•专论与综述•

环境监测数据分析和监测网设计中 SPSS 10.0 的应用

陆志波, 陆雍森
(同济大学环境科学与工程学院, 上海　200092)

摘　要: SPSS 10.0 是目前功能齐全、便于应用的优秀统计软件。文章介绍了该软件在环境监测数据分析和监测网设计中的应用。文中简述了该软件的功能及优点, 并通过实例, 展示其在环境监测数据分析和监测网设计中的应用, 这对于优化监测数据分析结果和监测网的设计均有较好的借鉴作用。

关键词: SPSS; 监测数据; 网络设计; 偏相关分析; 曲线拟合; 聚类分析
中图分类号: X830.3 X84　　文献标识码: A　　文章编号: 1006- 2009(2002)03- 0012- 05

Application of SPSS 10.0 to Environmental Monitoring Data Analysis and Network Design

LU Zhi-bo, LU Yong-sen
(School of Environmental Science and Engineering, Tongji University, Shanghai, 200092, China)

Abstract: SPSS10.0 is contemporary statistical software with versatile functions. The paper introduces the application of SPSS10.0 to environmental monitoring data analysis and network design. At first the functions and merits of the software are introduced, then the application of SPSS10.0 to data analysis and monitoring network design is described through three cases. SPSS10.0 is useful software for optimization of data processing and monitoring network design.

Key words: SPSS; Monitoring data; Network design; Partial correlation analysis; Curve estimation; Cluster analysis

1　环境监测与数理统计软件

1.1　环境监测的数据处理与表达

环境监测的目的是准确、及时、全面地反映环境质量现状及发展趋势, 为污染源总量控制、环境规划与管理等提供科学依据。环境质量的时空分布变化是受自然过程和人群活动影响的函数, 它们的变化规律是通过大量监测数据的偶然性变化表现出来的, 是一类随机变量。所以, 研究随机变量变化规律性的数理统计方法是环境监测工作中常用的数学工具。

环境监测, 例如日常空气质量监测、水质监测和噪声监测中取得的大量监测数据需要按用户的需求进行处理、解释和表达, 并将其转变为管理部门和广大公众了解环境质量或污染变化规律的有用信息。实现数据统计处理的计算机化, 建立在长期监测基础上的信息系统和决策支持系统是各级环境监测站和环境管理部门迫切要求解决的。传统的数据分析方法耗费的时间多, 特别是随着数据量的增加, 即使应用成熟软件如 Excel, FoxPro, FoxBase 工作量仍很大, 且用户界面可读性较差, 给非数据库专业人士的使用带来不便。SPSS 10.0 不但用户界面友好, 而且结果显示清楚, 会使用户有清新感觉。

1.2　SPSS 10.0 统计软件

SPSS 是 20 世纪 80 年代初由美国 SPSS 公司开发的目前世界上最优秀的统计分析软件之一, 已广泛应用于自然科学和其他领域, SPSS 10.0 for Windows 是其最新版本, 在环境监测中还很少应用, 特别是地方环境监测部门对它尚不了解。将这个优秀软件应用于实际工作中, 不但能够大大减少数据分析人员的工作量, 而且提高了数据处理、结果存贮和应用的准确性、实用性、可信性, 并便于上级部门应用于环境决策。

该软件主要菜单包含十大方面:

(1) File(文件) 菜单。与大多数 Windows 程序相同, 这里包含了对数据文件进行新建、打开、保存、

收稿日期: 2002- 01- 28; 修订日期: 2002- 04- 23

作者简介: 陆志波(1979—), 男, 浙江宁波人, 助教, 在读硕士生, 主要从事环境监测系统设计、环境评价与管理、GIS 应用、生活垃圾分类收运规划的研究。

页面设置、打印预览等 15 条命令。其中“Open”命令选项中的“Data”项可以打开 Excel、Access、Dbase、Lotus、SYLK、Text 等多种格式的数据文件,为充分利用原有的数据库系统提供了方便。“Cache(缓存) Data”命令允许用户将数据文件复制到缓存中,从而加快了分析运算的速度。

