高等学校教学用书

石油数学地质

李汉林
赵永军 编著

石油大学出版社

石油数学地质

李汉林
赵永军 编著

石油大学出版社

内 容 提 要

本书的基本内容是介绍石油数学地质的基本原理和主要方法。全书共分十五章，主要介绍常用地质多元统计分析、地质有序数列分析、马尔可夫概型分析、蒙特卡罗法、盆地数值模拟、沉积过程数学模拟、石油资源量及含油气有利地带的预测等内容。本书除介绍基本原理和方法之外，一般附有 FORTRAN 源程序和典型算例供上机实习用。本书一般可供 60 学时讲授，标有"*"的章节作为加选内容。

本书是石油地质专业类本科生及研究生教材，也可作为地球物理、油气开发等相关专业及有关科研人员的参考用书。

图书在版编目(CIP)数据

石油数学地质/李汉林，赵永军编著，-东营：石油大学出版社，1998.9

ISBN 7-5636-1147-9

I.石… II.①李… ②赵… III.石油-数学地质-高等学校-教材 IV.P618.130.2

中国版本图书馆 CIP 数据核字(98)第 25065 号

石油数学地质

李汉林　赵永军　编著

出版者：石油大学出版社(山东 东营，邮编 257062)

印刷者：山东省昌邑市印刷厂印刷

发行者：石油大学出版社(电话 0546—8392563)

开　本：787×1092　1/16　印张：28.125　字数：715 千字

版　次：1998 年 9 月第 1 版　1998 年 9 月第 1 次印刷

印　数：1—2000 册

定　价：29.50 元

前　言

高等学校教学用书《石油数学地质》，自1986年由石油工业出版社出版以来，受到石油院校师生和现场生产、科研人员的欢迎，为我国石油地质人才的培养、石油数学地质方法的推广普及和我国石油数学地质的发展作出了重要贡献。

近年来，石油数学地质学科已从理论研究向应用的方向发展，在某些方向取得了显著进展，产生了一些新的理论和方法，并已能直接为国民经济和地质基础理论研究服务。如含油气盆地数值模拟、地质过程的数值模拟等。另外，原教材在方法基本原理、实际应用方面阐述较简单，所附源程序也已和计算机的迅速发展不相适应。鉴于上述原因，在综合多年教学、科研实践资料和国内外大量科研成果的基础上，我们对1986年出版的《石油数学地质》进行了较全面的修改。

在本书中，保留了原书中多元统计分析、地质数据时间序列分析、蒙特卡罗法等大部分内容，并对其进行了详细的论述和补充了较多的国内外应用实例，以便开阔读者的视野和石油数学地质方法的应用思路。从数学地质的发展及实用角度考虑，增编了数据预处理、马尔可夫概型分析、盆地数值模拟、地质过程的数值模拟等内容，在基本概念、原理及过程、实际应用等方面进行了较详细的介绍。本书在原教材的基础上，改进和增编了FORTRAN语言源程序，完善和补充了有关实例，还结合内容编写了有关习题。作为数学地质重要内容之一的计算机绘图，由于近年来的迅猛发展，已成为体系较完善的分支学科，而且在部分石油院校中已作为独立的课程开设，因此本书中不再介绍这方面的内容。

本书共分十五章，其中第一、三、四、五、六、七、九、十章由李汉林执笔，第二、八、十一、十二、十三、十四、十五章由赵永军执笔(第十五章主要节选自赵旭东编著《石油数学地质概论》)，最后由李汉林、赵永军共同审查定稿。

