

应用多元统计分析理论预测城市 大气污染

黄 国 和

(厦门市环境保护科学研究所)

摘要 本文分析研究地区 31 个环境, 单元中 4 个环境因子与 5 个大气污染指标的相关性。通过对聚类中各因子最佳分割点的确定, 对各子聚类的合并判断以及对各指标的判别筛选, 探索大气污染的发生、发展与环境中诸因子存在状态的定量关系, 从而实现对该区大气污染状况的定量化预测。

关键词 多元统计分析; 大气污染预测。

大气污染是环境问题的一个重要方面。大气污染状况受环境中多种因子的影响, 且各因子又相互联系和制约^[1-4]。显然, 一般的单变量统计分析方法难以客观反映环境中各因子与大气污染状况的相关关系^[5,6]。本研究探讨一套多指标聚类算法。

一、方法和原理

将研究区域划为 n 个环境单元, 对每个单元 r 都监测 p 个污染指标 $\bar{y}(y_1, y_2, \dots, y_i, \dots, y_p)$, 同时调查与这 p 个指标有关的 q 个因子的状态 $\bar{x}(x_1, x_2, \dots, x_k, \dots, x_q)$ 。于是, n 个环境单元中的有关信息便可表示为矩阵 $X = (x_{rk})_{n \times q}$ 和 $Y = (y_{ri})_{n \times p}$ 。本研究的目的在于分析 $\bar{x}$ 中各因子、 $\bar{y}$ 中各指标以及 $\bar{x}$ 和 $\bar{y}$ 间的相关关系, 根据各环境单元所相应的数字信息, 对其进行自然分类, 得出有关指标与因子间关系的聚类图与聚类参数, 实现对各指标的预报。

将聚类 a (含 n_a 个样本) 分割为两个子聚类 b 和 c (分别含 n_b 和 n_c 个样本, $n_b + n_c = n_a$), 若这一分割最优, 则统计量 $U = |W|/|T|$ 应为最小^[7,8]。如果 U 值很大, 则表明不能将聚类 b 与 c 分开, 应将它们归并为较大的聚类 a 。于是, 聚类 b 和 c 的分割或归并便可通过 F 检验来判断^[9]

$$F(p, n_a - p - 1) = \frac{1 - U}{U} \cdot \frac{n_a - p - 1}{p} \quad (1)$$

按照确定的各因子最佳分割点, 首先对聚类 a 中 n_a 个样本相应于各因素 x_k 的取值 $x_{k,i}^{(a)}$ 按大小进行排序, 即

$$x_{k,1}^{(a)} \leq x_{k,2}^{(a)} \leq \dots \leq x_{k,i}^{(a)} \leq \dots \leq x_{k,n_a}^{(a)} \quad (2)$$

然后, 根据排序量 $\{l^k\}$, 逐步计算 $\bar{y}$ 的组内离差阵与总离差阵

$$t_{ij}(n_a^k) = A_{ij}^{(a)}(n_a^k) - n_a^k \cdot B_i^{(a)}(n_a^k) \cdot B_j^{(a)}(n_a^k), \quad (3)$$

$$w_{ij}(l^k, n_a^k) = t_{ij}(n_a^k) - v_{ij}(l^k, n_a^k), \quad (4)$$

其中

$$v_{ij}(l^k, n_a^k) = \frac{n_a^k \cdot l^k \{ [B_i^{(a)}(l^k) - B_i^{(a)}(n_a^k)] [B_j^{(a)}(l^k) - B_j^{(a)}(n_a^k)] \}}{n_a^k - l^k} \quad (5)$$

$$A_{ij}^{(a)}(\alpha) = \sum_{l=1}^{\alpha} y_{il}^{(a)} \cdot y_{jl}^{(a)}, \quad (6)$$

$$B_{\beta}^{(a)}(\alpha) = \frac{1}{\alpha} \sum_{l=1}^{\alpha} y_{a,l}, \quad (7)$$

