



当代
杰出青年
科学文库

空间变异 理论及应用

张仁铎 著



科学出版社
www.sciencep.com

内 容 简 介

本书较全面地概括了国内外有关空间变异理论及地质统计学应用方面的知识和新成果,其中也包括不少作者近年来的研究工作.全书共包括9章,主要介绍随机函数和变异函数,各种不同的克里金估计方法,协克里金法,拟协克里金法,估值的不确定性和采样策略,随机场产生的方法以及变量的随机模拟,多孔介质的空间变异性及尺度效应.还提供教学用计算机程序.

本书既可作为水文地质、土壤科学、农田水利、流域管理、农业生态学、植物学、生物学、采矿业、石油工程、土木工程、环境工程、地理学、地理信息系统、遥感、气象学、统计学等专业研究生和大学高年级学生的参考书,也可供相关专业科研、教学、工程技术人员参考.

图书在版编目(CIP)数据

空间变异理论及应用/张仁铎著. —北京:科学出版社,2005

(当代杰出青年科学文库)

ISBN 7-03-014848-7

I. 空… II. 张… III. 空间-随机变量-应用-地质学:统计学

IV. P628

中国版本图书馆 CIP 数据核字(2005)第 000734 号

责任编辑:李 锋 张 扬 罗 吉/责任校对:张怡君

责任印制:钱玉芬/封面设计:陈 敬

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>

中国科学院印刷厂印刷

科学出版社发行 各地新华书店经销

*

2005年1月第 一 版 开本: B5(720×1000)

2005年8月第二次印刷 印张: 12 1/2

印数: 1 501—2 500 字数: 237 000

定价: 48.00 元(含光盘)

(如有印装质量问题,我社负责调换(科印))

前 言

统计学方法是用来定性或定量地描述众多自然现象内在规律的最有效的工具之一.然而,大多数统计学方法都没有考虑和利用到所收集的资料中空间位置所包含的信息.对于我们研究的环境科学中的许多空间变量,比如降雨的分布、植被覆盖密度的分布、土壤物理化学特性的分布、污染物在某一区域的浓度分布等等,在空间上可能是相互关联的,这种空间连续性是许多自然现象的重要特性.地质统计学提供了必要的理论和方法来描述这种空间连续性,在运用传统的统计学方法的同时充分利用这种空间连续性,因而使地质统计学方法更合理,所得的结果也更精确,从而比传统的统计学方法更具优越性.

地质统计学是从 20 世纪 60 年代起,特别是在 70 年代以后,由克里金(Krige)等人在对南非金矿储量评估的基础上发展起来的.为了解决与空间变异性相联系的矿藏估计的一系列问题,比较系统的理论和方法就逐步建立起来了.这些理论包括了对空间随机分布变量的相关性进行定量描述和空间插值的一整套特别的统计方法.地质统计学正在被广泛地应用于地质学、水文学、土壤科学、气象学、农业、遥感、石油工程、生态、海洋、资源与环境以及其他研究“时空变量”的领域.随着地理信息系统(GIS)的广泛运用,与之相伴的地质统计学的知识和计算方法就必不可少.

“空间变异理论及应用”是作者在美国怀俄明大学讲授的研究生课程之一.该课交叉地列为三个学院的研究生课程,选课的研究来自十多个院系的各种学科,比如水文学、地质学、土壤科学、流域管理、农业生态学、植物学、生物学、昆虫学、石油工程、土木工程、地理学、地理信息系统、遥感、气象学、自然资源与环境、统计学、人类学等等.同时,作者在空间变异理论及地质统计学的应用方面进行了十多年的研究工作,在国际科学刊物上发表了 30 多篇学术论文.本书就是在作者多年教学和科研的基础上完成的,既是一本教材,可供涉及时空变量的相关领域的高年级大学生或研究生的教学之用,又是一本专著.因书中包括了作者在地质统计学在环境科学中的应用方面多年的研究成果,可用作相关读者进行研究的参考书.本书的姊妹篇是用英文写作的,两书可供双语教学之用.本书除了包括课堂教学的内容外,还提供教学用计算机程序以帮助学生实践地质统计学的基本概念及其应用.