由于编者的水平有限，缺点错误在所难免，不当之处，恳请读者批评指正。

编著者

1998.5. 东营

目　录

第一章 绪 论

§1 数学地质的产生及其现代含义

地质学是一门以地壳为研究对象，有着悠久历史的自然科学。但是，与其它学科相比，它的定量化程度直到目前仍然是比较低的，造成这种状况的根本原因是地质因素的多样性、地质过程的不可再现性以及它所遗留下来的地质信息的片面性。也就是说，地质学的研究内容，几乎都是发生在地球历史中的一些地质过程，而目前可以观察或测量的一些地质现象都是经历了长期的地质演化过程以后，所遗留下来的地质过程的残留记录，人们只能根据这些残留记录提供的部分地质信息去推断早已发生的地质过程。因此，有相当长的一个时期，地质学的研究方法是：首先对地质现象进行观察、记录和描述，收集实际的地质资料，然后再进行分类归纳和逻辑推理，最后得出相应的地质认识或地质论断。目前的多数地质理论就是由上述研究方法得出、而后又经大量事实所证实了的具有规律性的地质论断。

对于同一个地质问题，应用传统的定性研究方法，会使人们收集的地质资料有所差异，甚至是在同一批地质资料的条件下，也会因地质学家思维方法的差异而得出不同的结论；另外，这种传统的定性研究方法，也就决定了地质学在相当长的一个时期内只能是一门定性的科学。为改变地质学的上述状况，以适应生产发展的需要，必然促使一些人把地质学和其它基础自然科学相结合，以期对所研究的地质问题作出合理的地质解释。本世纪以来，地质学与物理学、化学以及力学相结合，结果产生了地球物理学、地球化学和地质力学以及相应的地质勘探方法。这些新的边缘学科的产生和发展，极大地促进了地质学的发展，使古老的地质学产生了质的飞跃，表现出强大的生命力。

由于测试仪器的不断改进和更新，新的边缘学科所产生的新的物、化探手段不断增加，从而可以观察到从宏观到微观的一切地质的、物理的和化学的特征，使不同类型的地质信息巨增，特别是出现了大量的数值型资料，这就使地质人员无法应用定性的研究方法去处理、分析和利用这些资料，甚至连对资料的阅读都成为十分困难的事，因而大量有用的地质信息，特别是那些数值型的信息就被浪费掉。为了充分利用地质勘探中所获得的各种找矿信息，加强地质研究，减少勘探风险，这就必须在地质学中引入定量的研究方法。地质学中应用数学方法大约始于19世纪40年代初期，但定量化的进程却非常缓慢。

欲想得到地质问题的定量结论，这不仅需要引入定量的数学方法，而且还应有记忆能力强、处理速度快的资料处理设备。50年代初，计算机的批量生产和数字绘图仪的问世，为在地质学中引入定量研究方法创造了条件。找矿难度的不断加大、不同类型地质信息的巨增以及计算机技术的普及，又进一步促进了地质学与数学的结合。从本世纪50年代末期开始，数学方法和电子计算机在地质学中获得广泛应用，到60年代末期形成一门新的边缘学科——数学地质。多数的数学地质工作者认为：数学地质是地质学与数学和计算机科学相互渗透、紧密结合而逐步形成的一门地质学的边缘学科。它是以数学为方法，以计算机为主要的研究手段，定量研究地质学基础理论和定量探矿法的一门方法性科学。随着其它科学的发展和生产的需要，数

学地质的理论和方法也在不断地发展和完善，它已从理论研究向应用的方向发展，并已能直接为国民经济服务，其最终目的是实现地质学的定量化和智能化。

石油数学地质是数学地质在石油勘探、开发及石油资源评价等领域中的应用，其主要任务是研究与石油地质学有关的一些问题的定量化，近年来它的主要研究内容是石油资源评价方法及其相应的软件。

数学地质的基本工作过程可以概括为：由地质学家提出地质问题，分析问题的地质因素，建立相应的地质概念模型；选择合适的数学方法，将定性的地质概念模型转化为定量的数学模型并研制相应的应用软件；对计算机输出的定量结果及地质图形资料进行地质解释，并在此基础上确定或修改给出的地质概念模型及相应的数学模型，以期解决所提出的地质问题。上述工作流程如图 1-1 所示。

