同克立格法和泛克立格法等。

三、地质数据库

数据库是20世纪60年代末出现的数据管理技术,比较完善的数据库软件系统是70年代初建立的。地质数据库在美国、加拿大、德国等西方国家发展很快,20世纪80年代后在许多国家已普遍应用。比较完善的地质数据库有:

(1)美国矿产资源信息库(CRIB)。该信息库是美国地质调查所建立的矿产资源数据库。数据库内存储了美国4万多个矿床和矿产产地以及其他国家6千多个矿床和矿产产地的资料。数据文件内容包括矿床位置、地质特征、储量、产量等多种数据。

(2)北美石油数据系统(PDS)。该数据库存储了美国和加拿大的10万余个油气田的有关资料。数据文件内容包括油气田的产量、生产井数、储量、圈闭类型、储集层厚度、油气性质、地层温度、地层压力、岩性等多种数据。

我国已建立了不同级别的各种类型地质数据库，如物理勘探资料数据库、油气化学勘探数据库、井下地质数据库等。

四、地质过程的数学模拟

地质系统在漫长的地质历史时期中经历了复杂的地质变化过程。地质过程模拟就是应用数学的方法、利用计算机模拟再现地质作用的过程。模拟的方法分为确定型模型和随机型模型。确定型模型常用的有微分方程、积分方程和代数方程等。如用常微分方程模拟生油岩干酪根热降解的成烃模型，用积分方程模拟泥岩孔隙度变化的函数关系、恢复地层的古厚度，盆地模拟中使用的超压方程和热流方程均为三维偏微分方程。随机型模型是指应用概率统计、随机过程等作为数学方法的模拟。如应用马尔可夫链分析模拟地层的沉积层序，在石油资源定量评价中应用蒙特卡罗法模拟石油资源量等。

任何一个地质过程都不可能是单一的确定型过程或随机型过程，往往有可能是两种地质过程在时间和空间上的叠加。因此，在实际地质过程模拟时，应能综合地利用两种模型的模拟结果，逼近地质事件的历史演化过程。

五、矿产资源定量预测和评价

随着社会对矿产资源需求量的不断增加和迅速开采，找矿的难度和风险也越来越大。从20世纪70年代末以来，矿产资源的定量预测和评价已成为数学地质研究的重要内容。利用数学地质方法对油气资源进行定量预测和评价，其任务是要计算探区的油气资源量、预测有利油气聚集带、减少勘探风险、提高探井的成功率、对探区的油气资源远景作出综合评价。常用的油气资源定量评价方法有蒙特卡罗模拟法、油田规模序列法、胡伯特(Hubbert)模型法、指数增长模型法。

第三节　数学地质的发展简史

数学地质的思想来源很早，开始于18世纪中叶，直到20世纪60年代才逐步形成一门独立的地质学边缘学科，其发展历史可分为三个大的阶段。

一、创立阶段(1840—1945年)

这一阶段，在地质学中初步尝试应用数学在个别方面进行少量分散的研究，数学方法的应用逐步扩展到一些地质学的分支。1840年，英国地质学家莱伊尔(Lyell)通过对古生物化石的统计分析，对古近系、新近系地层进行了划分，确定了古近系、新近系岩石地层的层序。1914—1934年，俄国人列文森—列星格(Левинсон-Лессинг)通过对岩石的岩浆系数的频率分布研究解决了岩浆岩的分类问题。1939年，西姆波森(G. G. Simpson)等编写了《定量动物学》一书，为古生物统计学发展奠定了基础。

20世纪40年代以后，开始引入双变量的相关分析，用于研究两个地质变量之间的关系，如岩石中某种矿物的性质与其化学成分之间的关系等。1944年，维斯捷列乌斯在苏联科学院报告集上发表了《分析地质学》一文，提出了用定量方法研究地质问题的初步思想。他本人也从事了30多年的数学地质工作，成为苏联数学地质学科的创始人和国际数学地质协会第一任主席。

二、形成阶段(1946—1970 年)

