● 高等学校试用教材

● 卢崇飞　高惠璇　叶文虎　编著

环境数理统计学应用及程序

● 高等教育出版社

高等学校试用教材

环境数理统计学应用及程序

卢崇飞　高惠璇　叶文虎 编著

高等教育出版社

内 容 提 要

《环境数理统计学应用及程序》一书共有五章正文二个附篇。正文详细阐述了多元分析方法的原理及其在环境保护实际工作中的应用实例，并对每个方法都给出了用 BASIC 语言编写的计算机程序。附篇(一)由浅入深地介绍了矩阵、概率论与数理统计的基本概念和基本原理，是学习正文的必要的数学基础。附篇(二)是与正文相配的全套用 FORTRAN 语言编写的计算机程序。本书中给出的程序全部可以在 IBM PC-XT 机或其兼容机上使用。

本书既可作为大学本科教材，亦是有关专业的教师、研究生和实际工作者的参考书。

高等学校试用教材

环境数理统计学应用及程序

卢崇飞　高惠璇　叶文虎　编著

高等教育出版社出版

新华书店上海发行所发行

吴江伟业印刷厂印装

开本 787×1092　1/16　印张 28.75　字数 652,000

1988 年 6 月第 1 版　1988 年 6 月第 1 次印刷

印数 0001—2,200

ISBN 7-04-000922-6/K·41

定价 5.80 元

序

概率统计的出现是人类认识史上一次飞跃

人类对大量偶然现象存在统计规律性的认识，始于中华，远在公元前2000年，已发现大量新生婴儿中，两性存在稳定的比例，约各占1/2。至十八世纪，英国女皇安娜的御医，皇家学会会员阿尔布特诺特研究了伦敦1629年—1710年出生的的婴儿，发现男女比例精确地维持平衡，并以此作为神的存在的论据。他的同事德尔享则进一步提出"男孩的比例略微超过女孩，是因上帝为了男人可能遭到一些危险(如战争、航海等)而制定的平稳补充"。古典概率的奠基人柏努里与拉普拉斯否定了这种看法，认为是自然法则。某个家庭生男或生女带有偶然性（当然也可从生理上去探究其必然的原因)，但大量的家庭却生出了不以人的意志为转移的男女各半的必然性。这是通过每次或男或女的偶然性表现出来的，这种统计规律性所表现的必然性，已不是单个家庭的属性，而是在另一个层次上即人类的属性。

在压力较大的氧气瓶中，氧分子不停地运动，互相碰撞并冲击器壁，某分子在某时刻碰壁是偶然的，但大量分子同时存在所表现出的压力稳定性，已不是单个分子的属性，而是在另一个层次上即大量分子总体运动所具有的统计规律性。再如，在一定的生产条件下，抽查一个产品，它是正品还是废品带偶然性，但大量抽查则会发现废品数具有稳定的百分比，这是大量产品所具有的统计规律性。这是在另一个层次上即整个生产水平的表现。

上述"另一个层次"实质上是在高一级层次上体现必然与偶然的辩证关系。偶然性根源于宇宙间相互联系互相制约的无限多样性。例如某时某地某点SO_2浓度，它从污染源出发，经大气搬运，湍流扩散，途中不知经过多少不同尺度漩涡的携带，也不知与其他分子碰撞了多少万万次，其全部原因是不可穷尽的，它是否落入取样器是偶然的。但是它在运动过程中的某一步，又都符合必然的力学规律。这些无穷无尽的必然交叉，就构成了事先无法预断的偶然性。跟踪追究每个分子在每个瞬时的必然运动规律，是无法实现的繁琐哲学。但是SO_2的污染受风速、风向及地面粗糙度的影响，实际大量监测所得到的数据是整体的统计规律，还是可信的。

拉普拉斯曾说过："假如有一个无所不知的大天才，能够写出适合于宇宙在给定时刻状态的所有方程，并且把它积分出来，那么他就能够完全准确地预见宇宙在无穷时间过程的全部演化"。这是属于机械论的自然观，以为了解了每个局部也就等于了解了整体，而没有认识到在高一级的层次上有新质。正如人并不等于人身上所有原子的总和，人有思想，思想已不是单个原子的属性。但也难怪拉普拉斯这位古典概率的奠基人，由于时代的局限性，他研究了统计规律性，却不认识其重大意义，因为古典概率形成的时代，正是经典力学走向全盛的时代，强调初始条件决定了整个运动。当时整个自然科学界正被机械论所统治。到19、20世纪，机械论的自