$i, j = 1, 2, \dots, p; k = 1, 2, \dots, q$; 对于每个 k , 令

$$l^k = 1^k, 2^k, \dots, n_a^k - 1. \quad (8)$$

那么, 对于每个因子 x_k , 可得分割点 l_0^k , 满足

$$U(l_0^k, n_0^k) = \text{Min}_{l_0^k=1, k}^{n_0^k-1} \{U(l_0^k, n_0^k)\} \quad (9)$$

对于 q 个因子, 可得分割点 k_0 , 满足

$$U(l_0^{k_0}, n_0^{k_0}) = \text{Min}_{k=1}^q \{U(l_0^k, n_0^k)\}. \quad (10)$$

于是, 聚类 a 的最佳分割点便为 $l_0^{k_0}$, 相应的因子值为 $x_{k_0}^{(a)}, l_0^{k_0}$. 如果

$$F(p', n_0^{k_0} - p' - 1) = \frac{1 - U(l_0^{k_0}, n_0^{k_0})}{U(l_0^{k_0}, n_0^{k_0})} \cdot \frac{n_0^{k_0} - p' - 1}{p'} \geq F_1, \quad (11)$$

便可按因子 x_{k_0} 的状态对聚类 a 进行分割:

$$l_0^{k_0} \leq l_0^{k_0} \text{ 的样本归于聚类 } b (< c), \quad (12)$$

$$l_0^{k_0} > l_0^{k_0} \text{ 的样本归于聚类 } c. \quad (13)$$

式中 p' 表示已选的指标数. 于是 x_{k_0} 就是对指标有显著影响的重要因子, 应予入选. 反之, 若(11)式不被满足, 则不能对聚类 a 进行分割. 于是, 进一步判断其它聚类可否被分割, 即对 $a = 1, 2, \dots, A$ 进行判断. 当不再有允许分割的聚类时, 则开始考虑聚类的归并.

设对 A 个聚类分别考虑其中子聚类 b 和 c 归并的可能性. 首先计算各 t_{ij} 和 w_{ij}

$$\begin{aligned} t_{ij}(n_b, n_c) &= A_{ij}^{(b)}(n_b) + A_{ij}^{(c)}(n_c) \\ &- \{[n_b \cdot B_i^{(b)}(n_b) + n_c \cdot B_i^{(c)}(n_c)] \\ &\cdot [n_b B_i^{(b)}(n_b) + n_c \cdot B_i^{(c)}(n_c)]\} / \\ &\{n_b + n_c\}, \end{aligned} \quad (14)$$

$$w_{ij}(n_b, n_c) = t_{ij}(n_b, n_c) - V_{ij}(n_b, n_c), \quad (15)$$

其中

$$\begin{aligned} v_{ij}(n_b, n_c) &= \{n_b \cdot n_c \cdot \{[B_i^{(b)}(n_b) - B_i^{(c)}(n_c)] \\ &\cdot [B_i^{(b)}(n_b) - B_i^{(c)}(n_c)]\} / \{n_b + n_c\}, \quad (16) \\ &i, j = 1, 2, \dots, p. \end{aligned}$$

于是, 作 F 检验, 如果

$$F(p', n_b + n_c - p' - 1) = \frac{1 - U(n_b, n_c)}{U(n_b, n_c)} \cdot \frac{(n_b + n_c) - p' - 1}{p'} < F_2, \quad (17)$$

则可将聚类 b 和 c 归并为一个新聚类 a . 反

之, 若(17)式不被满足, 则不能将聚类 b 和 c 归并. 于是, 进一步判断其它聚类可否归并, 即对 $b = 1, 2, \dots, A-1$ 和 $c = 2, 3, \dots, A$ 进行判断. 当不能再分割, 也不许归并时, 则计算结束. 否则, 需返回去考虑聚类的分割.

上述分割与归并计算结果最后形成一个聚类树. 树上的每个分割节点与一个入选的因子值 $x_{k_0}^{(a)}, l_0^{k_0}$ 相对应. 那么, 应用于预测计算时, 便可将新样本的 x_{k_0} 值与聚类树各分割节点上的 $x_{k_0}^{(a)}, l_0^{k_0}$ 值进行判断比较, 使之归入相应的枝端聚类, 从而实现预测.