这一阶段,单变量、双变量统计方法和计算机技术在地质学中得到广泛的应用,地质统计学、地质数据库、地质过程的定量模拟开始得到应用,数学地质成为独立的地质学分支学科。1954 年,绍(D. M. Show)等人应用统计的方法研究地球化学等问题。1958 年,克鲁拜因(W. C. Krumbein)首次在地质杂志上公布电子计算机程序,成为美国数学地质的奠基人。1962 年,法国学者马特隆(G. Matheron)在南非工程师克立格(D. G. Krige)等人工作的基础上,创立了地质统计学,并在金属矿床储量评价中得到应用。从 1961 年开始,美国的亚利桑那大学召开了一系列电子计算机在矿产工业中应用的讨论会。1967 年,在美国石油地质学家协会上建立了电子计算机数据储存和索取委员会,成立了国际地质科学联合会地质数据储存、自动处理和索取委员会(COGEODATA)。1968 年,在布拉格召开的国际地质会议上成立了国际数学地质协会(IAMG)。在这期间出版了各种专门的学术刊物和专著。1969 年,国际数学地质协会创办了国际性《数学地质》(Mathematical Geology)期刊,之后又出版了《计算机与地球科学》(Computers & Geosciences)、《地质计算程序公报》等刊物。另外出版的专著有《电子计算机在地层分析中的应用》(J. W. 哈博等,1968)。由于计算机技术的发展,地质过程的计算机模拟开始得到应用,数学地质发展成为地质学的分支学科。

三、提高阶段(1971 年至今)

这一阶段,多元统计分析方法应用于地质学的各个研究领域,地质统计学、地质过程的计算机模拟得到广泛应用,建立了大量的地质数据库,新的数学理论和方法不断地与地质学相结合,扩大了数学地质的研究领域和研究方向。

多元统计方法是最早引入到地质学中的数学方法,也是数学地质最为成熟的内容,在沉积岩的分类、地层的划分与对比、沉积相和沉积环境的划分与判别、油气构造的识别、油源对比、油气水层的判别、有利油气区块评价、油气资源预测与评价等方面都有应用。

地质统计学得到迅速发展,不仅用于矿石品位和矿床储量计算,而且应用于油气勘探中的沉积砂体分布、储集层建模、油气资源量预测等领域。

利用计算机技术开展对地质过程的模拟,建立各种地质过程的数学模型,如构造发育模型、地层沉积模型、石油生成模型等,通过这些模型来模仿再现地质过程的发生、发展及演变过程。

各类地质勘探数据库都相继建立,包括美国地质调查所建立的矿产资源数据库(CRIB)、北美石油数据系统(PDS),我国建立的石油勘探数据库等。计算机绘图技术也得到了普遍应用,极大地提高了地质研究的水平和工作效率。在这期间,新的数学理论和方法不断与地质学相结合,如模糊数学、灰色理论、分形理论、人工神经网络、地质专家系统、地理信息系统、数字地球等都在地质学中得到应用,扩大了数学地质研究领域和研究方向,提高了数学地质研究的水平,促进了数学地质学科的发展。

第二章　地质数据的表达方式及其特征

[本章学习目的和要求] 了解地质数据、地质变量的基本概念、特征及其类型，掌握地质变量线性变换的一般性方法，掌握地质数据的统计分布特征概念及其意义。

第一节　地质数据简介

一、地质数据的概念

地质数据是表示地质信息的数、字母和符号的集合。地质数据是用来表示地质客观事实这一地质信息的。从广义上讲，地质数据可以是定量的、定性的数据，也可以是文字的说明，甚至是图形的显示，它几乎等同于原始的地质观测结果或地质资料；从狭义的角度来看，地质数据主要是指定量的和定性的数据。

二、地质数据的类型

地质数据按其特点可以分为观测数据、综合数据和经验数据三大类。

(一)观测数据

观测数据是指利用各种观测手段对地质研究对象进行观测或度量所获得的数据，是地质数据的主要类型。这类数据一般没有经过任何加工和处理，所以也称为原始数据。观测数据根据其特点可分为定量数据和定性数据。

1. 定量数据

定量数据是指能用数值大小来表示的观测数据，包括间隔型数据和比例型数据两类。

1)间隔型数据

间隔型数据有明确的数量概念，可以用数值形式表示。这类数据的特点是无自然零值，但有负值存在，不仅能比较大小，而且可以定量地表示数据间的差异，即任何两个间隔型数据之差是有意义的，如两个地层分层数据之差就表示某段地层的厚度，海平面的高程值也是典型的间隔型数据。