全书共包括 9 章:第 1 章复习统计学的一些基本方法;第 2 章讨论随机函数和变异函数;第 3 章作为全书的重点,讨论了各种不同的克里金估计方法;第 4 章介绍协克里金法并给出研究实例;第 5 章用研究实例比较详细地介绍拟协克里金法;

第 6 章讨论估值的不确定性和采样策略;第 7 章讨论随机场产生的方法以及变量的随机模拟;第 8 章讨论多孔介质的空间变异性及尺度效应;第 9 章包括上机练习材料和课外作业练习.

全书的出版得到武汉大学“水资源与水电工程科学国家重点实验室”的重点资助,同时得到两项国家自然科学基金(编号:50279038 和 50239090)的部分资助,一并致谢.

张仁铎

2004 年 9 月

目 录

前言

第 1 章 统计学方法简述	1
1.1 单变量	1
1.1.1 频率表和直方图	1
1.1.2 累积频率表和直方图	2
1.1.3 正态和对数正态分布	3
1.1.4 常用的统计量	4
1.2 双变量	7
1.2.1 两个变量分布的比较	7
1.2.2 相关性	9
1.2.3 线性回归	10
1.2.4 双变量正态分布	11
第 2 章 随机函数和变异函数	13
2.1 随机函数	13
2.1.1 统计学与地质统计学的概念	13
2.1.2 空间变异结构	13
2.1.3 平稳性和二阶平稳	14
2.2 变异函数	14
2.2.1 固有假设	14
2.2.2 有效的变异函数模型	15
2.2.3 样本变异函数	18
2.2.4 变异函数的方向性和变程	18
2.3 交叉证实法	19
2.3.1 选择变异函数模型的实用原则	19
2.3.2 交叉证实法的步骤	20
2.3.3 测量值对变异函数的影响	22

第3章 克里金估计方法	23
3.1 概述	23
3.2 简单克里金法(SK)	23
3.3 普通克里金法(OK)	27
3.4 克里金法的特性	36
3.5 块克里金法(BK)	37
3.6 对数正态克里金法(LK)	39
3.7 指示性克里金法(IK)	39
3.8 泛克里金法(UK)	40
3.8.1 主要定义	40
3.8.2 漂移的估计	43
3.8.3 泛克里金法点的估计值	44
3.8.4 泛克里金法块上的估计值	45
3.8.5 应用泛克里金法的例子	46
3.9 离析克里金法(DK)	49
3.9.1 离析克里金法	50
3.9.2 农田土壤中 EC 的管理方案	51
第4章 协克里金法	57
4.1 概述	57
4.2 协克里金法方程	57
4.3 协变异函数的计算	64
4.4 用近红外反射率来提高土壤质地的预测精度	65
4.4.1 变异函数和协变异函数	65
4.4.2 结果	67
4.5 协克里金法在土壤和植物的微量元素估计中的应用	69
4.5.1 资料数据	71
4.5.2 统计分析	71
4.5.3 地质统计学分析	72
第5章 拟协克里金法	75
5.1 拟协变异函数	75
5.2 拟协克里金方程式	76
5.3 运用对称的拟协变异函数来估计土壤化学物质的空间分布	78
5.3.1 基本资料	78

5.3.2 硝酸盐和电传导度之间的协克里金法估值	79
5.3.3 钙和电传导度之间的协克里金法估值	81
5.4 运用不对称的拟协变函数来提高土壤溶质的估值精度	84
5.4.1 材料和方法	84
5.4.2 结果与讨论	86
第 6 章 估值的不确定性和采样策略	89
6.1 估值的不确定性	89
6.2 采样设计策略	90
6.2.1 样本设计	91
6.2.2 判断采样法	93
6.2.3 简单随机采样法	93
6.2.4 分层(区)随机采样法	94
6.2.5 系统采样法	95
6.2.6 分区系统非准直采样法	95
6.2.7 有目标的或直接的采样法	96
6.2.8 交替采样法	96
6.2.9 地质统计采样法	97
6.3 样本形状的影响和最佳样块大小	98
6.3.1 概述	98
6.3.2 非均匀性指数	98
6.3.3 样块形状的影响	100
6.3.4 最佳样块大小	101
第 7 章 随机变量的模拟	105
7.1 概述	105
7.2 随机场的模拟	105
7.2.1 转动带法	105
7.2.2 条件模拟	107
7.2.3 随机过程模拟	107
7.3 随机偏微分方程的迭代解法	109
7.3.1 理论基础	110
7.3.2 数值结果	112
7.3.3 应用	114
7.4 吸附溶质在非饱和土壤中运移的随机分析	117