(2) Edit(编辑)菜单。提供了复制、剪切、全选、粘贴、特殊粘贴等编辑命令。其中的“Option”选项允许用户对系统的默认设置进行修改,使之更加适合不同用户的要求,这其中包括默认文件夹、字体、图表输出形式、结果显示位置的控制等等。

(3) View(视图)菜单。包含显示 / 隐藏的状态栏、工具栏、字体栏、网格线、属性标签 5 条命令。

(4) Data(数据)菜单。提供了日期定义、插入变量或观测量、为观测量进行排序、合并或者分割数据文件等 11 条命令。

(5) Transform(变换)菜单。允许用户对数据进行初步的整理,其中的变换菜单提供了计算公式、随机选择、加权、观测量排序、变量分类、缺失值替换、建立时间序列等 10 条命令。

(6) Analyze(分析)菜单。包括 13 个子菜单,是SPSS 10.0 的主要功能模块,分别具有如下功能:①Reports(报告),是对于观测量的基本统计分析,综合统计分析等 4 个命令;②Descriptive Statistics(描述性统计),包括频数分布、描述性和探索性统计分析、列出联表分析等 4 项命令;③Compare Means(均值比较),包括求均值、3 种 t 检验和方差分析等 5 项命令;④General Linear Model(一般线性模型),包括单变量、重复测量等 4 项命令;⑤Correlate(相关分析),包括双变量相关和偏相关等 3 项命令;⑥Regression(回归分析),包括线性回归、非线性回归和曲线拟合等 9 项命令;⑦Log linear(对数线性),包括一般对数线性分析、模型选择对数线性分析等 3 项命令;⑧Classify(聚类分析),包括 K- 均值聚类、分层聚类和判别分析等 3 项命令;⑨Data Reduction(数据简化),包含因子分析命令;⑩Scale(等级分析),包括可靠性分析和多元分级等 2 项命令;⑪ Nonparametric Tests(非参数检验),包括 χ^2 检验、2 个独立样本分析等 8 项命令;⑫ Survival(残余分析),包括生命表、柯克斯回归等 4 项命令;⑬ Multiple Response(多元响应),包括 Define Sets 等 3 个命令项。

(7) Graphs(图表)菜单。SPSS 10.0 具有很强的制图功能,远远超过 Excel 的图表功能,这些图形可以在各种统计分析过程中通过对于相应的“Plot”选项进行设置,得到需要的理想图形,也可以直接由菜单项“Graphs”图形菜单产生。生成的图形包括 Bar(条形图)、Line(线形图)、Area(面积图)、Pie(饼图)、High- Low(如极差图)、Pareto(帕雷托图)、Control(工序控制图)、Box plots(箱线图)、Error Bar(误差条形图)、Scatter(散点图)、Histogram(直方图)等,而 Times Series(时间序列)选项可以生成自相关图、偏相关图和互相关图。

(8) Utilities(公用程序)菜单。这里包括变量集的定义、菜单栏定制内容等 7 条命令;

(9) Windows(窗口)菜单。在这里用户可以最小化所有窗口,也可以实现在不同窗口中的切换。

(10) Help(帮助)菜单。包括帮助主题、使用指南、统计教练、语法向导等 6 条命令。

此外,SPSS 10.0 面向用户使用突出的优势是:①Windows 的窗口方式和界面友好的对话框;②拥有全面生动的帮助。因此 SPSS 10.0 使原来需要几十分钟乃至几个小时完成的工作可以在几分钟内轻松完成。