然观被打破了。动摇机械论最有力的科学进展，发生在这一时期的物理学中，玻尔兹曼和吉布斯把统计规律引入了物理学，将气体分子运动论与热力学规律联系起来，建立了统计力学。玻尔兹曼从理论上证明了不可逆方向的熵函数和状态概率的自然对数成正比，揭露了热力学运动和机械运动的本质差别。“层次”的概念引入了物理学。这是在一门科学领域内突破机械论的束缚，运用总体的辩证的认识方法。它揭示出对于组成系统的大量分子的任何一个行为作出详细推断是不可能的，但大量分子的同时存在却表现出系统总体性质十分确定的规律。对此，传统的方法已无能为力，必须引进概率统计方法。因此很多自然科学家称之为“一种可贵的进步倾向”。控制论的创始人维纳甚至认为：“必须把20世纪物理学的第一次大革命归功于吉布斯，而不是归功于爱因斯坦、海森柏或普朗克”。维纳的提法未免过分，但他对概率统计在科学认识史上的地位之评价还是中肯的。他说，概率不仅作为物理学的数学工具，而且作为物理学的部分经纬被人们接受下来了。

概率论与数理统计，作为一门数学，就是从量的方面研究必然与偶然的关系，研究大量偶然现象的统计规律性。它从17世纪中叶起步，到现在，诸分支发展极其迅速，可以说不仅作为一门数学工具，而且作为部分经纬广泛渗透到自然科学、社会科学和工农业生产的众多领域。特别是在新兴的环境科学中，更是前途广阔，大有用武之地。

环境科学是一门极为重要的综合性与实践性都很强的横向科学，它涉及到几乎我们社会生活的各个方面，传统科学（自然科学与社会科学）的各个领域，而且偶然因素众多，大气污染受排放源、风速、风向、及地面粗糙度的影响。水污染受污染源及水体物理化学生物学特性的影响等等。环境科学随机性很强，环境科学本身就是多种学科综合的新的“层次”。由于人类干预物质循环，特别是大规模现代工业的发展，引起物质分布的改变，造成有害物质的积聚即污染，导致原有生态系统的急剧变化甚至崩溃。为此，人类又发展了环境科学来进行反报复，这是一个作用，反作用、反反作用的极其复杂的过程，不仅涉及自然科学的众多领域，而且也是社会科学问题，因此环境科学所研究的不仅是化学，物理、生物等等学科的特定主题，而且是新的层次上的大系统。例如我们研究污染源之来源、形成、扩散、分布、变化转换与归宿，进行预测预报与环境质量评价，制定监测与控制措施，检验治理效果等等，从数量角度看，就是要研究环境变量之间的关系，寻求其规律，据此施加人的干预。而所谓环境变量，诸如大气中的SO_2、NO_x、CO、TCH、飘尘等等的浓度；水土中的甲基汞与镉的含量，虫鱼生物之数量；噪声分贝数；汽车通行量与尾气量；职工发病数以及气象因子等等，多因时因地而异，受多种因素的影响，其必然规律是通过偶然性来表现的，数学上称之为随机变量。环境科学涉及多种随机变量，是一个有输入输出及反馈的复杂系统。要想攀登这个领域的高峰，除去各门知识的研究探讨外，还需要研究随机变量（自身及相互关系）统计规律性的数学分支——概率论与数理统计，它与环境科学结合，姑且命名为环境数理统计学。

卢崇飞同志一直热心于推广数理统计学在各个领域中的应用，为数理统计学的普及作出了很大的贡献。近十年来，他更全力以赴地致力于数理统计学在环境保护工作中的应用，本书是他多年实践的体会与总结。

应用，是一个十分艰辛的创造性的劳动，决不是什么有贡献无水平的工作。应用，是一个把科学技术转化为生产力的工作，在应用过程中不但要对原学科的内容融会贯通，而且要对所应用的领域有相当深入的了解和熟悉，不但要熟悉其主要内容和思维方法，而且更要能正确地把握住该课题的实质。