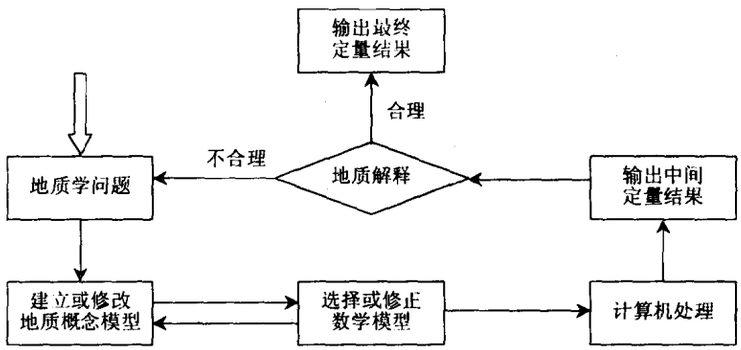

图 1-1　数学地质基本工作流程

§2　数学地质的主要研究内容

一、地质多元统计

地质多元统计是应用宏观统计方法研究地质问题的方法的统称，其中的大多数方法是从数理统计中直接移植来的，少数方法是根据地质工作的实际需要，在移植的基础上逐步发展衍生出来的。

地质多元统计是数学地质的基础，也是石油数学地质的主要方法。本世纪 70 年代以前，数学地质的主要研究内容就是地质多元统计方法及其在地质工作中的应用。

目前，地质多元统计方法已比较完善，它对地质学的定量研究起了极大的促进作用。但是，它在地质学领域中应用的广度和深度还远远不够，而且多数常规地质人员对地质多元统计方法也并不太熟悉。

目前常用的地质多元统计方法有回归分析、趋势面分析、聚类分析、判别分析、因子分析、对应分析、相关分析、时间序列分析、线性映射和马尔可夫概型分析等。

在近年的全国性数学地质学术讨论会上，地质多元统计方面的学术论文数量最多，一般占论文总数的 50%～60%，这就表明地质多元统计仍然是数学地质的重要组成部分，对地质研究工作还在起着重要的作用。地质多元统计在长时间内能够存在并不断发展，这是由地质学本身的特点决定的。因为任何一个地质问题都是非常复杂的，即地质问题具有时间久、空间广和因素多这三个基本特征。因而，地质人员总试图借用统计分析方法从已知信息中获得一些规律性的认识，以便从量的角度研究和分析地质问题，这就决定了地质多元统计在地质研究工作中

具有广阔的应用领域。

二、矿产资源预测

随着社会对矿产资源需求量的不断增加和迅速开采，使找矿的难度和风险也越来越大。从本世纪70年代开始，矿产资源的定量预测就已成为数学地质的重要研究内容之一。实践证明，数学地质定量预测方法的有效性越来越明显，并已得到地质界的普遍重视。

对一个探区进行矿产资源预测，要解决两个基本问题：一是有无矿产？有多少？二是如果有可供开采的矿产，到哪里找？此外，对探区进行经济评价与勘探决策也应作为地质评价的延伸性工作。

矿产资源预测在石油部门称为石油资源评价，其目前的主要内容包括：

(1) 探区石油资源量的估算；

(2) 确定探区中的有利勘探地区；

(3) 石油勘探的经济分析。

目前，我国有关地质找矿部门根据各自的实际需要，针对所找矿产的地质特征，都在大力研究矿产资源的定量预测理论及其相应的计算方法和应用软件。例如，石油部门近年来大力开展盆地数值模拟的理论和方法研究，并已在油气资源评价中起了重要的作用；冶金部门近十年来一直大力发展“地质统计学”。地质统计学是南非金矿矿山工程师克里格(D. G. krige)首先提出的预测金矿矿床的一种方法，后来经过法国的应用数学家马特隆(G. Matheron)加以理论化，而成为一个比较完整、自成体系的研究固体矿床，预测品位空间分布的专门方法。近年来，已在法国及法语系国家形成一支有影响的学派。由于这种方法是由克里格首先提出的，故称其为克里格法。地质统计学的主要研究内容是区域化变量，它是通过变异函数，研究固体矿床的空间变化、勘探方法与储量误差三者之间的数量关系。

