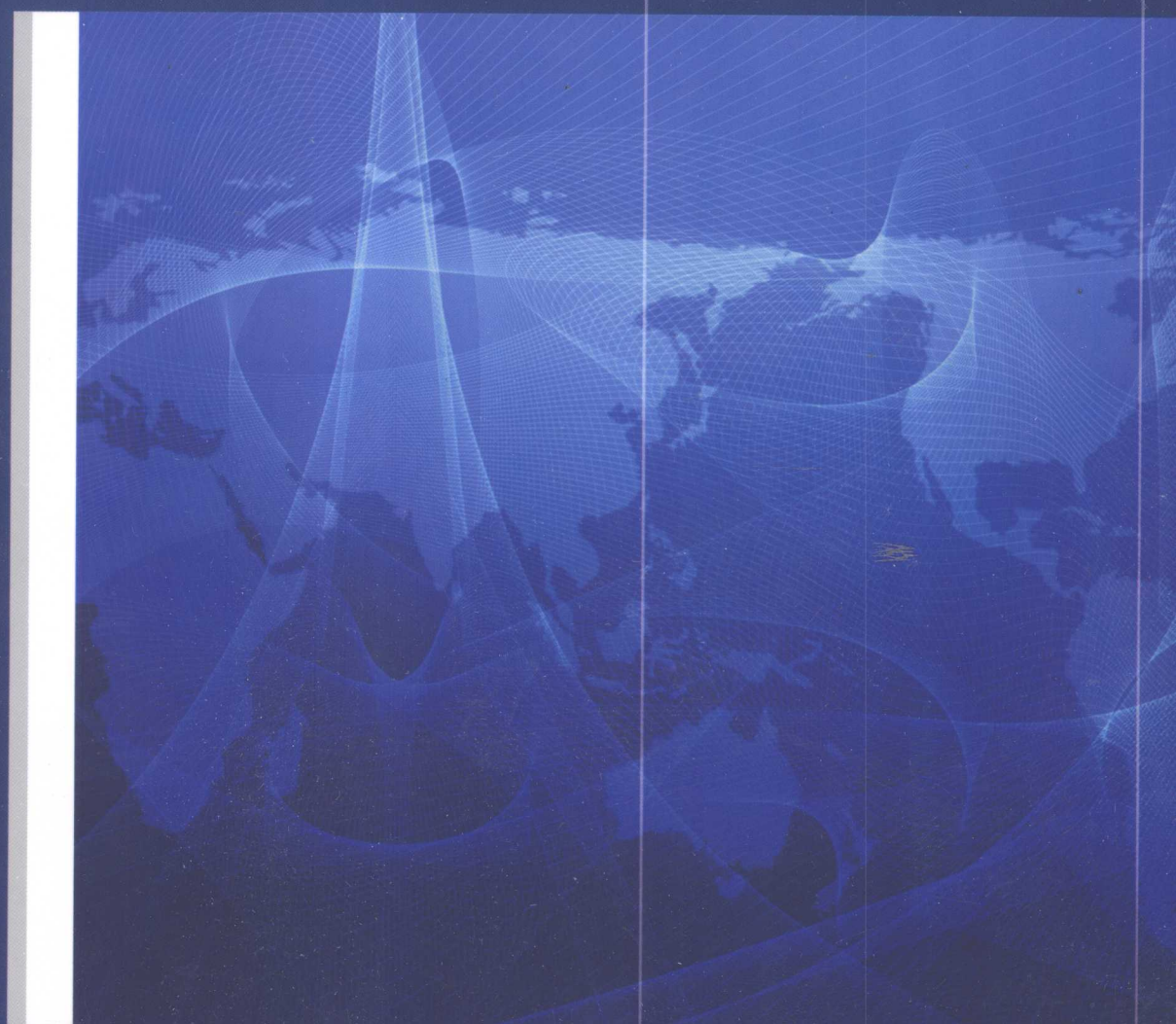
TEACHING MATERIALS FOR COLLEGE STUDENTS
高 等 学 校 教 材

石油数学地质 第二版

Petroleum Mathematic Geology

■ 主编 李汉林 赵永军 王海起



中国石油大学出版社

石油数学地质 第二版

李汉林　赵永军　王海起　主编

中国石油大学出版社

图书在版编目(CIP)数据

石油数学地质/李汉林主编. —2版. —东营:中国石油大学出版社,2008.11
ISBN 978-7-5636-2450-8

Ⅰ.石… Ⅱ.李… Ⅲ.石油天然气地质:数学地质—高等学校—教材 Ⅳ.P618.130.2-05

中国版本图书馆CIP数据核字(2008)第183793号

中国石油大学(华东)规划教材

书　　名:石油数学地质　第二版
作　　者:李汉林　赵永军　王海起

责任编辑:刘　洋(电话 0546—8392860)
封面设计:赵志勇

出 版 者:中国石油大学出版社(山东 东营　邮编 257061)
网　　址:http://www.uppbook.com.cn
电子信箱:shiyoujiaoyu@126.com
排 版 者:中国石油大学出版社排版中心
印 刷 者:青岛锦华信包装有限公司
发 行 者:中国石油大学出版社(电话 0546—8392791)
开　　本:185×260　印张:11.5　字数:278千字
版　　次:2008年11月第2版第1次印刷
定　　价:19.00元