2 SPSS 10.0 在环境监测中的应用潜力

2.1 用于监测数据的分析

2.1.1 数据集合的参数检验与区间估计

一般环境监测站都有统一格式的数据库,可以存贮历年的各类环境监测资料供查阅、打印报表并具有有限的统计计算功能,如求均值、方差等。这类简便数据库可以与 SPSS 联合运用。在 SPSS 10.0 中可以非常方便地实现 t 检验、F 检验和置信区间的估计,具体实现的过程是选择软件中的“Analyze”——“Compare means”等选项,按照提示可以方便地实现大量数据的检验与区间估计。

2.1.2 数据集合的非参数检验

非参数检验是不依赖于总体分布的统计推断方法,是指在总体分布不服从正态分布且分布类型不明时,用来检验数据资料是否来自同一个总体假设的一类检验方法。在环境监测数据处理中,由于各种影响因素较多,碰到的非参数检验的机会也较多,在 SPSS 10.0 中可以方便地实现卡方检验(Chi-square Test),只需要选择“Analyze”菜单中的“Nonparametric Tests”即可按照提示操作。

2.1.3 方差分析

在分析环境监测数据的过程中,要求识别“处

理”和“误差”的作用, 而误差中又包括随机误差及系统误差。分析者在得到一批数据后, 应从反复的试验和观测中, 分析出哪些因素是对该结果起主导作用应予重视的。例如, 在试验一种分析方法时, 用不同的药剂不同配比, 在不同温度和(或)反应时间条件下进行多次试验, 才能确定出最佳的操作条件, 而在这个选择过程中可用方差分析及试验设计方法。如果只是验证一种实验试剂在不同反应温度下的作用, 则可采用单因素方差分析法, 在 SPSS 10.0 中的实现过程就是选择“Analyze”——“Compare Means”——“One Way ANOVA”; 如果是多种因素相互作用, 则要用到多因素方差分析, 其实现过程是选择“Analyze”——“General Linear Model”——“Univariate”。

2.1.4　相关分析

在环境监测数据分析中, 经常需了解各环境参数的相互关系。这可以在现有的各种数据的基础上, 建立一个或几个环境要素之间的相关关系, 用于环境状况的研究。在 SPSS 10.0 中, 可以进行相关分析以及偏相关分析, 用相关系数矩阵的形式描述有关参数的亲疏程度, 其实现过程是“Analyze”——“Correlate”。

2.1.5　回归分析(包括多元回归和逐步回归)

回归分析是处理环境监测数据的常用手段。由于环境数据的影响因素较多, 常会碰到多元回归问题, 使用 SPSS 10.0 可以轻松地实现多元回归以及多个元素的逐步回归分析, 而且其自带的绘图功能能直观地反映出采用不同回归模型的区别及其优缺点, 还可以同时与实测数据在图上对比。便于用户选择不同的模型是其另一功能, 其实现的过程是“Analyze”——“Correlate”。

2.1.6　聚类分析

如果一个监测网有 n 个站位, 每个站位监测 m 个参数, 则可依据各种参数数据的特征对站位进行分类, 也可以对按 m 个参数在不同站位的分布特征对参数之间的关系进行分类, 以便对监测网进行调整或从中找到规律, 对所研究的环境问题做出合理的解释。聚类分析可以客观地找到变量间的亲疏关系, 然后将全部变量归并成不同的类别, 并以分类树形图表示。SPSS 10.0 可以自动进行数据的标准化, 在结果显示中, 通过选择可以得到非常理想的树形分类图, 其实现过程是“Analyze”——“Classify”。

2.1.7　因子分析

治理环境污染, 首先要了解污染源的情况。直接监测污染源是主要途径, 但当污染源众多而且复杂时, 难以直接监测, 于是人们转向监测污染源的排放物, 再由排放物监测数据去推算出污染源的类型及其成因。这时可以将污染源作为若干个待求的因子, 建立起污染源因子与污染物元素数据间的数学模型, 再由该数学模型推导出两者间应满足的关系式, 然后进行判断, 得出结果。在 SPSS 10.0 中的实现过程是“Analyze”——“Data”——“Factor”。