三、地质数据库

地质数据库是计算机技术为地质人员服务的一个范例，也是一个国家地质现代化的一个重要标志。地质数据库的出现，使地质人员从繁琐的收集和整理地质资料等非研究性的工作中解脱出来，把精力集中到研究中去。它已成为科学管理地质信息的重要手段，从而为地质问题的定量研究和地质资料处理自动化创造了有利的条件。

数据库是分门别类的存储在一起的相关数据的集合。数据的存储独立于使用它的程序，对于插入新数据，修改和检索原有数据均能按照公用的和可控的方式进行。因而，一个完善的数据库应包括数据的存储、检索、更新、处理、显示、通讯及网络等多种功能。

数据库是本世纪60年代末出现的最新数据管理技术，比较完善的数据库软件系统是70年代初完成的。地质数据库在美国、加拿大、法国、德国等西方国家发展很快，80年代以来，在许多国家已普及应用。目前，世界上大约已建成500多个大、中型地质数据库，这些数据库已涉及到地质学的各个分支领域，其中某些大型数据库已在一个国家甚至许多国家形成网络系统。比较著名的地质数据库有：

1. *计算机化矿产资源信息库*(CRIB)

它是美国地质调查所的矿产资源数据库。库内存储了美国的4万多个矿床和矿产产地以及其他国家的6千多个矿床和矿产产地的资料。数据文件中包括：矿床位置、地质特征、储量、产量等多种数据。用户可以通过计算机网络系统在全世界500多个城市用电话查询和索取这个数据库中的地质数据。

2. 北美石油数据系统(PDS)

该数据库包括了目前公开投入使用的10个石油地质数据库。库中存储了美国和加拿大的10万余个油气田的有关资料。文件内容包括:油气田的产量、生产井数、储量、圈闭类型、储层时代、储层厚度、油气性质、地层温度、地层压力、岩性等多种数据。

3. 井史控制系统(WHCS)

该数据库属于美国的"石油信息公司"。它是世界上目前最大的一个油井数据库,存储了美国100多万口油气井的有关数据。

虽然地质数据库的数据结构、层次的命名术语等方面各有不同,但是,数据库目前大的层次基本上可以归纳为4个级别,即子库、文件、项目、数据。

近年来,国内利用微型计算机管理地质数据的工作进展很快。但是,因为微型计算机的功能与存储空间都很有限,所以建立的数据库大体上相当于上述4个级别中的文件。因此,建立在微型计算机基础上的数据库,一般只适用于某些专项地质工作。

四、地质过程的数学模拟

应用数学模拟的方法研究地质历史的演化过程,是探索地质基础理论的重要途径之一。近十年来,地质过程的数学模拟已发展成为数学地质的重要组成部分,其发展速度较快,例如,通过盆地模拟研究石油地质演化史已成为一个热门课题,并直接服务于油气资源评价。在国际数学地质协会第25周年大会上,许多国家的多篇论文都涉及到地质过程的数学模拟这个问题。例如,瑞典Fookes介绍了模拟沉积旋回性的程序EFPRO-EUSTASY,该程序综合考虑了水动型海平面变化,沉积物供给、堆积速率,以及构造沉降、抬升这三种因素,已成为商业程序由一家美国公司出售。捷克地质调查所用PDI(IES)盆地模拟系统研究了西北维也纳盆地外卡帕推覆体的热演化史,其中仔细地估计了逆冲过程中岩石的物性参数及压实作用的变化。这一研究表明,盆地模拟已由张性盆地进入到压性盆地。俄罗斯Svalova通过建立在流变性和热力学方程基础上的力学—数学模型推导了地幔底辟与沉积盆地形成的关系,指出从较深处快速上升的地幔底辟将导致地壳浅部沉积盆地的形成,否则会导致地壳浅部上隆,这一结果对地幔底辟将导致沉积盆地形成的传统概念是一个挑战。意大利AGIP石油公司介绍了以地质统计学的方法模拟油田的三维岩相分布及各岩相内部孔隙度和渗透率的频率分布,进而开展采油史动力学模拟,以此指导油田开发方案的设计。美国斯坦福大学的哈博教授及他的学生们自60年代以来悉心研究沉积盆地形成过程的模拟。他们研制的SEDSIM系统在模拟沉积物剥蚀、搬运及沉积的动力学过程时考虑了河流的流量、波浪作用、沉积物搬运的力学机制,盆地环流、压实、粒间水的挤出以及均衡补偿等作用。