目录

第一章 绪 论

§1 数学地质学的发展简史

数学地质学是采用数学理论和方法，以计算机为主要技术手段，定量化、智能化、可视化地研究地质过程中所产生的地质现象和资源状况的一门地质学边缘学科。

1840年至1935年，是数学在地质学中初步应用和在个别方面进行少量分散研究的时期。1840年莱伊尔利用古生物化石的统计分析对第三系进行划分。1890年皮尔逊(Karl Pearson)编写了《数学进化论贡献》丛书，内有古生物化石的统计分析。1914年至1934年，列文生-列星格(Левинсон-Лессинг)通过考察岩石岩浆系数的频率分布，研究了安山岩、玄武岩、英安岩、流纹岩的分类。1929年勃林克曼(R. Brinkmann)进行了一些生物地层学方面的统计研究工作。

1936年至1945年，数学方法的应用范围由地质学个别问题逐渐扩展至地质学的一些分支。1939年西姆波森(G. G. Simpson)等编著了《定量动物学》一书，为古生物统计学的发展奠定了基础。美国人克鲁拜因(W. C. Krumbein)从1934年开始进行沉积作用和地层的统计分析工作，成为美国数学地质学的奠基人。1944年前苏联维斯捷列乌斯(А. Б. Вистелиус)在前苏联科学院报告集上发表了《分析地质学》一文，提出用定量方法研究地质学问题的初步思想。从此，他从事数学地质工作30余年，成为前苏联数学地质学的创始人和国际数学地质协会的第一任主席。

1946年至1960年，数学方法应用于地质学的许多分支，单变量、双变量统计方法被普遍应用。前苏联已有人研究金属矿床元素的统计分布特点。1954年绍(D. M. Shaw)等应用统计方法研究地球化学问题。1956年初，切叶思(F. Chayes)应用均值、方差、标准差于岩石学研究中。1958年克鲁拜因开始从事区域地层统计分析方面的工作。在此期间，电子计算机的应用和数字绘图仪的诞生，为地质学与数学的结合创造了条件。1958年克鲁拜因首次在地质学杂志上公布电子计算机地质应用程序。

1961年至1970年，数学方法和电子计算机在地质学中开始广泛应用。亚利桑那大学从1961年开始召开了一系列“电子计算机在矿产工业中的应用”讨论会。第二代计算机的成批生产和应用导致数学地质学文献数目激增。1964年达特蒙斯大学第一次成功地应用了计算机分时系统，美国堪萨斯地质调查所召开了第一次“电子计算机在地球科学中的应用讨论会”，《美国石油地质工作者公报》杂志设立了“电子计算机应用”的专门编委。堪萨斯地质调查所从1966年开始连续出版电子计算机程序集。1967年在美国石油地质工作者协会中建立了电子计算机数据存储和索取委员会。同年成立了国际地质科学联合会的地质数据储存、自动处理和索取委员会。1968年在巴黎召开的国际地质会议上成立了国际数学地质协会，并开始出版《国际数学地质协会杂志》和《地质计算程序公报》；美国地质调查所首次公布其电子计算机贡献文集；电子计算机在地球科学应用方面的第一本书出版。在这一阶段，多元统计方法在地质学中大量应用，数学地质学发展成为一门独立的学科。

1971年后，数学地质学科向更高水平发展。地质过程的数学模拟在数学地质学中占据愈来愈重要的地位，愈来愈多的数学方法应用于地质学中；地质统计学取得明显进展，由法语国家向英语国家逐渐推广，并且水平不断提高；地质多元统计有形成独立分支的趋势。数学和地质学的不断结合推动了数学地质学的进展。

20世纪70年代末，中国成立了数学地质学分会，先后召开全国数学地质学术讨论会6次，专题学术会议12次，推动了我国数学地质学的发展。

§2 数学地质学的主要研究内容与方法

数学地质学主要研究概率论与数理统计、地质统计学、地质建模技术、模糊数学、灰色理论、非线性科学、图形处理技术等内容。

(1) 概率论与数理统计。主要包括随机样本分析、随机抽样技术、蒙特卡罗技术、趋势分析、回归分析、判别分析、因子分析、聚类分析、最优分析等，这部分内容以概率论和数理统计理论为基础，对地质数据进行统计分析，得到相应的统计结果并进行地质解释。这些常规的数学地质学方法在地质领域已有较广泛的应用。

(2) 地质统计学。是在地质分析和统计相互结合的基础上，用随机函数的形式体系来评价和探索区域化地质变量的理论和方法。主要包括变异理论、克立金技术等。

(3) 地质建模技术。是运用计算机技术，将空间信息管理、地质解释、空间分析和预测、地学统计、实体内容分析和图形可视化工具等结合起来，对地质对象以及相关的工程活动进行再现和分析的技术。主要包括模型与模拟、建模性质、模型分类、数字油田、勘探数据库与数据银行、地质模型等内容。