2)比例型数据

比例型数据有明确的数量概念，可以用数值形式来表示。比例型数据之间不仅其差值有实际意义，而且比值也有实际意义。比例型数据与间隔型数据的另一个区别是比例型数据中存在绝对零值而没有负值。大多数定量数据，如储量、产量、有机碳含量等，都属于比例型数据。这类数据所反映的数据概念是完整的，意义最明确，因而也是最重要的一类定量数据。

2. 定性数据

定性数据是指不能用数值表示而只能用符号或代码表示的观测数据。这类数据不具备数

量上的概念，包括名义型数据和有序型数据两类。

1)名义型数据

名义型数据没有明确的数量概念，数据之间也没有次序关系，只能用符号或代码形式表示。名义型数据是通过区分不同的对象或个体并赋予不同的代码后形成的。如描述岩石颜色的红、绿、灰、黑色，可用符号 A,B,C,D 来表示；又如砂岩、泥岩、石灰岩等岩石类型可用符号 S,M,L 来表示。符号 A,B,C,D 和 S,M,L 就是名义型数据。名义型数据之间只存在“相等”或“不相等”的关系，如红色等于红色(A=A)，砂岩不等于石灰岩(S≠L)。

2)有序型数据

有序型数据没有明确的数量概念，但数据之间有次序关系，常用等级符号或代码来表示。如鉴定岩石相对硬度的摩氏标准，将岩石硬度由小到大分为十个级别，即滑石、石膏、方解石、萤石、磷灰石、长石、石英、黄玉、刚玉、金刚石，分别用符号 1,2,3,4,5,6,7,8,9,10 表示；又如生油岩干酪根的范氏分类将干酪根分为Ⅰ，Ⅱ，Ⅲ三种类型。有序型数据除有相等、不相等关系外，还有“大于”或“小于”的关系。

(二)综合数据

综合数据是由观测数据经有限次算术运算后得到的具有明确地质意义的数据，如甲烷系数、奇偶优势指数、时间—温度指数(TTI)等。另外，随机变量的各种数值特征，如平均值、标准差、变异系数、相关系数等，都可认为是综合数据。

(三)经验数据

经验数据是指在研究地质现象的地质过程变化规律后，经过归纳或根据经验公式计算而得到的数据。经验数据通常是大量地质信息的综合反映，其地质意义往往是十分明确的，但具体的地质影响因素及它们之间的相互关系却是不确定或不清楚的。石油资源定量评价工作中经常使用经验数据，如单储系数、排烃系数、聚集系数等。

由于每个地质研究人员工作经历的局限性，经验数据往往具有较明显的地域性特征。因此，在使用经验数据时，要特别注意对比地质条件的相似性，不加选择地引用往往会导致错误的结论。

三、地质数据的特点

由于地质系统、地质条件和地质作用复杂多变，各种技术测试手段之间存在着较大差异等原因，造成了地质数据本身的许多特点，概括起来有以下几个方面：

(1)地质数据的类型多，性质不一，反映的地质内容十分广泛。

(2)地质数据的数量和精度相差大，数据的量纲、量级不统一。

(3)地质数据往往反映了多种地质因素综合作用的结果，具有混合分布的特征。

(4)定量数据是地质数据的主要类型，对地质定性数据的定量化研究和应用尚不够成熟。

上述特点说明，地质数据不是单一性质的集合，而是具有多种来源的复杂数据集合。这些特点是客观存在和不易改变的，使用数据时要特别注意适用性，对不同的使用目的要选用不同的数据，同时还要科学的对数据进行选择和处理，只有这样才能有效地使用地质数据，使得到的计算结果具有明确的地质结论。

四、地质数据的误差

任何观测手段都不可能得到与实际地质情况完全吻合的观测值。这是因为在野外观测、

样品采集、分析化验、仪器读数和数据整理过程中，由于工作人员的主观因素、测量和分析仪器精度的限制、周围环境和人为过失的影响，必然会对观测数据产生误差。误差是衡量数据质量的重要标志，按其性质可分为三种类型。

(一)随机误差

随机误差是在数据的观测或测量过程中由不可控制的、无规律的偶然因素而引起的误差，一般近似服从均值为0的正态分布。这类误差的大小和正负各不相同，不能人为地加以控制，当观测次数增加时，误差的均值趋于0。随机误差往往导致观测数据在一定范围内出现波动，称为观测数据的波动性或统计涨落性。