由上述地质过程模拟的例子可知,地质过程数学模拟的关键是确定定性表征地质过程的地质概念模型和定量描述地质概念模型的数学模型。地质概念模型是指在对地质体系深刻理解和抽象思维的基础上,以定性方式表达地质体系发生和演化过程及其量间关系的模型,而地质数学模型是指用定量方法描述地质体系发生、演化过程及其量间关系的模型。地质概念模型是建立数学模型的基础,把概念模型转化为数学模型是对地质体系认识的深化和概括。为了分析概念模型与数学模型的可靠性,经常采用试验方法对模型进行验证。例如,用水槽试验模拟沉积过程;用泥巴试验、光弹试验模拟构造演化过程等。

数学模型按其使用的数学方法又可分为确定型与随机型模型。然而,任何一个地质过程都不可能是单一的确定型过程或随机过程,而是两种地质过程在时间和空间上的叠加。因而,一个完善的数学模型应该是由上述两种模型构成的复合模型。目前,人们侧重于用单一的确定型

或随机型模型来研究地质历史演化过程，对地质体系认识上的不完全或者把复杂的地质体系进行简化描述是造成使用单一模型的根本原因。

地质过程的数学模拟，实际上就是在对地质体系分析、归纳和逻辑思维的基础上，先提出一个表征地质体系的地质概念模型，并把它转化为一个数学模型，然后通过计算机对数学模型进行反复运算，以再现地质过程的发生和演化史，进而实现对地质过程定量描述所做的一种试验。

五、地质绘图自动化

地质绘图占据了地质科技人员大量的时间，使他们不能有更多的时间集中精力综合研究所进行的地质科研课题，因此地质绘图自动化一直是地质科技人员的一个追求目标，也是数学地质的主要研究内容之一。

随着计算机技术和图形学的发展，计算机自动绘制地质图的工作进展很快，已有形成一门独立学科的明显趋势。对于地质工作中的绝大部分图件，目前计算机不仅可以绘制，而且能够绘制出精美的彩色图件。可以说，在不久的将来就会实现地质绘图的自动化。

上述五个方面就是现阶段数学地质的主要研究内容。这五个方面既相互独立又相互联系。虽然每个方面的研究内容各有侧重，但它们都是为了一个统一的目的，那就是加快地质学的定量化进程，最终实现地质学研究的定量化和智能化。

一个学科的定量化程度表征着它的成熟和完善程度。当今科学技术正处在飞速发展的时代，任何一个学科都在汲取数学的成熟方法以及最新成就，或者依据本学科的需要向数学界提出新的问题，从而促进数学的发展并服务于本学科。地质学的定量化就是用数学的语言描述地质学中的定义、概念和规律等，从而使地质学由定性描述转变到全面的定量描述。

实现地质学的定量化是十分困难的。这除了地质学自身的复杂性外，还存在着其他一些难点，例如同一地质概念的多种含义问题，观测手段的精度问题，地质数据的代表性问题等都在很大程度上阻碍着地质学实现定量化。因此，地质学的定量化将需要一个相当长的历史阶段，需要几代人的不懈努力方能逐步实现。

第二章　地质数据与地质变量

§1　地质数据

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。它是用来表示地质客观事实这一地质信息的。从广义角度来看，地质数据可以是定量、定性数据，也可以是文字说明，甚至是地质图形。从狭义角度来看，地质数据主要是指定量的和定性的地质数据。