具体算法为, 首先根据下式将新样本逐步归到聚类树的相应分支中, 最后该样本将归入某一枝端聚类. 判别规则为:

$$x_k \leq x_{k_0, l_0^{k_0}}^{(a)} \text{ 的样本划归聚类 } b (< c), \quad (18)$$

$$x_k > x_{k_0, l_0^{k_0}}^{(a)} \text{ 的样本划归聚类 } c. \quad (19)$$

那么, 设新样本归入枝端聚类 b_0 , 则相应的各指标预测值为

$$\hat{y}_i = \bar{y}_i^{(b_0)} \pm r_i^{(b_0)} \quad \forall i \quad (20)$$

式中 $\bar{y}_i^{(b_0)}$ 为第 i 个指标在子聚类 b_0 中的平均值; $r_i^{(b_0)}$ 为指标 y_i 在子聚类 b_0 中的凝聚半径. 即

$$\bar{y}_i^{(b_0)} = \frac{1}{n_{b_0}} \sum_{l=1}^{n_{b_0}} y_{il}^{(b_0)} \quad (21)$$

$$r_i^{(b_0)} = \frac{1}{2} \{ \text{Max}_{l=1}^{n_{b_0}} [y_{il}^{(b_0)}] - \text{Min}_{l=1}^{n_{b_0}} [y_{il}^{(b_0)}] \}. \quad (22)$$

在整个聚类过程中始终落选的因子, 对聚类不提供附加信息, 故可不予考虑.

二、实例研究

研究地区含 31 个环境单元 (以网格划分). 对各单元中 3 个环境污染指标 (SO_2 、 NO_x 和降尘) 及与之有关的 4 个环境因子 (工业煤耗、人口密度、交通密度、饮食服务点密度) 均进行全面的调查、监测和统计衡算, 结果见表 1.

表 1 X 和 Y 的数值集合

样本编号	SO ₂ (mg/m ³)	NO _x (mg/m ³)	降 尘 (t/km ² · mo)	工业煤耗 (10 ⁴ t/km ² · yr)	人口密度 (10 ⁴ 人/km ²)	交通密度 (辆/km ²)	饮食服务点密度 (家/km ²)
1	0.007	0.010	5.61	0.084	0.039	35.7	23
2	0.020	0.004	5.68	0.640	0.053	3.0	8
3	0.012	0.011	9.52	0.102	0.051	8.3	12
4	0.023	0.011	9.53	0.004	0.043	12.0	22
5	0.033	0.013	7.60	0.050	0.093	26.3	24
6	0.062	0.016	9.53	0.432	0.185	31.2	41
7	0.041	0.015	10.37	0.120	0.203	33.8	39
8	0.071	0.014	8.07	0.430	0.194	50.8	22
9	0.011	0.015	5.37	0.099	0.079	27.2	12
10	0.008	0.011	5.53	0.028	0.093	42.9	24
11	0.021	0.027	5.05	0.108	0.618	43.7	44
12	0.078	0.021	10.45	2.312	0.439	39.8	74
13	0.099	0.022	11.83	1.403	0.330	48.1	77
14	0.120	0.024	9.21	0.245	0.451	19.3	68
15	0.062	0.016	11.45	0.673	0.208	15.1	7
16	0.092	0.044	10.24	2.319	0.245	9.1	12
17	0.044	0.040	9.76	0.933	0.619	17.3	64
18	0.073	0.036	6.34	1.566	2.255	16.7	96
19	0.120	0.081	6.86	0.217	6.732	46.8	276
20	0.009	0.005	6.43	0.704	1.164	16.1	40
21	0.014	0.007	7.18	0.017	1.233	25.8	39
22	0.120	0.078	8.68	0.144	5.544	42.5	260
23	0.045	0.044	7.62	0.323	1.271	32.0	77
24	0.024	0.026	6.65	0.144	0.152	10.7	5
25	0.056	0.053	8.91	0.503	1.255	7.1	57
26	0.028	0.029	6.51	0.080	1.632	8.3	54
27	0.024	0.006	4.11	0.120	1.442	6.8	135
28	0.014	0.006	6.06	0.078	1.268	7.3	140
29	0.030	0.030	7.30	0.045	0.247	6.4	80
30	0.021	0.009	7.03	0.121	0.224	6.7	29
31	0.012	0.008	6.36	0.099	1.379	9.4	6