2.1.8　时间序列分析

监测一个区域的环境质量或一个工厂的污染物排放状况常是定时、长期和连续进行的, 监测数据随时间不断变化, 这种变化可以通过时间序列分析对未来环境状况或污染物排放状况的发展趋势进行预测估计。在 SPSS 10.0 中可以通过选择“Graphs”——“Time Series”来实现时间序列分析。

2.2　用于监测系统的设计

空气质量、水质和其他环境要素监测系统的任务是为达到监测目标, 按照监测精度的要求设计采样方案, 建立监测网和采集具有时空代表性的样品数据。具体说, 就是决定采样站的数目、位置以及采样频率、采样时间和采样技术等。如何在经费投入有限的情况下优化环境监测系统的设计是一个非常重要的问题。在实际工作中, 可以对历年监测资料重新分析, 利用聚类分析、方差分析和因子分析方法对监测站的位置布设和采样频率进行调整, 确定必设点位和采样频率, 取消一些重复性点位和多余的采样次数, 避免浪费, 使监测结果更具有科学价值。结合应用 SPSS 10.0 中多种功能分析, 例如将数据分析、统计检验聚类分析、因子分析和相关分析等结合起来运用, 可达到优化环境监测布点和频率的目的, 为更加有效地监测区域内各种污染源, 改善城市环境质量提供科学依据。

3　应用案例

3.1　环境监测数据的多元相关分析

3.1.1　原始数据整理

经过多年监测, 发现一条河的河水中砷的年平均质量浓度可能与河水年平均流量以及上游一个矿山向河道中的日平均倾倒废矿砂量有关, 对多年监测数据整理后得到表 1 中的结果。现依据表 1 所列数据确定是否可建立河水中砷质量浓度与流量及矿砂倾倒量的相关关系。

表 1　原始数据

年　　份	1991	1992	1993	1994	1995	1996	1997	1998	1999
砷质量浓度 y　$\rho/(\mu g\cdot L^{-1})$	15	23	24	21	22	22	17	26	19
流　　量 x_1　$V/(m^3\cdot s^{-1})$	4	11	10	7	8	7	6	12	7
倾 倒 量 x_2　$Q/(t\cdot d^{-1})$	2	3	7	6	5	5	0	7	1

3.1.2　多元相关

应用 SPPS 10.0 软件可进一步了解砷的质量浓度与河水流量和倾倒量的二元相关关系

$$y = 11.417 + 0.927x_1 + 0.542x_2$$

经 t 检验, 按 $\alpha = 0.01$ 水平, 均有显著性意义。同时还可以生成三维散点图。

3.2　水质变化趋势外推

3.2.1　原始数据收集

一个区域地下水过度开采, 水中金属离子质量浓度不断升高, 经过 1990 年—2001 年共 12 年监测, 取得该区域钙(Ca^{2+})的年平均质量浓度数据见表 2, 现在要推测 2002 年和 2003 年钙的质量浓度。

表 2　原始数据

年份编号	1	2	3	4	5	6	7	8	9	10	11	12
年　　份	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
钙质量浓度 $\rho/(mg\cdot L^{-1})$	20.70	17.70	17.60	20.50	20.90	20.90	22.40	25.70	26.60	29.40	31.00	34.10

3.2.2　曲线拟合

选择 Analyze 菜单项中的 Curve estimation 程序, 采用三次曲线进行拟合, 外推得出预测结果。其拟合方程是:

$$y = 21.3465 - 1.9606x + 0.4043x^2 - 0.0129x^3$$

式中: x ——为年份编号;

y ——为预测的 Ca 离子质量浓度, mg/L。

由此得出 2002 年和 2003 年 Ca 离子的预测质量浓度分别为 35.83 mg/L 和 37.72 mg/L, 其误差为 10^{-4}, 同时生成拟合图, 见图 1。