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。这类数据一般未进行任何加工处理，所以也称为原始数据。观测数据根据其本身的特点可分为定量数据和定性数据。

1. 定性数据

定性数据是指不能用数值描述，只能用符号或代码描述的观测数据。这种数据不具备数量上的概念，它包括名义型数据和有序型数据两类。

(1) 名义型数据

名义型数据没有数量上的概念，并且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。例如描述岩石颜色的红、绿、灰、黑，可用符号 A、B、C、D 表示，又如，若不深究岩性的详细概念，砂岩、泥岩、灰岩等岩石类型可用符号 S、M、L 表示等等。符号 A、B、C、D 和 S、M、L 就是名义型数据。

名义型数据量之间只存在“相等”或“不相等”的关系，如红色等于红色（$A=A$），砂岩不等于灰岩（$S\neq L$）。

(2) 有序型数据

有序型数据虽没有数量上的概念，但数据之间有次序关系，常以等级符号或代码形式表示。例如干酪根的范氏分类法将干酪根分为Ⅰ、Ⅱ、Ⅲ三个级别，分别用符号 1、2、3 表示，又如鉴定岩石相对硬度的摩氏标准，将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。

有序型数据之间除有相等、不相等关系外，还有“大于”或“小于”关系。如滑石硬度小于石英硬度，从生油潜力看，Ⅰ型干酪根大于Ⅲ型干酪根。

2. 定量数据

定量数据是指能用数值大小来描述的观测数据。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。例如以基准海平面起算的地层分层数据就是典型的间隔型数据。

间隔型数据之间除了具有相等、不相等以及大于、小于关系外，还可以定量说明数据之间的差异，这种差异具有实际意义。如某地层底界和顶界分层深度之差等于该地层厚度。

(2) 比例型数据

比例型数据也有明确的数量概念，可以用数值形式表示。比例数据之间不仅其差值具有实际意义，而且比值也有实际意义。它和间隔型数据的另一个区别是比例型数据是以0为边界的定量数据，即比例型数据是由大于等于0的实数组成的数据集合，而间隔型数据中可能出现负值。如某井各地层厚度数据，其差值表示两个地层的厚度差，比值反映了其中一个地层厚度是另一个地层厚度的百分之多少。

比例型数据所反映的数据概念最完整、意义最明确，因而是最重要的一类数据。

(二) 综合数据

综合数据是指由定量数据(或经定量化处理后的定性数据)，经有限次算术运算后得到的具有明确地质意义的综合性数据，例如总烃含量、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、极差、相关系数等都可认为是综合数据。

(三) 经验数据

经验数据是指在大量研究了地质现象和规律后，经过归纳或根据经验公式计算而得到的经验值。它们通常是大量地质信息的综合反映。经验数据的地质意义往往是十分明确的，但经验数据受到哪些地质因素的影响，以什么方式影响，经验数据和地质因素之间的数学关系是什么，这些问题往往是不确定的或不清楚的。

石油资源评价工作中经常使用经验数据，如单储系数、聚集系数、排烃系数等等。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，使用经验数据时要特别注意对比地质条件的相似性。不加选择地引用将导致错误的结果。

三、地质数据的特点

由于地质系统、地质条件和地质作用的复杂多变，各种技术测试手段之间的较大差异等原因，造成了地质数据本身的许多特点，主要包括以下几个方面：

① 地质数据的类型多，性质不一，反映的地质内容十分广泛，数量的多少和数据的精度相差悬殊，量纲变化大。

② 地质数据往往反映了多种地质因素综合作用的结果，具有混合分布特征。

③ 定量数据仍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成熟。

上述特点说明了地质数据不是单一性质的集合，而是属于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。使用地质数据时要特别注意其适用性，对不同的使用目的要选用不同的数据，同时还要加强和改进数据的加工和处理技术，只有这样才能有效使用地质数据，使数学地质方法取得较好的地质效果。