遗憾的是在本书的校订加工过程中，卢崇飞同志因病住院，未能参加讨论，所以这项工作是由叶文虎在高惠璇，马小明二同志的全力协助下完成的。在校订中，我们改正了一些文字，调整了一些章节，补写和改写了一些章节。特别是为了使广大读者能迅速地把书中所介绍的各种方法运用于实际，高惠璇同志在校订书中的BASIC程序的同时，还写出了全部相应的FORTRAN程序。本书全部BASIC程序均在Apple机上实现；FORTRAN程序均在IBMPC/ZT机上实现。

本书介绍的主要内容是多元统计分析，研究随机变量的统计规律性。其中包括：逐步回归法，多对多双重筛选逐步回归法，定性因子的逐步判别法，多指标环境质量分类法（分类之后亦可用0，1赋值法进入回归式与判别式），因子分析法（它研究潜在因子），时间序列分析法（随着监测技术的提高，被越来越广泛的应用），为学习以上内容打基础的基础知识以及书中介绍的各种方法的BASIC和FORTRAN算法程序。

由于我们水平有限，书中可能出现一些错误，请读者指正。

编　者　1987

目　录

第一章　环境变量关系式的建立法

第二章 环境质量类别的定性判别分析法

第三章 多指标环境质量分类法

第四章 因 子 分 析

第五章 时间序列分析法

附　篇　(一)

矩阵、概率论及数理统计基础

附 篇 (二)

环境统计方法 FORTRAN 程序包

第一章　环境变量关系式的建立法

为简便起见，姑且将环境科学中有关的变量统称之为环境变量。本章介绍根据实测数据建立环境变量之间关系式的一些数学方法。

任何学科的研究，都要揭示客观世界内在的本质联系，当然不能满足于定性的关系，应尽可能建立定量关系。这种关系常用函数形式表现，如 $y=f(x_1, x_2, \cdots, x_m)$；其中 x_1，x_2，$\cdots$，x_m 称为自变量，y 称为因变量。例如，y 是某点 SO_2 的浓度，自变量是多个排放源的排放量以及多种气象因子。建立了函数关系式，即可定量分析各自变量作用的大小，算得什么情况下污染严重，什么情况下问题不大，并据此制定防治对策。如果自变量在时间上是提前量，例如建立了现在的自变量与 5 天后的 y 的关系式。则可预测预报 5 天后 SO_2 的浓度值，从而提前采取防治措施。

建立关系式通常有两类方法。一类是利用已有的物理、化学、生物等等的数量规律用逻辑推理导出；把问题化为建立微分方程并求解，比如用于大气与水中的湍流扩散方程等。第二类方法是根据实测数据来建立经验公式，或称为拟合关系式。如果由物理、化学等专业知识，已经给出函数的形式类型，只是其中若干参数待定，则第二类方法是根据实测数据估出参数，从而具体确定函数关系。如果给不出函数的形式类型。则第二类方法只能由实测数据来拼凑某种意义下最佳的函数关系式。这种拼凑归根结蒂也是设法先给出函数的形式类型然后再定参数。本章介绍第二类方法，即使用最小二乘法原则，使拟合的函数值与实测值之差的平方和达极小。这在数学上称为回归分析法。本章介绍：(1)一般回归分析法。(2)逐步回归分析法。此法借助电子计算机做大量的运算和统计检验，来筛选自变量。从而也起到构造函数形式的作用。(3)多对多双重筛选逐步回归法。此法属多元统计分析，因变量是多指标的向量，逐步筛选自变量，并将多个因变量分类构造关系式。(4)数量化理论(I)。此法是将自变量中加入定性的因子。如气象因子中的晴、阴、雨；防治措施甲、乙、丙、丁；地区分类 A、B、C、D…。至于因变量是定性指标时，如污染分为级别，人发病或不发病等。这类模型属判别分析，将在第二章介绍。回归分析法特别是逐步回归法在多种领域有广泛的应用，在环境科学中大有用武之地。(5)已知函数形式是非线性的。其中若干参数待定。用 Marquardt 方法反复迭代估参数。