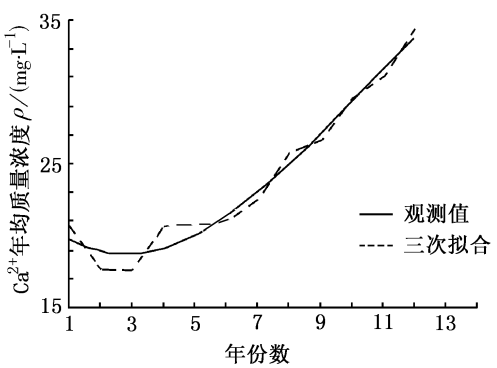

图 1　某区域地下水 Ca 离子年均质量浓度趋势

3.3　水质监测网站位的调整布置

3.3.1　原始数据收集

某河流监测系统中有 5 个水质趋势监测站位, 根据多年监测资料统计, 取得以下 3 个参数的多年平均值, 见表 3。

表 3　5 个站位水质污染物监测的原始数据　　mg/L

污染物	1	2	3	4	5
BOD_5	1.5	1.3	2.1	2.7	3.0
COD	3.2	4.0	3.9	5.5	5.0
NH_3N	0.8	1.2	2.0	2.2	2.1

3.3.2　聚类分析

由于经费限制, 现在要将 5 个站位减至 3 个。

应用 SPSS 10.0 中分层聚类分析(Hierarchical Cluster Analysis)的最优分割法可以处理上述数据, 如果采用欧氏距离计算, 可得到树形图, 见图 2。

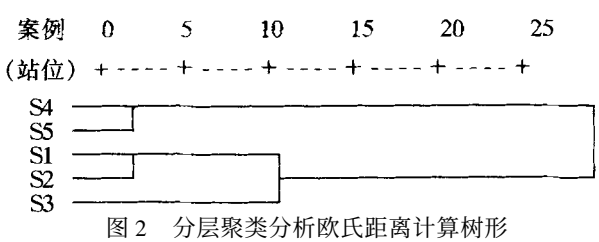

图 2　分层聚类分析欧氏距离计算树形

图中 S1– S5 分别代表 5 个站位, 从聚类树型图中可以清楚地了解到这 5 个站位的最优二分割为{S1, S2, S3}和{S4, S5}, 最优三分割为{S1, S2}, {S3}, {S4, S5}, 即 S1 和 S2、S4 和 S5 的水质监测结果较接近, 可以用 1 个站位代表, 因此, 建议取 S1, S3, S5 为调整后的河流监测站位。

3.4　分层聚类法调整大气监测网的站位

3.4.1　原始数据收集

某市现有监测网近 3 年的大气污染监测数据, 经整理后得到表 4, 用模糊聚类法调整监测网。

表 4　各采样站空气污染物平均浓度数据　　mg/m^3

站点编号	SO_2	NO_X	O_X	CO	TSP
1	0.013	0.011	0.013	0.660	0.170
2	0.067	0.022	0.014	1.060	0.517
3	0.050	0.017	0.011	0.817	0.290
4	0.023	0.020	0.018	0.730	0.300
5	0.055	0.021	0.012	0.960	0.300
6	0.270	0.030	0.020	0.737	0.360

3.4.2　聚类分析

在SPSS 10.0 中将 6 个站位作为变量参数(variable), 5 个空气监测内容作为案例(case), 选择 Analyze——Classify——Hierarchical Cluster Analysis 分析运算, 例题中选用的聚类方法是类间平均锁链法(Between- groups linkage), 对距离的测度方法选择相关系数距离法(Pearson- correlation), 得到的模糊矩阵如下:

模糊矩阵(Proximity Matrix)

	站点 1	站点 2	站点 3	站点 4	站点 5	站点 6
站点 1		0.973	0.995	0.988	0.998	0.928
站点 2	0.973		0.991	0.996	0.984	0.952
站点 3	0.995	0.991		0.997	0.999	0.952
站点 4	0.988	0.996	0.997		0.994	0.942
站点 5	0.998	0.984	0.999	0.994		0.947
站点 6	0.928	0.952	0.952	0.942	0.947	