7.4.1	概述	117
7.4.2	水流运动和溶质运移问题的描述	118
7.4.3	土壤水力特性和化学吸附特性的空间变异性	119
7.4.4	矩分析方法	120
7.4.5	模拟的设置	121
7.4.6	浓度分布的空间矩和宏观弥散度	122
7.4.7	土壤含水量对溶质扩散的影响	123
7.4.8	吸附系数对溶质扩散的影响	124
7.4.9	相关结构对溶质扩散的影响	124
7.5	对重油污染土壤修复的随机模拟	125
7.5.1	数值模拟的设置	126
7.5.2	土壤渗透度的随机场	126
7.5.3	重油恢复以及饱和度的平均值计算	127
7.5.4	平均饱和度在各相间的时间分布	127
7.5.5	重油浓度的时空分布规律	129
7.5.6	治理时间	129
第 8 章	空间变异性及尺度效应	132
8.1	样本方差随尺度的变化关系	132
8.1.1	概述	132
8.1.2	理论	132
8.1.3	数值结果	136
8.1.4	地质统计学参数和非均匀性指数之间的关系	140
8.1.5	方差与采样区域尺度的关系	140
8.2	随尺度变化的土壤水力传导度	141
8.2.1	概述	141
8.2.2	理论	141
8.2.3	结果	143
第 9 章	上机练习材料和课外作业	145
9.1	程序和文件	145
9.2	输入数据文件	145
9.3	基本统计量	146
9.4	样本变异函数的计算	148
9.5	交叉证实法	155

9.6 克里金法	160
9.7 协克里金法	167
9.8 上机练习题和课外作业题	169
9.8.1 课外作业 1	169
9.8.2 课外作业 2	170
9.8.3 课外作业 3	171
9.8.4 课外作业 4	171
9.8.5 上机练习 1	172
9.8.6 上机练习 2 (样本变异函数值计算)	172
9.8.7 上机练习 3 (样本变异函数值计算)	172
9.8.8 上机练习 4 (交叉证实)	172
9.8.9 上机练习 5 (点克里金法)	173
9.8.10 上机练习 6 (块克里金法)	173
9.8.11 上机练习 7 (协变异函数模型的计算)	173
9.8.12 上机练习 8 (协克里金法)	174
9.8.13 学期项目报告	174
9.8.14 上机练习小结	175
参考文献	176
附录 地质统计学常用词汇(中英文对照)	186

第 1 章 统计学方法简述

1.1 单变量

这一章主要是帮助读者复习统计学的一些基本知识和方法,先讨论一个变量(单变量)的统计方法,然后讨论两个变量(双变量)间的关系.

1.1.1 频率表和直方图

为了表述方便,我们以图 1.1 中的数据为例来说明一些概念,图 1.1 表示在 $10\text{m} \times 10\text{m}$ 网格中的变量 Z_1 (化合物浓度)的 100 个测量值.一种最常用的表达数

81 +	77 +	103 +	112 +	123 +	19 +	40 +	111 +	114 +	120 +
82 +	61 +	110 +	121 +	119 +	77 +	52 +	111 +	117 +	124 +
82 +	74 +	97 +	105 +	112 +	91 +	73 +	115 +	118 +	129 +
88 +	70 +	103 +	111 +	122 +	64 +	84 +	105 +	113 +	123 +
89 +	88 +	94 +	110 +	116 +	108 +	73 +	107 +	118 +	127 +
77 +	82 +	86 +	101 +	109 +	113 +	79 +	102 +	120 +	121 +
74 +	80 +	85 +	90 +	97 +	101 +	96 +	72 +	128 +	130 +
75 +	80 +	83 +	87 +	94 +	99 +	95 +	48 +	139 +	145 +
77 +	84 +	74 +	108 +	121 +	143 +	91 +	52 +	136 +	144 +
87 +	100 +	47 +	111 +	124 +	109 +	0 +	98 +	134 +	144 +