(4) 模糊数学。模糊性现象是一种普遍存在的现象，而模糊数学就是研究和处理这些模糊现象的一种数学理论和方法。其基本研究内容有三个方面，即模糊数学的理论以及它与精确数学、随机数学的关系。

(5) 灰色理论。又称灰色系统理论，与研究“随机不确定性”的概率统计和研究“认知不确定性”的模糊数学相比，灰色理论的研究对象是具有“部分信息已知，部分信息未知”特点的小样本、贫信息的不确定系统。灰色系统主要研究方法包括灰色系统建模方法、灰色系统控制理论、灰色关联分析方法、灰色预测方法、灰色规划方法、灰色决策方法等。

(6) 非线性科学。非线性科学研究的是变量间存在的非线性关系。自然现象的复杂性决定了用非线性描述的必要性，它使所进行的描述更为合理可信。地质学中的非线性理论和方法包括分形理论、耗散结构理论、混沌理论、突变理论、协同学、非线性动力学等。

(7) 图形处理技术。是用计算机图形化手段对地质信息和研究成果进行展示和分析的一种技术，其主要内容包括图形算法、计算机显示技术等。随着计算机图形技术的发展，计算机绘图工作进展迅速，在20世纪80年代后期基本上发展成为一门独立的学科。

§3 未来展望

自然科学的一般发展趋势是由定性研究向定量研究发展，而数学地质学正是地质学由定性研究向定量研究发展的主要手段，虽然目前在地质学研究中并不占主导地位，但由于它和尖端科学技术在地质学中的应用密切相关，可以预见数学地质理论和方法必将在地质学研究中发挥十分重要的作用。今后数学地质学的发展主要有以下几个方向：

(1) 传统数学地质学方法的进一步改进与应用。以多元统计分析为主的传统数学地质学方法,目前已在许多地质定量研究中得以应用,事实也证明了这些方法的有效性。但这些传统数学地质学方法在取得成功的同时,也暴露出不少问题。如数学模型与地质概念模型的吻合程度因所研究的对象不同而有较大差异,甚至有些偏差过大,导致处理结果无法解释地质现象。虽然原因是多方面的,但数学模型本身的局限性也显而易见。今后,在合理完善数学模型的基础上,传统的数学地质学方法依然有强大的生命力。

(2) 非线性科学在地质学中的进一步应用。地质系统是一个复杂的巨系统,具有非平衡性、非线性、多尺度性、突变性、自组织性、自相似性、有序性和随机性等特点,因此,为了研究和解决地质系统的重大理论和实际问题,必须应用非线性理论和方法。一个被称为"非线性地质学"的新发展方向越来越引起人们的重视,将逐渐得到推广和应用。

(3) 地理信息系统(geographic information system,GIS)在地质学中的推广应用。地理信息系统是计算机软硬件支持下的空间数据输入、存储、检索、运算、显示和综合分析的应用技术系统,该系统的研究重点是空间实体及其相互关系,主要用途是分析和处理在一定地理区域中分布的各种现象和过程。近年来,地理信息系统已逐渐应用于地质研究。由于该系统自身的特点及优势,今后必将成为数学地质学研究的重要方向。

(4) 人工智能在地质学中的应用将进一步发展。除了继续建立各种地质专家系统外,将更重视多种人工智能技术的综合应用,特别是应用人工神经网络和遗传算法,使人工智能在地质学中的应用达到一个新的高度。

(5) 地质数据库共享技术。在计算机网络技术迅速发展的基础上,利用数据库共享技术,可以充分使用分布在各地的通用勘探数据库、图形库及数据银行中的数据资源,随时随地提供最新研究资料。

(6) 数学新理论新方法、计算机新技术的不断引进。对一些用目前的数学方法难以描述的地质现象,随着数学理论和计算机技术的发展,问题将得以合理解决,从而提高数学地质学方法的应用价值。

(7) "数字油田"技术的推广应用。利用测绘、数据库、地理信息系统、因特网、虚拟现实等技术,可将油田勘探、开发、生产过程中的资料采集、处理、存储、显示等环节实现数字化和可视化。结合分析和运算功能,可更加准确地预测油气资源的分布状况,实现决策的合理性及提高研究工作效率。

第二章　地质变量与地质数据

§1　地质变量

一、地质变量的概念及其分类

1. 地质变量的概念

地质变量是反映某地质现象在时间或空间上变化规律的量。如生油岩的厚度、地层的埋藏深度、生油岩中有机质的丰度等。

2. 地质变量的分类

造成地质现象的因素的复杂性，导致表示不同地质特征的地质变量各不相同。但是，可以根据它们所取数据的性质及方法，将其分为观测变量（定性变量和定量变量）和综合变量。

观测变量是可以直接进行观测、分析或度量的地质变量。如地层的厚度、石油的密度和粘度、岩石的颜色等。

综合变量是把两个或两个以上的观测变量按一定的方式进行组合而得到的具有综合意义的地质变量。如区分天然气成因类型的甲烷系数 $M(M = C_1 / \sum_{i=1}^{5} C_i)$，当 $M > 99\%$ 时，认为是生物成因气，否则认为是热解成因气。又如有机质转化率（总烃与有机碳之比）。