(二)系统误差

系统误差是由于观测系统本身所引起的误差。如仪器不准确、测量方法不合理、测量条件或环境不同、观测者习惯不同等因素引起的误差都属于系统误差。系统误差往往使观测数据整体上偏大或偏小，可以采用一定的方法校正观测数据，降低或消除这类误差。

(三)过失误差

过失误差是指在数据观测和数据整理过程中，受到各种干扰和人为过失等因素影响所产生的误差。这类误差使地质数据失去了真实性和代表性，称为数据失真。如样品的污染、分析仪器的瞬时故障，观测系统的测错、传错、记错等，都可能使观测数据失真。失真的数据是难以校正的，对数据的计算结果会产生严重的影响。在实际工作中，要尽量避免出现数据的过失误差。

五、地质数据的选择

由于地质勘探的程度不同，不同研究区域的地质数据在数量和质量上会存在着较大差异。因此，必须根据研究的对象和目的，对地质数据进行收集和选择。选择地质数据时，应注意以下问题：

(1)应根据研究目的和要求选择地质数据。

(2)各类地质数据反映的地质意义要明确，可靠性要强。

(3)地质数据的量纲要一致，量级不能差异太大。

(4)地质数据在平面上的分布要合理，应尽可能保持数据的均匀性。

(5)地质数据的数量应能满足数学模型的要求，能反映数据的统计规律性。

六、原始数据矩阵

地质数据的数量比较大，为便于对地质数据进行处理，可将地质数据用数据矩阵表示。通常矩阵的每一列是一个变量的多个观测值，而每一行则是包含多个变量观测值的一个样品。如果一组地质数据包含 p 个变量的 n 次观测值(n 个样品)，则可用下列 n 行 p 列的原始数据矩阵 $\boldsymbol{X}$ 表示：

$$\boldsymbol{X}=[x_{ij}]_{n\times p}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

数据矩阵 $\boldsymbol{X}$ 中，包括了对 p 个变量的 n 次观测值，x_{ij} 表示第 i 个样品的第 j 个变量的观

测值。

【例 2-1】 某探区已发现 5 个地质圈闭，为了描述这些圈闭的地质特征，选用了圈闭面积、闭合高度、长短轴比、埋藏深度共 4 个地质变量，数据如表 2-1 所示。请将这些数据表示为原始数据矩阵。

表 2-1 地质圈闭数据表

地质变量 / 圈闭编号	圈闭面积 10^2m^2	闭合高度 m	长短轴比	埋藏深度 m
1	1000	500	1.5	2000
2	250	150	1.0	2200
3	100	70	3.0	1500
4	10	200	2.0	1800
5	40	100	5.0	2500

将表 2-1 中的数据整理为以下 5 行 4 列的原始数据矩阵：

$$\boldsymbol{X}=[x_{ij}]_{5\times4}=\begin{bmatrix}1000 & 500 & 1.5 & 2000\\ 250 & 150 & 1.0 & 2200\\ 100 & 70 & 3.0 & 1500\\ 10 & 200 & 2.0 & 1800\\ 40 & 100 & 5.0 & 2500\end{bmatrix} \tag{2-1}$$

第二节 地质数据的预处理

地质数据的类型多，时空分布不均匀，且存在着数据失真，所以直接采用观测数据进行计算是不合适的。因此，在正式计算之前要对观测数据进行预处理。这是地质数据定量计算过程中不可缺少的一个重要环节。

一、可疑数据的鉴别与处理

地质数据失真会导致平均值过高，不能反映数据的总体特征。失真的地质数据也称为可疑数据或外来值。要查明造成数据失真的原因是不容易的，在实际工作中不能随意地将可疑数据舍去或保留，要用统计学的方法对可疑数据进行鉴别和处理。常用的检验方法有以下两种。