图 1.1 在 $10\text{ m} \times 10\text{ m}$ 网格中的变量 Z_1 的 100 个
数据值(单位为 $\text{ppm}^{①}$)

① $1\text{ppm} = 1 \times 10^{-6}$

据的统计方式就是使用频率表和与之相关的直方图,频率表记录落于某些区间的频率.一般来说,设随机事件 A 在 n 次试验中出现 n_A 次,那么比值

$$f_n(A) = \frac{n_A}{n} \quad (1.1)$$

叫做事件 A 在这 n 次试验中出现的频率.表 1.1 总结了图 1.1 中 100 个 Z_1 值的在各取值区间的频率分布,表 1.1 中结果也可以用直方图来表示,如图 1.2 所示.

表 1.1 变量 Z_1 值的频率(区间宽度为 10 ppm)

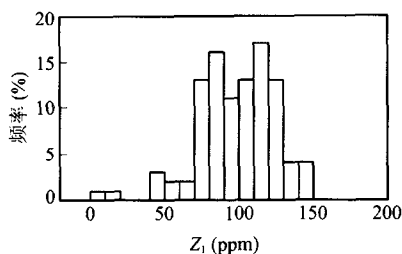
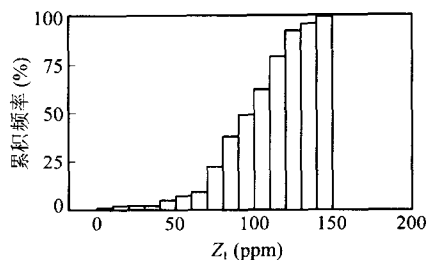
区 间 (ppm)	数 据 个 数	百 分 数 (%)
$0 \leq Z_1 < 10$	1	1
$10 \leq Z_1 < 20$	1	1
$20 \leq Z_1 < 30$	0	0
$30 \leq Z_1 < 40$	0	0
$40 \leq Z_1 < 50$	3	3
$50 \leq Z_1 < 60$	2	2
$60 \leq Z_1 < 70$	2	2
$70 \leq Z_1 < 80$	13	13
$80 \leq Z_1 < 90$	16	16
$90 \leq Z_1 < 100$	11	11
$100 \leq Z_1 < 110$	13	13
$110 \leq Z_1 < 120$	17	17
$120 \leq Z_1 < 130$	13	13
$130 \leq Z_1 < 140$	4	4
$140 \leq Z_1 < \infty$	4	4

1.1.2 累积频率表和直方图

绝大多数的统计学书里都按上升的区间值来计算累积频率,但在一些实际应用中,比如研究矿藏储量和环境污染分布时,人们对高于某一下限值(阈值)的累积频率更感兴趣,在这种情况下,应按递减的区间来计算累积频率.

表 1.2 是在表 1.1 的基础上完成的,这里不是记录在某些区间的数值个数,而是记录了低于某些截断值的总的数值个数,及其与总的数据个数的比值,相应的累积频率是数值在 0 和 100 % 之间的一个不减函数.累积频率分布也可以用直方图来表示,如图 1.3 所示.

如果随机变量 X 是连续变化的(即频率区间趋于 0),由频率和累积频率分布就可分别引入概率密度和概率分布函数的概念,并有以下关系:

图 1.2 100 个 Z_1 值的频率直方图图 1.3 100 个 Z_1 值的累积频率直方图

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f(u) du \quad (1.2)$$

式中 $F_X(x)$ 是随机变量 X 低于或等于截断值 x 的累积频率, P 是概率分布, $f(x)$ 是概率密度函数。

表 1.2 变量 Z_1 值的累积频率(区间宽度为 10 ppm)

区 间 (ppm)	数据个数	百 分 数 (%)
$Z_1 < 10$	1	1
$Z_1 < 20$	2	2
$Z_1 < 30$	2	2
$Z_1 < 40$	2	2
$Z_1 < 50$	5	5
$Z_1 < 60$	9	9
$Z_1 < 70$	7	7
$Z_1 < 80$	22	22
$Z_1 < 90$	38	38
$Z_1 < 100$	49	49
$Z_1 < 110$	62	62
$Z_1 < 120$	79	79
$Z_1 < 130$	90	90
$Z_1 < 140$	96	96
$Z_1 < \infty$	100	100