3. 地质变量的观测值

用各种化学、物理以及直接观测的方法获得的地质变量的各种数据和其他形式记录的资料统称为地质变量的观测值。

二、地质变量的特征

1. 具有明确的地质意义

地质意义主要是指对地质变量所代表的特定研究对象的认识，主要包括：对地质变量所代表的石油地质特征的认识，如地层的时代、地层温度、圈闭闭合面积等；对地质变量所代表的盆地地球化学特征的认识，如有机质类型和丰度、干酪根成熟度等；对地质变量所代表的地球物理特征的认识等。

2. 具有明显的统计性质

很多地质变量是随机变量，因此，它们的观测值具有明显的统计意义，如观测值的平均值是地质变量数学期望的估计值，而观测值的方差反映了地质变量在研究区域上的变异。

3. 具有相关性

地质变量之间具有一定程度的相关性，如岩石的渗透率与有效孔隙度密切相关。

三、地质变量的选择

分析研究地质变量的目的是想通过它们预测地质体的特征及有关的地下资源。用什么样的地质变量才能较好地实现研究目标，这就是地质变量的选择问题。例如，要想通过一些地质变量预测某沉积单元的油气资源量，就要选择与油气资源量相关的生油条件、储集条件、保存条件、圈闭条件等地质变量。一般来说，地质变量的选择应遵循以下基本原则：

(1) 基于地质概念模型。以相关地质学科理论为指导,分析、建立研究问题的地质概念模型,依据地质概念模型选择相应的地质变量。

(2) 结合地质变量间的相关性。地质变量之间存在着不同程度的相关性,应选择与矿藏形成或地质体特征等有密切联系的控矿因素和找矿标志。

(3) 地质变量要有代表性。地质变量的代表性是指所选择的地质变量能否较好地表征某地质作用过程的进行程度,或者变量的观测区与未观测区之间的相似程度。

(4) 地质变量要有明确的地质意义。拟定的地质变量,特别是构造的综合型地质变量要有确切的地质含义。如在油气资源评价中,生油岩体积与沉积岩体积之比表示评价区的生油条件,而近油源圈闭面积与沉积岩面积之比则表示评价区的圈闭条件,总烃与有机碳之比表示有机质转化率等。

§2 地质数据

一、地质数据的概念

用以代表地质体或其他自然产物特性的实物样子称为样品。地质样品的采集对象有岩体、地层、矿体、油气、生油岩、储集层、土壤及各种松散的沉积物、地表水及地下水、植物、空气等。用各种物理、化学方法以及直接观测的方法获得的用以表示样品特性的各种数据和其他形式记录的资料统称为地质数据或样品变量观测值。

二、地质数据的分类

地质数据是地质样品的变量观测值。因此,从狭义上讲地质数据分为定性数据和定量数据,从广义上讲它可以是定性数据、定量数据、图形或其他形式记录的资料等。根据地质数据的来源,地质数据分为观测、综合、经验数据三类。

1. 观测数据

指对样品(或采样对象)用各种物理、化学或直接观测的方法获得的表达样品(或采样对象)特性的数据。这种源于样品、没有经过任何加工处理的数据,又称为原始数据。依据数据的性质,又分为定性数据和定量数据两类。

(1) 定性数据。定性数据是指用符号或代码表示的没有数量概念的观测数据。可将其分为名义型和有序型两类:

① 名义型数据是没有数量概念和次序之分,但彼此之间有"相等"或"不相等"关系的定性数据。如岩石的红、绿、灰、黄色可以用字母 A,B,C,D 表示,又如砂岩、泥岩、灰岩可以用 S,N,H 代替,它们之间有 $A=A,A\neq B,S=S,S\neq N$ 的关系。

② 有序型数据是没有数量概念,但彼此之间具有次序关系的定性数据。如Ⅰ,Ⅱ,Ⅲ型干酪根可用数字 1,2,3 表示,它们之间有Ⅰ型干酪根的生烃潜力优于Ⅱ型干酪根的关系。

(2) 定量数据。定量数据是指用数值来描述的观测数据。可将其分为间隔型数据和比例型数据两类:

① 间隔型数据是有明确数量概念和地质含义的定量数据。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。它们之间具有相等、不等以及大于、小于关系,其差异具有实际的地质意义。如某地层底界和顶界分层深度值之差等于该地层的厚度。

② 定量数据的比值构成比例型数据。这类数据本身及它们的差值都有实际意义。比例型数据是大于等于 0 的实数组成的数据集合,这是它与间隔型数据的一个重要区别。如

两地层厚度的比值反映其中一个地层厚度是另一个地层厚度的百分之几，或者反映某种沉积环境，或者反映生油条件等。

2. 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定量数据。这种数据具有明显的地质意义，例如总烃含量、时间-温度指数、生油岩厚度与沉积岩厚度的比等。另外，随机变量的各种数字特征，如平均值、标准差、极差、相关系数等都可视为综合数据。