(一)肖维纳(Chauvent)检验法

肖维纳检验法的具体计算步骤为：

(1)计算观测值的平均值 $\bar{x}$ 。

(2)计算单次观测的概率误差 p：

$$p=0.6745\sigma$$

式中 σ——观测值的标准差。

(3)计算可疑数据与平均值之差 D 和比值 D/p。

(4)根据表 2-2 中的观测次数 n 与其对应的 D'/p' 决定数据的取舍。

表 2-2　观测次数与对应偏差/概率误差表

n	5	10	15	20	50	100
D'/p'	2.5	2.9	3.2	3.3	3.8	4.2

(5) 若 $D/p>D'/p'$，则舍去该观测数据。

【例 2-2】 某一生油岩有机碳含量数据如表 2-3 所示。请用肖维纳检验法检验其中哪一个是外来值。

表 2-3　有机碳含量数据表

TOC,%	D	D^2
1.17	1.254−1.17=0.084	0.0071
1.15	1.254−1.15= 0.104	0.0108
1.16	1.254−1.16= 0.094	0.0088
1.60	1.60−1.254= 0.346	0.1197
1.19	1.254−1.19= 0.064	0.0410

$$\sum_{i=1}^{n}\mathrm{TOC}=6.27$$

$$\sum_{i=1}^{n}D_i^2=0.1505$$

$$\bar{x}=\frac{6.27}{n}=\frac{6.27}{5}=1.254$$

$$\sigma=\sqrt{\frac{0.1505}{4}}=0.194$$

$$p=0.6745\sigma=0.6745\times0.194=0.1309$$

下面检验 1.60 是否为外来值：

$$\frac{D}{p}=\frac{0.346}{0.1309}=2.64$$

由表 2-2 中可知，当 $n=5$ 时，$D'/p'=2.5$。故 $D/p>D'/p'$，即 $2.64>2.5$，故 1.60 为外来值，应予舍去。

(二)格罗伯斯(Grubps)检验法

当数据 $x_1,x_2,\cdots,x_n$ 由小到大排序服从正态分布时，可用 U 统计量来检验数据是否为外来值。

$$U=(x_n-\bar{x})/S \tag{2-2}$$

$$\bar{x}=\sum_{i=1}^{n}x_i/n,\ S=\sqrt{\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n-1}}$$

式中　U——极值减去平均值形式的统计量。

当 $U>U_{n,\alpha}$ 时，则 x_n 为外来值。不同显著性水平 α 和不同 n 下的临界值由表 2-4 查得。

例 2-2 中，$U=(x_5-\bar{x})/S=\dfrac{1.60-1.254}{0.194}=1.78$。给定 $\alpha=0.01$，则 $U_{n,\alpha}=U_{5,0.01}=1.749$。$U>U_{5,0.01}$，即 $1.787>1.749$。在高度显著性检验下，x_5 为外来值，应予舍去。

同理，采用U统计量还可检验最小值x_1是否为外来值：

$$x_1 = 1.15$$

$$U = \frac{1.254 - 1.15}{0.194} = 0.536$$

$$U_{5,0.01} = 1.749$$

$U < U_{n,\alpha}$，即$0.536 < 1.749$，故x_1为非外来值。

表 2-4　格罗伯斯检验临界值表

n \ α	0.01	0.025	0.05	0.10	n \ α	0.01	0.025	0.05	0.10
3	1.155	1.155	1.153	1.148	15	2.705	2.549	2.408	2.247
4	1.492	1.481	1.463	1.425	16	2.747	2.585	2.443	2.279
5	1.749	1.715	1.672	1.602	17	2.785	2.62	2.475	2.309
6	1.944	1.887	1.822	1.729	18	2.821	2.651	2.504	2.336
7	2.097	2.02	1.938	1.828	19	2.849	2.676	2.527	2.358
8	2.198	2.104	2.011	1.89	20	2.884	2.708	2.557	2.358
9	2.323	2.215	2.109	1.977	21	2.912	2.733	2.58	2.408
10	2.41	2.29	2.176	2.036	22	2.939	2.758	2.603	2.429
11	2.485	2.355	2.234	2.088	23	2.963	2.781	2.624	2.449
12	2.55	2.412	2.285	2.134	24	2.987	2.802	2.644	2.467
13	2.608	2.461	2.331	2.175	25	2.997	2.792	2.682	2.450
14	2.659	2.507	2.371	2.213	—	—	—	—	—

二、原始数据的均匀化和简缩

(一)原始数据的均匀化(网格化)

在对原始地质数据进行定量处理时，往往应先对数据进行网格化处理，即将地质数据标注在规则的矩形网格交点上。网格化的地质数据一般是在平面分布的和位置有关的定量数据，如某个盆地中某地层的厚度数据。象限距离加权平均法是一种简单而又常用的网格化方法。