1.1.3 正态和对数正态分布

如果某一变量是正态(高斯)分布,那么会对这一变量的理论分析和估计方法的处理带来很多便利.因此,对于收集到的数据,我们往往希望知道它们的分布与

正态分布有多接近,正态分布图就是用来解决这一问题的.将累积频率点在正态分布图纸上,如数据是正态分布的,那么累积频率图就应是一条直线.正态分布的概率密度函数用以下的数学方式来表达:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (1.3)$$

式中 σ^2 是方差, μ 是均值, $f(x)$ 是一条以均值为中心的对称的钟形曲线.

在实际应用中,许多变量并不接近正态分布,通常情况下,它们的样本值有许多很小的值和一些很大的值.虽然正态分布不适合用来表征这种不对称分布,但与之相联系的另一种分布——对数正态分布有时却是一种较好的选择.如果将变量进行对数变换后的新变量是正态分布,这种变量就称为对数正态分布.即如变量 $Y = \ln(X)$ 是正态分布, X 就是对数正态分布,概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \quad (1.4)$$

式中 σ^2 是 Y 的方差, μ 是 Y 的均值. Y 和 X 的均值和方差有如下的关系:

$$X \text{ 的均值 } E(X) = \exp(2\mu + \sigma^2) \quad (1.5)$$

$$X \text{ 的方差 } \text{Var}(X) = \exp(2\mu + \sigma^2) [\exp(\sigma^2) - 1] \quad (1.6)$$

应当注意,对样本值的分布选择一个理论模型并不是进行变量估计的必要步骤,也就是说,样本显示的正态性不一定能保证好的估计值,反之亦然.有的方法更多地依赖于这种正态性,有的却不然.甚至对于一些在正态性假定的基础上建立起来的方法,即使在样本资料显示非正态分布情况下,也可能是有用的.

1.1.4 常用的统计量

频率分布(直方图)的重要特征可以通过一些样本数据资料的统计量来表征,这些统计量可归为三类:对样本数据分布位置的量度,对样本数据分布分散情况的量度和对样本数据分布形状的量度.第一类统计量给出变量分布的各部分所在的位置,比如均值、中值和众数给出了分布的中心所在,分布的其他位置由各种四分位数来表征;第二类统计量包括方差和标准方差等,主要用来描述数据值的分散度,这些统计值越大,表示数据越分散;对形状量度的统计量包括倾斜系数和变异系数等,倾斜系数用来描述分布的对称性,而变异系数则用来描述一些分布的尾部长度.

1.1.4.1 位置的量度

均值:均值 m 就是样本数据值的算术平均

$$m = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.7)$$

式中 n 是数据的个数, $x_1, x_2, \dots, x_n$ 是数据值. 图 1.1 中的 100 个 Z_1 值的均值为 97.55 ppm.

中值: 如果将观测值按递增的顺序排列, 中值 M 就是这些数据的中心点, 一半的观测值在中值之下, 一半在中值之上. 将数据排列为 $x_1 \leq x_2 \leq \dots \leq x_n$, 那么中值就可以用下式计算:

$$M = \begin{cases} x_{\frac{n+1}{2}}, & \text{如 } n \text{ 是奇数} \\ (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2, & \text{如 } n \text{ 是偶数} \end{cases} \quad (1.8)$$

均值和中值都用来量度分布的中心位置, 但均值对于一些特别高的数据值很敏感, 比如图 1.1 中 145 ppm 如变成 1450 ppm, 均值就变成了 110.60 ppm, 但中值却不受这一变化的影响, 因为中值只取决于多少数据高于或低于它, 而不取决于高多少或低多少. 图 1.1 中 100 个 Z_1 值的中值为 100.50 ppm.

众数: 众数是最常出现的观测值. 直方图中频率最高所对应的数据区间就包含了众数值, 在图 1.1 中, 110 至 120 ppm 区间频率最高. 在此区间, 111 ppm 比其他数据出现的次数都多, 所以是众数.