3. 经验数据

经验数据是在研究地质现象和规律的基础上，根据大量实际资料和经验总结归纳出的数据，如单储系数、排烃系数、聚集系数等。经验数据是大量地质信息的综合反映，地质意义明确，但它受哪些主控因素的影响，以及各因素之间的作用关系等问题目前尚不清楚。另外，经验数据还具有较明显的地域性。因此，在油气资源评价等工作中使用经验数据时，要特别注意对比地质条件的相似性。

三、地质数据的主要特点及数据矩阵

1. 地质数据的主要特点

由于地质系统、地质条件和地质作用的复杂性，测试手段的差异等，导致地质数据有如下几个主要特点：

（1）地质数据类型多，性质不一，地质内涵丰富，量纲不统一，定量数据的数量级相差大，各类数据的数量和精度相差悬殊。

（2）地质数据往往是多种地质因素综合作用的结果，故具有混合分布特征。

（3）地质数据以定量数据为主，而定性数据的定量化研究和应用目前尚不成熟。

地质数据的特点决定了地质数据不是单一性质的数据集合，而是多种来源的混合数据集合，这一特点客观存在且不易改变。使用地质数据时，要注意它们的适用性，同时还要研究和改进数据加工和处理技术，发挥各种地质数据的作用，方可使地质研究获得良好的效果。

2. 数据矩阵

为便于数据处理，地质数据常用数据矩阵表示。假设有 n 个样品，每个样品有 m 个变量，那么样品变量的观测值可用以下数据矩阵 $\boldsymbol{X}$ 表示：

$$\boldsymbol{X}=(x_{ij})_{n\times m}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

式中 x_{ij} —— 第 i 个样品第 j 个变量的观测值。

常把 $\boldsymbol{X}$ 的第 j 列记为 X_j，它是第 j 个变量的 n 次观测值。有时也将数据矩阵记为：

$$\boldsymbol{X}=(x_{ij})_{m\times n}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

式中 x_{ij} —— 第 i 个变量的第 j 次观测值。

例如，地质圈闭的5个参数(表2-1)可以用5行4列的矩阵式(2-1)表示。

表2-1 地质圈闭数据

圈闭编号	闭合面积/(10^2 m^2)	闭合高度/m	长短轴比	埋藏深度/m
1	1 000	500	1.5	2 000
2	250	150	1.0	2 200
3	100	70	3.0	1 500
4	10	200	2.0	1 800
5	40	100	5.0	2 500

$$\boldsymbol{X}=(x_{ij})_{5\times4}=\begin{pmatrix}1\,000 & 500 & 1.5 & 2\,000\\ 250 & 150 & 1.0 & 2\,200\\ 100 & 70 & 3.0 & 1\,500\\ 10 & 200 & 2.0 & 1\,800\\ 40 & 100 & 5.0 & 2\,500\end{pmatrix} \tag{2-1}$$

§3 地质数据的预处理

地质数据的预处理是指在定量研究地质问题时，预先对原始数据进行的各种处理。其主要内容为定量数据的标准化、定性数据的定量化、原始数据的网格化、原始数据的简缩和增补、离群数据的识别与剔除等。

一、定量数据的标准化

定量数据的标准化是对变量的观测值进行标准化。其目的是消除或抑制不同变量观测值数量级的巨大差异，使它们在同一尺度范围下参与地质研究。标准化方法有标准差标准化、极差标准化、极差正规化、总和标准化、最大值标准化、模标准化和中心标准化等。其中最常用的是标准差标准化、极差标准化和极差正规化。

1. 标准差标准化

标准差标准化是变量 X_j 的每个观测值 x_{ij} 减去观测值的平均值 $\bar{X}_j$，再除以观测值的标准差 S_j，即数据矩阵 $\boldsymbol{X}$ 中第 j 列上的每个元素减去该列元素的平均值，再除以第 j 列元素的标准差，最终得到变量 X'_j。变换公式为：

$$x'_{ij}=(x_{ij}-\bar{X}_j)/S_j \qquad (i=1,2,\cdots,n;j=1,2,\cdots,m) \tag{2-2}$$

式中 x'_{ij}——标准化后的数据；

x_{ij}——标准化前的数据(原始数据，即第 i 个样品第 j 个变量的观测值)；

$\bar{X}_j$——第 j 个变量观测值的平均值，$\bar{X}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}(j=1,2,\cdots,m)$；

S_j——第 j 个变量观测值的标准差，$S_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{X}_j)^2}(j=1,2,\cdots,m)$。