四、地质数据的误差

任何的观测手段都不可能得到与实际情况完全吻合的观测值。这是因为在野外观测、样品采集、管理、分析化验、仪器读数、资料整理过程中，由于受到工作人员的主观因素、仪器的精度限制、周围环境或随机因素、人为过失等的影响，必然使观测数据产生误差。误差是衡量地质数据质量的重要标志，按性质可分为三种类型。

1. 随机误差

随机误差是指在观测或测量过程中由不可控制的、无规律的偶然因素引起的误差，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，当观测次数增加时，误差均值

趋于0。随机误差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。

2. 系统误差

系统误差是指由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，测量条件或环境的不同，观测者的不同习惯等因素引起的误差都是系统误差。这类误差往往使观测数据整体上偏大或偏小，可以用一定的手段校正观测数据，降低这类误差。

3. 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，称为失真。如样品的污染，仪器的瞬时故障，数据整理过程中的笔误等都可能使观测数据失真。过失误差的多少和大小一般反映了观测人员的水平高低，所形成的失真数据往往是难以校正的，它对数据处理结果会产生极其不利的影响。

五、数据矩阵

地质数据的数量一般比较多，出于数据处理的需要，可将地质数据用数据矩阵表示。矩阵的每一列是一个地质变量的多个观测值，每一行是包含多个地质变量观测值的一个样品。如果地质数据包含 m 个变量的 n 次观测值（n 个样品），则可用下列 n 行 m 列的数据矩阵 X 表示：

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

上述数据矩阵中包括了对 m 个地质变量的 n 次观测值，x_{ij} 表示第 j 个地质变量的第 i 次观测值。由于数学习惯的不同，也可能用每列表示一个样品，每行表示一个地质变量的多个观测值。这样，上述数据矩阵成为下列形式：

$$X=[x_{ij}]_{m\times n}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

其中 x_{ij} 表示第 i 个地质变量的第 j 次观测值。对于不同的数据矩阵表示方式，在处理时应注意区别对待。本书中一般按第一种方式表示数据矩阵。

例如，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项地质变量，5次观测数据如表2-1所示。

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

将上述数据整理为下列5行4列的数据矩阵：

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

§2 地质数据的预处理

由于观测数据的量纲不同及存在各种误差等原因，将原始地质观测数据直接用于计算往往是不合适的。因此在进行正式计算之前需要对观测数据进行预处理，形成供进一步计算使用的数据。地质数据的预处理是定量计算过程中不可缺少的一个重要环节，它已成为数学地质的重要内容之一。

一、定量数据的标准化

不同地质变量原始观测值的单位、量纲以及数值大小、变化范围是不相同的，如果对原始数据直接使用，可能突出观测值较大地质变量的作用，降低观测值较小地质变量的作用。为克服数据中存在的这种不合理现象，在进行计算之前要将各地质变量的观测值变换到某种规范尺度之下，即定量数据的标准化。严格讲，定量数据的标准化包括对变量和样品观测值的标准化，但一般情况下只考虑对变量观测数据的标准化，少数情况下使用对样品的标准化。另外，为计算方便，定量数据的标准化一般在数据矩阵的基础上进行。

1. 总和标准化

总和标准化是指将变量的每个观测值变换为它与该项变量所有观测值总和的比值。因此，在变换后，数据矩阵的元素值在[0,1]之间，且每个变量的所有观测值之和等于1。具体变换公式为：

$$x'_{ij}=\frac{x_{ij}}{x_{\cdot j}}\quad(i=1,2,\cdots,n\quad j=1,2,\cdots,m) \tag{2-2}$$