§1　一元线性回归分析法

(一) 一元线性回归模型参数的最小二乘法估计

这是最简单的模型。设

$$y=a+bx+\varepsilon$$

其中 a, b 是待估参数，x 是自变量，y 是因变量，ε 是随机误差，不妨设 ε 的数学期望为零，即 $Ey=E(a+bx+\varepsilon)=a+bx$。对自变量及与其对应的因变量进行 n 次独立观测，实测数据为

$$x_1,\ x_2,\ x_3,\ \cdots x_n$$

$$y_1,\ y_2,\ y_3,\ \cdots y_n$$

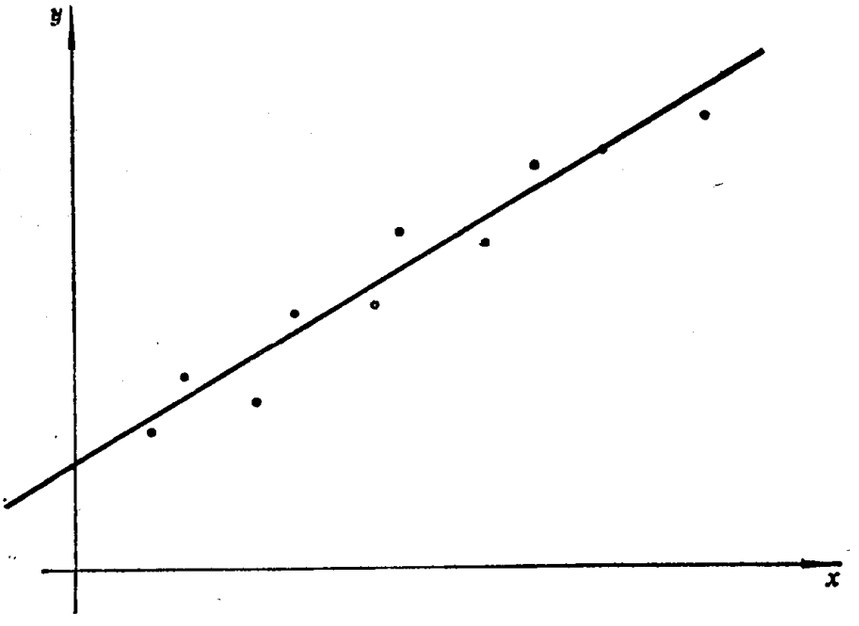

图 (1.1) 实测数据点图

通常先将实测数据在直角坐标系内点图，如图(1.1)。由图看出，用实测点基本上在一直线附近波动，从而采用线性模型。$y=a+bx+\varepsilon$。若点图的结果并非直线，则用其他方法处理，以后另作介绍。

现在的问题是如何估计 a 与 b。自然会想到，要如此地选择 a 与 b，使得 x_i 对应的直线上的值 $a+bx_i$ 与 x_i 对应的实测值 y_i 的误差

$$\varepsilon_i=y_i-(a+bx_i) \qquad i=1,\ 2,\ \cdots n$$

在某种意义下最小。上式亦可写为

$$y_i=a+bx_i+\varepsilon_i \qquad i=1,\ 2,\ \cdots n$$

ε_i 称为第 i 次观测的随机误差，一般不妨设$[\varepsilon_i]$相互独立，且期望 $E\varepsilon_i=0,\ i=1,\ 2,\ \cdots n$。

通常采用最小二乘法原则来估 a 与 b。即如此地选择(估计)b 与 a 使得上述误差的平方和

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2$$

达极小。于是问题化为解方程组

$$0=\frac{\partial Q}{\partial a}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]$$

$$0=\frac{\partial Q}{\partial b}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]x_i$$

此方程组称为正规方程组，其解 $\hat{a}$，$\hat{b}$ 称为 a 与 b 的最小二乘法估计量。上述方程组等价于方程组

$$na+\left(\sum_{i=1}^{n}x_i\right)b=\sum_{i=1}^{n}y_i \tag{1.1}$$

$$\left(\sum_{i=1}^{n}x_i\right)a+\left(\sum_{i=1}^{n}x_i^2\right)b=\sum_{i=1}^{n}x_i\,y_i \tag{1.2}$$

这是关于 a 与 b 的简单线性方程组，解得

$$\hat{a}=\bar{y}-\frac{\left[\left(\sum_{i=1}^{n}x_i\,y_i\right)-n\bar{y}\bar{x}\right]}{\sum_{i=1}^{n}x^2{}_i-n\bar{x}^2}\bar{x}$$

$$\hat{b}=\frac{\sum_{i=1}^{n}x_i\,y_i-n\bar{y}\bar{x}}{\sum_{i=1}^{n}x^2{}_i-n\bar{x}^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$。