变量 X'_j 叫做标准化变量，其特点是平均值 $\bar{X}'_j=0$，标准差 $S'_j=1$，故变量 X'_j 又叫做规格化变量。

对式(2-1)中的数据标准差标准化后，得到新的数据矩阵：

$$\boldsymbol{X}' = \begin{bmatrix} 1.949 & 1.916 & -0.707 & 0.000 \\ -0.081 & -0.350 & -1.016 & 0.587 \\ -0.487 & -0.867 & 0.354 & -1.468 \\ -0.731 & -0.026 & -0.354 & -0.587 \\ -0.650 & -0.673 & 1.768 & 1.468 \end{bmatrix}$$

2. 极差标准化

极差是第 j 个变量 X_j 观测值的最大值与观测值最小值的差，即：

$$\Delta X_j = \max_{1\leqslant i\leqslant n} x_{ij} - \min_{1\leqslant i\leqslant n} x_{ij} \qquad (j = 1,2,\cdots,m)$$

极差标准化是变量 X_j 的每一个观测值 $x_{ij}(i=1,2,\cdots,n)$ 减去 X_j 观测值的平均值 $\bar{X}_j$，再除以极差 ΔX_j。变换公式为：

$$x'_{ij} = (x_{ij} - \bar{X}_j)/\Delta X_j \qquad (i = 1,2,\cdots,n; j = 1,2,\cdots,m) \qquad (2\text{-}3)$$

式中 x'_{ij}——标准化后的数据；

x_{ij}——标准化前的数据(原始数据)；

$\bar{X}_j$——第 j 个变量观测值的平均值；

ΔX_j——第 j 个变量观测值的极差。

极差标准化后，变量 X'_j 的极差为 1。对式(2-1)中数据极差标准化后，得到新的数据矩阵：

$$\boldsymbol{X}' = \begin{bmatrix} 0.727 & 0.688 & -0.250 & 0.000 \\ -0.030 & -0.126 & -0.375 & 0.200 \\ 0.182 & -0.312 & 0.125 & -0.500 \\ -0.273 & -0.009 & -0.125 & -0.200 \\ -0.242 & -0.242 & 0.625 & 0.500 \end{bmatrix}$$

3. 极差正规化

极差正规化的变换公式为：

$$x'_{ij} = (x_{ij} - \min_{1\leqslant i\leqslant n} x_{ij})/\Delta X_j \qquad (i = 1,2,\cdots,n; j = 1,2,\cdots,m) \qquad (2\text{-}4)$$

式中 x'_{ij}——标准化后的数据；

x_{ij}——标准化前的数据(原始数据)；

ΔX_j——第 j 个变量观测值的极差；

$\min\limits_{1\leqslant i\leqslant n} x_{ij}$——第 j 个变量观测值的最小值。

对式(2-1)中数据极差正规化后，得到新的数据矩阵：

$$\boldsymbol{X}' = \begin{bmatrix} 1.000 & 1.000 & 0.125 & 0.500 \\ 0.242 & 0.186 & 0.000 & 0.700 \\ 0.091 & 0.000 & 0.500 & 0.000 \\ 0.000 & 0.302 & 0.250 & 0.300 \\ 0.030 & 0.070 & 1.000 & 1.000 \end{bmatrix}$$

由式(2-4)可知，变量 X'_j 最大值为 1，且 $x'_{ij} \geqslant 0$，即新数据在区间[0,1]内。

二、定性数据的定量化

定性数据的定量化是指把定性数据变换为数值。根据定性数据状态的多少，可分为二态和多态有序定性数据。两类定性数据的定量化方法都是对定性数据的状态赋值。

1. 二态定性数据的变换

只有两种对立状态的定性数据为二态定性数据。可用 0 和 1 表示这两个状态，从而实现定性数据的定量化。如某观测点有无某种化石，就只有两种可能，若有则用 1 表示，若无就用 0 代表。一般来说，按以下原则处理：

二态定性数据	状　态	肯定或有利	否定或不利
	赋　值	1	0

2. 多态有序定性数据的变换

多态有序定性数据是指状态多于两个，并且状态又可按一定次序排列的定性数据。如储层岩心的含油性，按含油程度可分为四级，采用等差方式赋值如下：

四态有序定性数据	状　态	不含油	油　斑	含　油	饱含油
	赋　值	0	1	2	3

又如，按颜色可将泥岩分为四级，为区分各级泥岩的生油能力，可采用非等差方式赋值如下：

四态有序定性数据	状　态	红　色	浅灰色	灰　色	黑　色
	赋　值	0	1	3	5

一般按以下原则处理：

多态有序定性数据	状　态	状态 1	状态 2	状态 3	…
	赋　值	x_1	x_2	x_3	…

三、原始数据的网格化

原始数据的网格化是指把平面上无规则分布的定量数据 z_i 分配到矩形网格的每个交点上(图2-1)，产生规则分布的定量数据。网格化的方法很多，如全点插值法、圆内插值法、曲面插值法、克立金法等。在此仅介绍既简单又实用的按象限取点距离加权平均法。