其树形分类图见图 3。

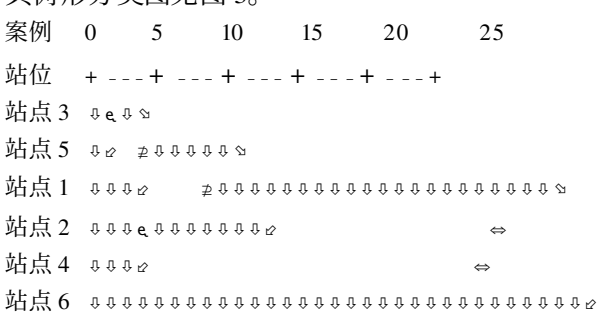

图 3　分层聚类分析相关系数距离法树形

从计算结果可以得出, 该市现有的大气污染监测点可以聚合为 5 类, 即 3 号和 5 号测点归为一类, 其余 2、4、6、和 1 号测点仍保持独立, 换言之, 可用 5 个监测点来监测该市的大气质量变化趋势。

4　结论

从以上介绍可见, SPSS 10.0 在环境监测网设计和数据分析中极具推广价值。

(1) SPSS 10.0 易学易用, 可以举一反三, 事半功倍, 提高个人及部门的工作效率, 缩短工作周期, 保证数据处理的准确性和精密性。

(2) SPSS 10.0 的应用范围广, 大可至一个国家或地区的历年监测数据的分析与预测, 小可至一组实验数据的处理, 对于较大数据量的处理, 显然更有优势。

(3) 相对于其他数据处理软件而言, SPSS 10.0更加面向用户, 直接用对话框选择的形式, 代替了冗长的命令行, 使用更方便, 适合非数据库专业人士使用。

(4) SPSS 10.0 可用于自动监测分析, 定期给出数据结果, 有利于实现环境监测计算自动化。

(5) 可以结合其他软件从数据库发展到信息系统与决策支持系统;

(6) 目前 SPSS 10.0 的输入和输出需用英文, 汉化后将更方便使用。

[参考文献]

[1] 陆雍森. 环境工程手册——环境监测卷[M]. 奚旦立 主编, 蒋展鹏主审, 北京: 高等教育出版社, 1998. 1- 181.

[2] 三味工作室. 世界优秀统计软件 SPSS 10.0 for Windows 实用基础教程[M]. 北京: 北京希望电子出版社, 2001. 1- 413.

[3] 张孟威, 康德梦. 环境问题的数学解法及计算机应用[M]. 北京: 中国环境科学出版社, 1989.

[4] 王林书. 概率论与数理统计[M]. 北京: 科学出版社, 2000.

[5] 奚旦立. 环境监测[M]. 修订版, 北京: 高等教育出版社, 1999.

[6] 李昭智, 李昭勇译. 决策支持与数据仓库系统[M]. 北京: 电子工业出版社, 2001.

[7] 张尧庭译. 离散多元分析理论与实践[M]. 北京: 中国统计出版社, 1998.

[8] 吴伯庆. FOXBASEPLUS 2.1 数据库系统在大气环境监测数据管理中的应用探讨[J]. 煤矿环境保护, 1995, 10(2): 48- 52.

[9] 许建华. 环境监督监测的数据统计处理[J]. 环境监测管理与技术, 1999, 11(4): 41- 42.

[10] 李志辉, 黄国华, 洪　楠. SPSS 7.5 for Windows 95/NT 统计软件包简介[J]. 中国卫生统计, 1998, 15(5): 49- 51.

[11] 赵晓明. SPSS 统计软件在环境监测实验中的应用[J]. 实验技术与管理, 1999, 16(6): 66- 68.

[12] 翟振武. 锐意进取　铺路搭桥——评《社会统计分析方法——SPSS 软件应用》[J]. 人口研究, 2000, 24(5): 77- 79.