在以某一个网格点为坐标原点的坐标系的4个象限中，各选一个距该点最近的数据点，假设其平面距离值分别为r_1, r_2, r_3, r_4，相应的数据值分别为z_1, z_2, z_3, z_4，考虑到距离越小，对该网格点的影响越大，因此取距离$r_i (i=1,2,3,4)$的倒数作为权。

该网格点的数据Z可按下列公式进行预测：

$$Z = \sum_{i=1}^{4} \frac{z_i}{r_i} \Big/ \sum_{i=1}^{4} \frac{1}{r_i} \tag{2-3}$$

对每个网格点进行上述计算，即可完成对数据的网格化工作。处理过程中，要特别注意$r_i = 0$的情况，此时网格点上的数据与z_i相同。

(二)原始数据的简缩

当研究区域内的地质数据样品数量很多，或是数据在研究区域内分布极不均匀时，有可能会出现反映相同地质特征的多个近似样品。出现这种情况，不仅会使计算工作量大大增加，而且无助于最终的成果解释，数据量大在计算过程中还会出现计算病态问题。因此，需要对数据进行简缩。

数据简缩的方法有分区加权平均法、分区滑动平均法。

1. 分区加权平均法

假如在一个研究区有 N 个地质数据，则可根据实际工作需要将研究区域分成大小相等或不相等的 m 个区块，要求在每个区块中至少有一个原始数据。如第 j 个区块中有 $n_j(j=1,2,\cdots,m)$ 个数据点，则有：

$$N = n_1 + n_2 + \cdots + n_j + \cdots + n_m$$

令第 j 个区块中每个数据点的权为 $1/n_j$，则每个区块中所包含的数据点的权和等于 1，而且有：

$$n_1 \frac{1}{n_1} + n_2 \frac{1}{n_2} + \cdots + n_j \frac{1}{n_j} + \cdots + n_m \frac{1}{n_m} = m$$

按加权平均在每个区块都可计算出一个综合数据点（重心），从而将原来很大的地质数据量简缩为 m 个有效数据点。

地质数据经常是多变量观测数据，如果每个样品由 p 个地质变量观测值组成，则第 j 个区块第 k 个地质变量观测值的简缩可用下式计算：

$$z_{jk} = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{jki} \quad (j = 1,2,\cdots,m\,;\, k = 1,2,\cdots,p) \tag{2-4}$$

式中 z_{jk}——第 j 个区块第 k 个地质变量的简缩观测值；

n_j——第 j 个区块中地质数据的个数；

z_{jki}——第 j 个区块中第 k 个地质变量的第 i 个数据。

2. 分区滑动平均法

分区滑动平均法与分区加权平均法一样，也要将研究区分成若干个区块，分区的原则与分区加权平均法相同，所不同的是分区滑动平均法要考虑简缩后数据点的位置。

如果第 j 个区块中有 $n_j(j=1,2,\cdots,m)$ 个数据点，每个数据点含 p 个地质变量的观测值，其中第 i 个数据点的坐标为 (x_{jki}, y_{jki})，变量观测值为 z_{jki}。第 j 个区块简缩后的有效数据点的坐标值及变量值可用以下三式计算：

$$x_{jk} = \sum_{i=1}^{n_j} x_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \quad (j = 1,2,\cdots,m\,;\, k = 1,2,\cdots,p) \tag{2-5}$$

$$y_{jk} = \sum_{i=1}^{n_j} y_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \quad (j = 1,2,\cdots,m\,;\, k = 1,2,\cdots,p) \tag{2-6}$$

$$z_{jk} = \sum_{i=1}^{n_j} z_{jki} \Big/ n_j \quad (j = 1,2,\cdots,m\,;\, k = 1,2,\cdots,p) \tag{2-7}$$

式中 x_{jk}, y_{jk}——第 j 个区块第 k 个地质变量观测值简缩后的横坐标与纵坐标；

z_{jk}——第 j 个区块第 k 个地质变量的简缩值；

x_{jki}, y_{jki}——第 j 个区块第 k 个地质变量观测值的第 i 个数据点的横坐标与纵坐标；

z_{jki}——第 i 个区块第 k 个地质变量观测值的第 i 个数据；

n_j——第 j 个区块中的地质数据个数。

按上述公式计算的坐标可能有 p 个，若需要一个统一的坐标点，则可根据地质变量的作用大小，采用加权平均法计算。