极小值: 数据中的最小值称为极小值, 图 1.1 中数据的极小值为 0.

极大值: 数据中的最大值称为极大值. 图 1.1 中数据的极大值为 145 ppm.

第 1 和第 3 四分位数: 与中值将数据系列分成两部分相似, 四分位数将数据系列分为四个部分, 如果数据以递增顺序排列, 那么四分之一的数据低于第 1 四分位数 (Q_1), 四分之一的数据高于第 3 四分位数 (Q_3).

与中值相似(对应于 50% 的概率,)四分位数很容易从概率曲线上读出, 对应于累积概率 25% 和 75% 的 x 轴坐标的读数就分别是 Q_1 和 Q_3 的值. 图 1.1 中 100 个 Z_1 数据的 $Q_1 = 81.25$ ppm, $Q_3 = 116.25$ ppm.

1.1.4.2 分散的量度

方差: 方差 σ^2 用以下公式计算:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 \quad (1.9)$$

式中 m 是均值, 因方差涉及到观测值与均值差的平方和, 所以对于特别高的值很敏感. 图 1.1 中数据的方差为 668 ppm².

标准方差: 标准方差 σ 为方差的平方根值. 图 1.1 中数据的标准方差值为 26.23 ppm.

四分位数全距: 四分位数全距也可用来量度测量值的分散情况, 四分位数全距

是上、下两个四分位数之差:

$$IQR = Q_3 - Q_1 \quad (1.10)$$

不像方差或标准方差,四分位数全距不用均值作为分布中心,因此在有一些特别高的数据值的情况下很有用.图 1.1 中数据的四分位数全距值为 35.00 ppm.

1.1.4.3 形状的量度

倾斜系数:最常用用来表征分布对称性的统计量为倾斜系数,用以下式来进行计算:

$$C_s = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m)^3}{\sigma^3} \quad (1.11)$$

式中 m 和 σ 分别为均值和标准方差. C_s 对特别高的值尤为敏感,因此通常不是用 C_s 的值而是用其正负号来描述分布的对称性.正的 C_s 值表示直方图的分布向右方(高的数据值方向)倾斜,即拖一个长的尾巴,在此情况下,中值比均值小;而负值则表示直方图向左方(低的数据值方向)倾斜,在此情况下,中值大于均值.如 C_s 接近于 0,则直方图大致对称,中值与均值相近.图 1.1 中 100 个 Z_1 值的 C_s 值比较接近于 0(-0.779),说明分布只是略为不对称.

变异系数:变异系数 C_v 是描述分布形状的另一方式,主要用于数据值都是正值及其 C_s 也为正值的情况下, C_v 定义为:

$$C_v = 100\% \sigma/m \quad (1.12)$$

在进行变量估计的情况下, C_v 对于估计的结果可提供一些预警的信息, C_v 值大于 100% 表示存在一些特别大的样本值,它们对变量的估计有很大的影响.图 1.1 中数据的变异系数为 26.9%,不是很大,反映在直方图高数值的部分没有拖一个长尾巴.

在土壤科学中,根据 C_v 值可对土壤性质的变异程度进行分类(Wilding, 1985): C_v 值在 0~15% 为小变异, 16%~35% 为中等变异, 大于 36% 为高度变异.一些典型土壤性质的 C_v 值列于表 1.3 中(Jury, 1986; Jury et al., 1987; Beven et al., 1993; Wollenhaupt et al., 1997), pH 和土壤孔隙率属于小变异变量,而与土壤水和溶质运移有关的变量则多具有高度变异性.

例 1.1 在一块 100 hm² 的田块中,黏粒含量的均值和标准方差分别是 25% 和 5%,而饱和和传导度的均值和标准方差分别是 1 cm/ha² 和 1.5 cm/ha²,哪一种土壤性质具有更大的变异性?