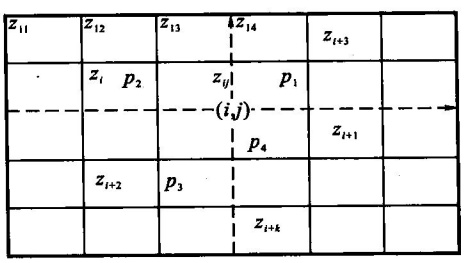


图 2-1　原始数据网格化示意图

以网格交点(i,j)为原点建立坐标系(图 2-1)，在各象限内各取一个距点(i,j)最近的数据点，记为 $p_i(i=1,2,3,4)$，各点上相应的数据值分别为 $z_i(i=1,2,3,4)$，可以用 p_i 点上的数据值 z_i 的算术平均值

$$\bar{z}=\frac{1}{4}\sum_{i=1}^{4}z_i$$

作为网格交点(i,j)上的估计值z_{ij}。

考虑到数据点距网格交点(i,j)越近，对网格点的估计值影响越大，因此取距离的倒数作为权求网格交点(i,j)的估计值。

假设数据点$p_i(i=1,2,3,4)$到网格交点(i,j)的距离为$d_i(i=1,2,3,4)$，那么网格交点(i,j)上的估计值为：

$$z_{ij}=\sum_{i=1}^{4}\frac{z_i}{d_i}\Big/\sum_{i=1}^{4}\frac{1}{d_i} \tag{2-5}$$

在按式(2-5)计算z_{ij}的过程中，当出现$d_i=0$时，则以z_i作为网格点(i,j)上的估计值。

对于某些网格点(如边界网格点)，不能在四个象限中都找到数据点，则在有数据点的象限内取距离近的点上的数据进行加权平均(至少有一个数据)。在按象限取点距离加权平均插值法中，每个象限内也可以取多个距插值点近的数据点进行加权平均，其过程与上类似。

对每个网格交点进行上述计算，即可完成对原始数据的网格化工作。

【例 1】如图 2-2 所示，已知四个数据点$p_1(4,4)$，$p_2(1,4)$，$p_3(1,2)$和$p_4(5,2)$，各点上的数据值依次为 3，2，2，2，求插值点$p(3,3)$上的估计值。

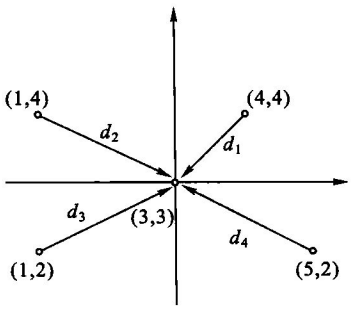


图 2-2　点的插值

解：

① 各点到点$p(3,3)$的距离：$d_1=\sqrt{2}$，$d_2=d_3=d_4=\sqrt{5}$。

② 式(2-5)的分母：$\sum_{i=1}^{4}\frac{1}{d_i}=2.048\ 7$。

③ 式(2-5)的分子：$\sum_{i=1}^{4}\frac{z_i}{d_i}=4.804\ 5$。

④ 插值点的值：$4.804\ 5\div 2.048\ 7=2.345\ 1$。

四、原始数据的简缩和增补

1. *原始数据的简缩*

当分布在研究区上的地质数据点很多(可能出现反映相同地质特征的多个近似数据点)时，或者是数据在研究区上的分布极不均匀时，不仅会使计算量增加，而且也无助于最终的成果解释，甚至在计算过程中还会出现不可预料的计算病态问题。因此，就需要对作用不大或相近、可有可无的多余数据予以舍弃，这就是数据的简缩。

数据的简缩方法一般包括分区加权平均法、分区滑动平均法和随机删点法。

(1) 分区加权平均法。

假设研究区内每个地质数据点有m个变量，根据实际需要将研究区划分成大小相等或不等的n个小区，并且每个小区内至少有一个数据点，那么第j个小区内第i个数据点上第k个地质变量的观测值为$z_{jki}(j=1,2,\cdots,n;k=1,2,\cdots,m;i=1,2,\cdots,n_j)$，而第$j$个小区内第$k$个地质变量的简缩值为：

$$z_{jk}=\frac{1}{n_j}\sum_{i=1}^{n_j}z_{jki}\qquad(j=1,2,\cdots,n;k=1,2,\cdots,m) \tag{2-6}$$

式中　z_{jk}——第j个小区第k个变量观测值的简缩值；

n_j——第j个小区地质数据点数；

z_{jki}——第j个小区第i个数据点上第k个变量的观测值。

按照式(2-6)对研究区内原始数据进行处理后，相当于每个小区内有一个有效数据点，从而将原来大量的数据点简化为 n 个有效数据点。

(2) 分区滑动平均法。

分区滑动平均法的分区方法和分区原则与分区加权平均法相同，但这种方法要考虑简缩后数据点的位置。

如果第 j 个小区内有 n_j 个数据点，每个数据点上有 m 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})，那么第 j 个小区简缩后的有效数据点坐标值及变量由式(2-7)，(2-8)给出：

$$\begin{cases} x_{jk} = \sum_{i=1}^{n_j} x_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \\ y_{jk} = \sum_{i=1}^{n_j} y_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \end{cases} \tag{2-7}$$