对各环境因子的分布状态与各环境污染指标含量间的关系作单变量相关分析, 结果见表 2。可见, SO₂ 含量与因子 x_1 、 x_2 、 x_3 和 x_4 均具有显著的正相关关系 ($p < 0.05$), 尤其 SO₂ 含量与 x_1 、 x_2 和 x_4 具更密切的相关关系, 其相关系数 $r \geq 0.462$ ($p < 0.01$); NO_x 含量与因子 x_2 、 x_3 和 x_4 具显著的正相关关系 ($p < 0.01$); 降尘量则与因子 x_1 和 x_3 具显著的正相关关系 ($p < 0.01$), 从各污染指标间的相关关系看, SO₂ 含量与 NO_x 含量间的单相关系数为 0.690 ($p < 0.01$), 与降尘量间单

相关系数为 0.543 ($p < 0.01$), 均显著相关; 但 NO_x 含量与降尘量之间的单相关系数为 0.193 ($p \sim 0.1$), 尚未达到显著水平。

从各因子间的相关关系看, x_2 与 x_3 的单相关系数为 0.623 ($p < 0.01$), x_2 与 x_4 的单相关系数为 0.897 ($p < 0.01$), 均显著相关。但 x_1 与 x_2 、 x_3 和 x_4 的相关系数 $|r| \leq 0.074$ ($p > 0.3$), 尚未达到显著水平; x_3 与 x_4 的相关系数为 0.297 ($p > 0.05$), 也未达到统计学显著性水平。