式中 x'_{ij}—— 变换后的数据；

x_{ij}—— 变换前的数据；

$x_{\cdot j}$—— 第 j 个变量观测值总和，$x_{\cdot j}=\sum_{k=1}^{n}x_{kj}$，$n$ 为样品总数。

对式(2-1)中所列数据矩阵按式(2-2)进行变换，计算出各列(变量观测值)的总和如下：

$x_{\cdot 1}=1400, x_{\cdot 2}=1020, x_{\cdot 3}=12.5, x_{\cdot 4}=10000$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}0.714 & 0.049 & 0.120 & 0.200\\ 0.179 & 0.147 & 0.080 & 0.220\\ 0.071 & 0.069 & 0.240 & 0.150\\ 0.007 & 0.196 & 0.160 & 0.180\\ 0.029 & 0.098 & 0.400 & 0.250\end{bmatrix}$$

当样品数只有二个时，我们可以把每个变量变换后的观测值看成是二维平面上的点，第一个样品中各变量观测值作为 x 坐标，第二个样品中的变量观测值作为 y 坐标。由于进行总和

标准化后的变量观测值之和为1，因此变换后的点必然落在直线 $x+y=1$ 上。若将变换前的原始点和坐标原点连成一条直线，可以证明，该直线和 $x+y=1$ 的交点就是总和标准化后的点。

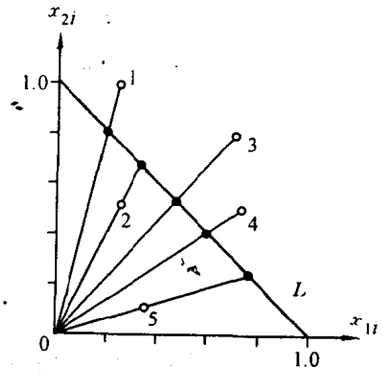

图 2-1 二个样品的总和标准化

例如，有2个样品，5个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\ 1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}0.20 & 0.33 & 0.47 & 0.60 & 0.78\\ 0.80 & 0.67 & 0.53 & 0.40 & 0.22\end{bmatrix}$$

变换前后观测值点的空间分布如图2-1所示。

2. 最大值标准化

最大值标准化是将每个变量的观测值除以该变量所有观测值中的最大者。进行最大值标准化后的观测值中最大值为1。具体变换公式为：

$$x'_{ij}=\frac{x_{kj}}{\max\limits_{1\leqslant k\leqslant n}x_{ij}}\qquad (i=1,2,\cdots,n\quad j=1,2,\cdots,m)\tag{2-3}$$

式中 x'_{ij}——变换后的数据；

x_{ij}——变换前的数据；

$\max\limits_{1\leqslant i\leqslant n}x_{ij}$——第 j 个变量观测值中的最大者。

对式(2-1)中所列数据矩阵按式(2-3)进行变换，1至4列(变量观测值)的最大值如下：

1000,500,5,2500

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}1.00 & 1.00 & 0.30 & 0.80\\ 0.25 & 0.30 & 0.20 & 0.88\\ 0.10 & 0.14 & 0.60 & 0.60\\ 0.01 & 0.40 & 0.40 & 0.72\\ 0.04 & 0.20 & 1.00 & 1.00\end{bmatrix}$$

当样品数只有二个时，可将每个变量变换后的观测值看成是二维平面上的点，由于最大值标准化后每个点必有一个坐标值为1，所以变换后点必然落在直线 $x=1$ 或 $y=1$ 上。

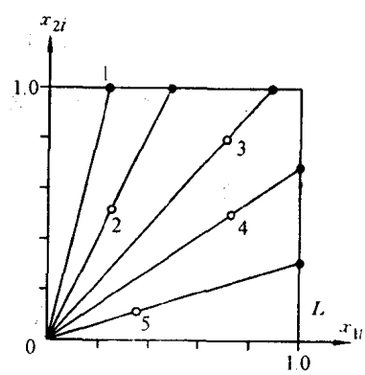

图 2-2 二个样品的最大值标准化

例如，有2个样品，5个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\ 1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}0.25 & 0.50 & 0.86 & 1.00 & 1.00\\ 1.00 & 1.00 & 1.00 & 0.67 & 0.26\end{bmatrix}$$

变换前后观测值点的空间分布如图2-2所示。