表 1.3 一些土壤和作物性质的变异系数(C_V)范围(Mulla and McBratney, 2002)

性 质	C_V (%)	变异性
pH	2~15	低
孔隙率	7~11	低
土壤容重	3~26	低至中
作物产量	8~29	低至中
砂粒含量(%)	3~27	低至中
0.01MPa的含水量	4~20	低至中
农药吸附系数	12~31	中
有机质含量	21~41	中至高
1.5MPa的含水量	14~45	中至高
黏粒含量	16~53	中至高
土壤硝态氮(NO_3^- -N)	28~58	中至高
土壤水渗透率	23~97	中至高
土壤中可利用钾	39~157	高
土壤中可利用磷	39~157	高
土壤电传导度	91~263	高
饱和水力传导度	48~352	高
饱和溶质运移速度	78~194	高
溶质弥散系数	79~178	高
溶质弥散度	78~539	高

1.2 双变量

上节讨论的方法可以用于多变量中的每一个单变量,然而,这会限制我们对于多变量的分析,许多时候,我们需要了解变量之间的相互关系和依赖性,因为这些是变量的重要和有用的特性.在上一节图 1.1 的基础上,图 1.4 显示了在同一位置上两个变量(Z_1 和 Z_2)的数据值,在这一节中,我们将以这两个变量为例来讨论描述双变量关系的一些方法.

1.2.1 两个变量分布的比较

在多变量分析中,我们往往需要比较它们的分布情况.直方图和一些统计量可以揭示变量间一些大的差别.变量 Z_1 和 Z_2 的直方图和统计量(表 1.4)显示了它们存在着一些主要差别: Z_1 的分布是负倾斜,而 Z_2 的分布是正倾斜,同时 Z_1 的值一般比 Z_2 的值大, Z_1 的均值比 Z_2 的均值大 5 倍以上, Z_1 的中值和标准方差也比 Z_2 的大.

81	77	103	112	123	19	40	111	114	120
+	+	+	+	+	+	+	+	+	+
15	12	24	27	30	2	18	18	18	18
82	61	110	121	119	77	52	111	117	124
+	+	+	+	+	+	+	+	+	+
16	7	34	36	29	7	4	18	18	20
82	74	97	105	112	91	73	115	118	129
+	+	+	+	+	+	+	+	+	+
15	9	22	24	25	10	7	19	19	22
88	70	103	111	122	64	84	105	113	123
+	+	+	+	+	+	+	+	+	+
21	8	27	27	32	4	10	15	17	19
89	88	94	110	116	108	73	107	118	127
+	+	+	+	+	+	+	+	+	+
21	18	20	27	29	19	7	16	19	22
77	82	86	101	109	113	79	102	120	121
+	+	+	+	+	+	+	+	+	+
15	16	16	23	24	25	7	15	21	20
74	80	85	90	97	101	96	72	128	130
+	+	+	+	+	+	+	+	+	+
14	15	15	16	17	18	14	6	28	25
75	80	83	87	94	99	95	48	139	145
+	+	+	+	+	+	+	+	+	+
14	15	15	15	16	17	13	2	40	38
77	84	74	108	121	143	91	52	136	144
+	+	+	+	+	+	+	+	+	+
16	17	11	29	37	55	11	3	34	35
87	100	47	111	124	109	0	98	134	144
+	+	+	+	+	+	+	+	+	+
22	28	4	32	38	20	0	14	31	34

图 1.4 在同一位置上两个变量(Z_1 和 Z_2)的数据值表 1.4 图 1.2 中 Z_1 和 Z_2 的主要统计量

统计量	Z_1	Z_2
n	100	100
m	97.6	19.1
σ	26.2	9.81
C_V	27	51
Min	0.0	0.0
Q_1	81.3	14.0
M	100.5	18.0
Q_3	116.8	25.0
Max	145.0	55.0

显示双变量资料的一种最常用的方式就是画出数据点的散布图,这是一种 x - y 图,即 x 坐标对应一个变量的数据值, y 坐标对应于同一位置上另一个变量的数据值.图 1.5(a)显示 Z_1 和 Z_2 的数据点散布图,虽然数据有一些离散,但总的趋势是 Z_1 高的值与 Z_2 高的值相对应,低的 Z_1 值与低的 Z_2 值相对应.

另外散布图还可以使我们对一些异常的数据点引起注意.在进行空间连续变量分析的早期阶段,有必要检查和清整数据,因为成功的估计方法取决于可信的原始数据.即使清理了数据,一些异常的数值也可能对变量的估计产生主要影响,散