显然, 由于变量间存在显著相关信息, 因

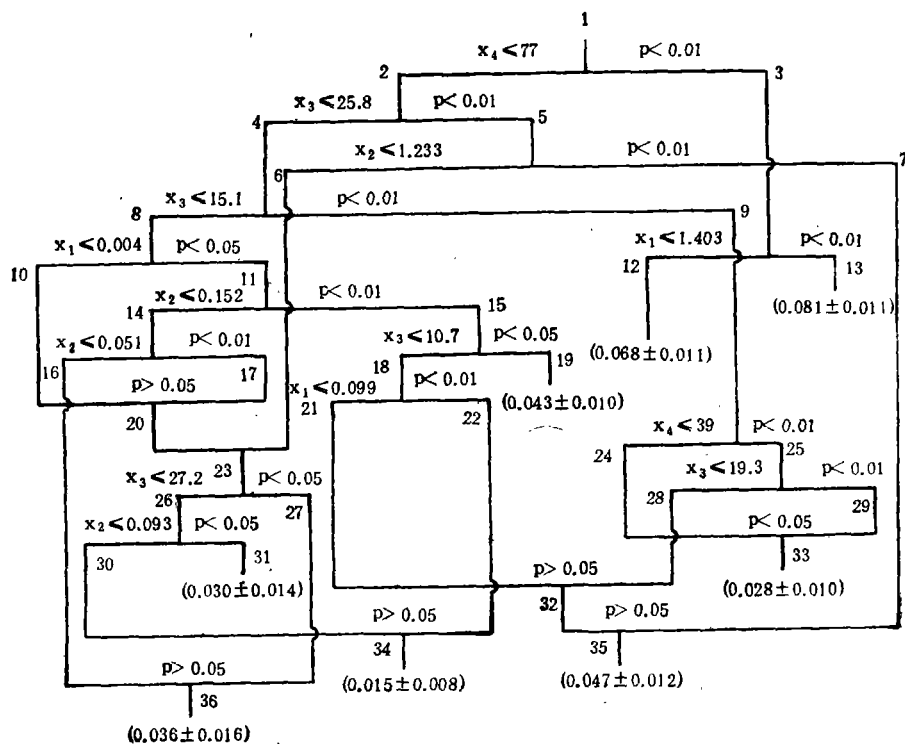
表 2 各变量间的相关系数

变 量 1	变 量 2	r	变 量 1	变 量 2	r
SO ₂	NO _x	0.690	降 尘	x ₁	0.466
	降尘	0.543		x ₂	-0.139
	x ₁	0.462		x ₃	0.458
	x ₂	0.507		x ₄	-0.073
	x ₃	0.397	x ₁	x ₂	-0.074
	x ₄	0.560		x ₃	0.064
NO _x	降尘	0.193		x ₄	-0.037
	x ₁	0.178	x ₂	x ₃	0.623
	x ₂	0.758		x ₄	0.897
	x ₃	0.734	x ₃	x ₄	0.297
	x ₄	0.699			

此, 只有用多元统计方法进行研究, 才可能客观和正确反映各环境因子与各污染指标间存在的关系^[9-12]。事实上, 城市大气中各污染指标含量在时空上均存在不同的分布水平^[13,14], 因此可以通过多指标逐步聚类方法, 对总体样本(调查结果)进行自然分类。如果各环境因子对各污染指标具有显著影响, 则

在研究计算中这些因子必将入选而构成聚类树上各分支的最佳分割点^[11,12]。

图 1—3 分别为通过大量上机计算得出的 SO₂、NO_x 和降尘量的聚类树。具体分割与合并的置信水平分布原则为 $p < 0.05$ 时才允许对聚类进行分割; $p > 0.05$ 表明两个聚类经 F 检验无显著性差异, 则可给予归并。