关系式 $\hat{y}=\hat{a}+\hat{b}x$

称为回归式，有时称为预报式。即给定 x 代入上式算得 $\hat{y}$ 作为 x 所对应真值 y 的预报值。

最小二乘法原则将误差取平方，这是人为的夸张。直观上更合理的原则应使误差的绝对值相加达最小，即如此地取 a 与 b 的值，使

$$\sum_{i=1}^{n}|\varepsilon_i|=\sum_{i=1}^{n}|y_i-(a+bx_i)|$$

达极小。

亦可如此地选取 a 与 b 的值，使

$$\max_{1\leqslant i\leqslant n}|\varepsilon_i|$$

达极小。随着电子计算机的发展，后述的两种原则已有专用程序。我们所介绍的最小二乘法原则，误差取平方是人为夸张，是退步，但在某种意义下可将各自变量的作用简单分解，类似于方差分析，是“退一步，进两步”。

(二) 检验回归系数 b 是否为零

前述最小二乘法原则，对各种类型的随机误差项皆适用。若进一步假定随机误差项遵从正态分布，则可对 b 是否为零进行统计检验。即

设 $y_i=a+bx_i+\varepsilon_i \quad i=1,2,\cdots,n$

其中误差项 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 独立同分布 $N(0,\sigma^2)$，σ^2 是未知参数。要检验假设 $H_0: b=0$。

若否定 $H_0: b=0$，则认为 x 与 y 有线性关系，若不能否定 H_0，则认为 x 与 y 无线性关系。必须注意到无线性关系不等于说无其他关系。例如设 $y=\sin x+\varepsilon$，实测得 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n)，此时若用线性关系式 $y_i=a+bx_i+\varepsilon_i$，$i=1, 2, \cdots, n$ 拟合，则可检验得 $b=0$。即 x 与 y 没有线性关系但它们之间有很强的其他关系。

检验 $H_0: b=0$ 的步骤如下：

由各种原因(自变量的作用及其他一切因素的作用)引起因变量的总波动。称之为总平方和 $S_{总}$：

$$S_{总}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}+\hat{y}-\bar{y})^2=\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)(\hat{a}+\hat{b}x)-\bar{y}]^2$$

$$=\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)]^2+\sum_{i=1}^{n}(\hat{a}+\hat{b}x_i-\bar{y})^2$$

$$+2\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)][\hat{a}+\hat{b}x_i-\bar{y}]$$

其中第三项

$$2\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)][\hat{a}+\hat{b}x_i-\bar{y}]$$

$$=2(\hat{a}-\bar{y})\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)]+2\hat{b}\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)]x_i=0$$

这是因为 $\hat{a}, \hat{b}$ 是正规方程的解，上式中两个"Σ"号正好是正规方程组中等号左边的项，皆等于零。$S_{总}$ 中的第一项 $\sum_{i=1}^{n}[y_i-(\hat{a}+\hat{b}x_i)]^2$ 是实测所得 y_i 与回归式算得的 $\hat{y}_i$ 的误差平方和，此项表现了随机误差(除自变量外的一切因素)的作用，称为误差平方和或残差平方和或剩余平方和，记为 $S_{误}$ 或 $S_{残}$ 或 $S_{剩}$。$S_{总}$ 中的第二项 $\sum_{i=1}^{n}[(\hat{a}+\hat{b}x_i)-\bar{y}]^2$ 是将各自变量代入回归式算得的 n 个 $\{\hat{y}_i\}$ 与 $\bar{y}$ 的离差平方和，此项表现了自变量代入回归式在总波动中的贡献，称之为回归平方和，记为 $S_{回}$。由正规方程组中的 (1.1)式，得 $\hat{a}+\hat{b}\bar{x}=\bar{y}$，代入 $S_{回}$，得

$$S_{回}=\sum_{i=1}^{n}[(\hat{a}+\hat{b}x_i)-\bar{y}]^2=\sum_{i=1}^{n}[\bar{y}-\hat{b}\bar{x}+\hat{b}x_i-\bar{y}]^2=\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

总之，$S_{总}$ 可简单地分解为有实际意义的两项

$$S_{总}=S_{回}+S_{剩}$$

显然 $S_{回}$ 所占的比重愈大，则说明回归式愈有意义。数学上可以证明，在随机误差项遵从正态分布的条件下，有

$$\frac{S_{剩}}{\sigma^2}\sim\chi^2_{n-2}\text{—分布}$$

当 $H_0: b=0$ 假设成立时