$$z_{jk} = \sum_{i=1}^{n_j} z_{jki} / n_j \tag{2-8}$$

$$(j = 1,2,\cdots,n;k = 1,2,\cdots,m)$$

式中 x_{jk}，y_{jk}——第 j 个小区第 k 个地质变量观测值简缩后的横坐标和纵坐标；

z_{jk}——第 j 个小区第 k 个地质变量的简缩值；

x_{jki}，y_{jki}——第 j 个小区第 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 j 个小区第 k 个地质变量观测值的第 i 个数据；

n_j——第 j 个小区地质数据点数。

按上述公式算出的坐标有 m 个，如果需要一个统一的坐标点，则可根据地质变量观测值的大小，采用加权平均的方法算出。另外，根据实际需要，也可采用其他的计算方法。

(3) 随机删点法。

对于探区内的局部数据点密集区，随机删去一些数据点，既可减少计算工作量，又可提高计算过程的稳定性。删除点的方法是对数据点编号，用随机抽样法删去其中的一些数据点。

2. 数据的增补

在一般情况下，探区内投入的工作量是不均匀的，特别是勘探早期阶段。因此在区域上会出现数据点空白区，在这种空白区往往需要补充一些数据点，这就是数据的增补。

在数据点空白区补充数据点时，可以用临近数据点上的数据外推，即根据数据的变化趋势补充适量的数据点，也可以用某种插值方法补充一定数量的数据点。值得注意的是：补点的目的是为了全区计算的稳定性，而原空白区的计算结果仅供参考。

此外，对于多变量的地质样品，由于分析化验项目不一定完全一致或其他原因，导致某些样品缺少某些变量的观测值。对于那些研究需要而又缺少观测值的变量，可以用该变量邻近区域上观测值的平均值作为该变量观测值的近似值。

五、离群数据的识别与剔除

相对研究区的观测数据来说，局部的异常高值和异常低值称为离群数据。这种数据往往直接影响到基于观测数据的数据处理过程和对计算结果的合理解释。对于某些已知因素造成的离群数据，可进行相应的数据校正。如在油气地表化探中，由地表自然地理条件、土

壤的类型和颜色等导致的有关指标含量的差异，经过相应的校正，即可以消除数据中的干扰。

如果离群数据是地质现象的真实反映，当它们对数据处理过程和计算结果产生消极影响时，应对这些数据进行适当处理。对于那些人为等因素造成的错误数据，理所应当地删除或重新进行观测。然而，判断造成离群数据的因素是很困难的。在实际工作中，总是假设数据是真实的，在此假设下讨论对离群数据的挑选和处理。

对离群数据进行处理的第一步工作是挑选离群数据，这就涉及离群数据的界限问题。下面简单介绍离群数据的界限确定和处理方法。

1. 类比法

以实际工作经验确定离群数据的界限，以此界限识别区域上的离群数据。斯米尔诺夫根据实际经验，总结出确定矿床品位离群数据的界限（表 2-2，其中离群品位高出平均品位的倍数项是经验数据）。

从矿床成因角度看，绝大多数的油气藏都属于与沉积岩有关的矿床。所以，确定油气勘探、开发中石油地质数据界限时，可以参照表 2-2 中的Ⅰ，Ⅱ矿床类型。

表 2-2　矿床品位离群数据的界限

矿床类型	组分分布性质	典型矿床	离群品位高出平均品位的倍数
Ⅰ	很均匀	大多数沉积矿床	2～3
Ⅱ	均　匀	复杂沉积矿床与变质矿床	4～5
Ⅲ	不均匀	绝大多数有色金属矿床	8～10
Ⅳ	很不均匀	大多数稀有金属矿床和金矿床	12～15
Ⅴ	极不均匀	某些稀有金属矿床和金矿床	＞15

2. 计算法

用经验公式确定离群数据的界限。沃洛多莫夫给出的计算离群数据界限的经验公式为：

$$ch = c_1 + (n-1)c_1(c_1 - c_2)/c_2 \tag{2-9}$$

式中　ch——正常数据的最大值，大于 ch 的数据即为离群数据；

c_1——校正前（包括离群数据）的样品平均值；

c_2——校正后（不包括离群数据）的样品平均值；

n——样品总数（包括离群数据）。

在一组数据中，离群数据一般只有少数几个，当个数太多时就不应是离群数据。在实际计算时，令$(c_1-c_2)/c_2=20\%\sim30\%$，由式(2-9)可求出离群数据的界限值。此界限值显然与样品数 n 有关。n 越大，ch 的偏离可能越大，故此法适合对小子样的检验。

3. 统计检验法

在观测数据来自同一个总体的前提下，统计检验法的思路是检验数据是否服从正态分布。若通过检验，则认为数据中不存在离群数据，否则认为数据中存在离群数据，这时就需要识别出其中的离群数据并对其进行有关的处理。统计检验的关键是构造一个合适的统计量及其所服从的分布，在此基础上确定相应的假设检验方法。

（1）正态分布的 χ^2 检验法。