图 1 SO₂ 的聚类树

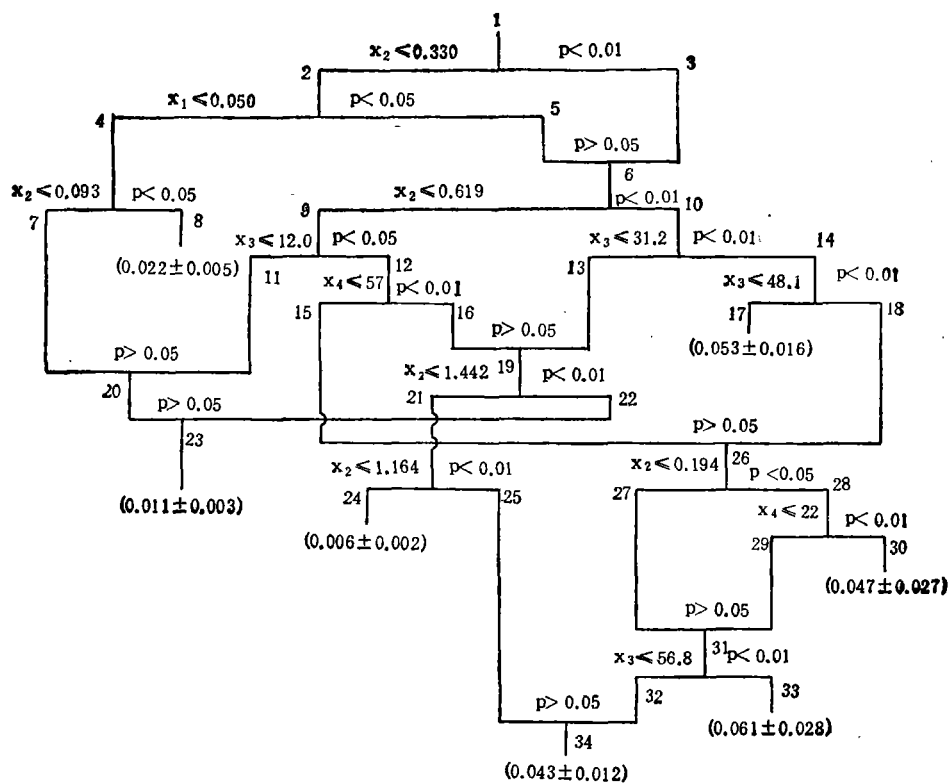


图2 NO_x 的聚类树

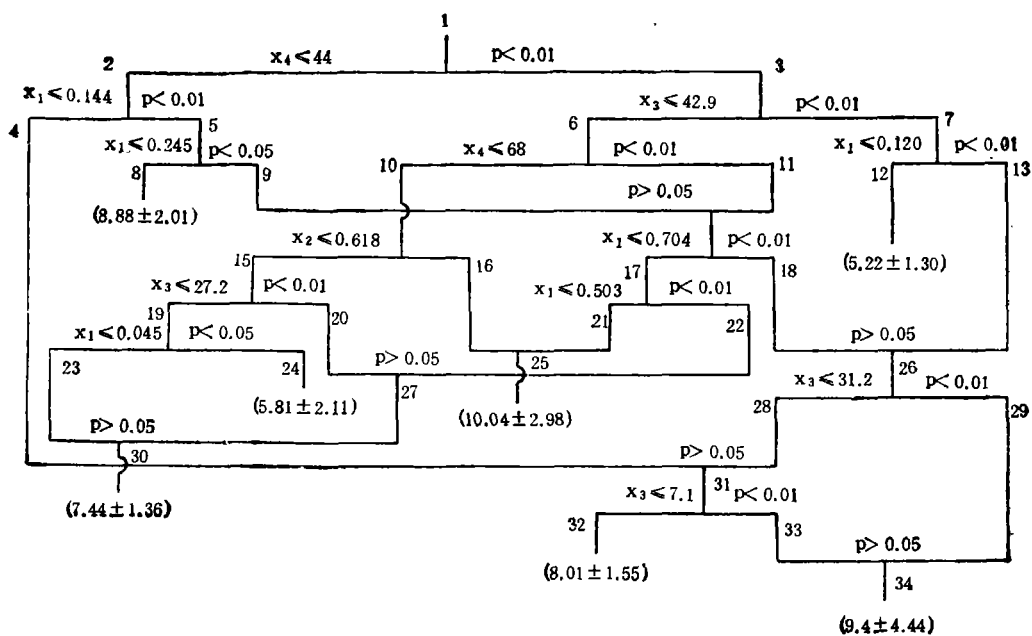


图3 降尘量的聚类树

图上各分割与合并节点上标出的便是判别时 F 检验的显著性水平。

由图 1—3 可见, SO_2 、 NO_x 和降尘含量与工业煤耗、人口密度、交通密度和饮食服务点密度等因子具有显著的相关关系。于是,借助于这些聚类树,便可由城市中各环境因子的分布状态预测出大气中 SO_2 、 NO_x 和降尘的含量。图 1—3 事实上是一个预测系统。对于新的样本,只要调查核算出未来时段该样本中各因子的分布状况,便可循各指标所相应的聚类树进行预测。以 SO_2 (图 1) 的预测为例,设已知 $x_1=0.044$, $x_2=0.087$, $x_3=20.7$, $x_4=48$, 那么,从第一个分割点看, $x_1 < 77$, 故归入聚类 2; $x_3 < 25.8$, 故归入聚类 4; $x_3 > 15.1$, 故归入聚类 9; $x_4 > 39$, 故归入聚类 25; $x_3 > 19.3$, 故该样本最后归入枝端聚类 33, 其 SO_2 预测值为 $0.028 \pm 0.010(\text{mg}/\text{m}^3)$ 。

三、结 论

1. 本研究所提出的预测大气环境质量的统计学方法体系综合了各有关环境因子的影响。在统计、判别和逐步聚类分析研究过程中,由于各因子均带有各指标分布状况的信息,因而使数值分析与预测结果具更高的准确度。

2. 由本方法计算得出的预测系统表现为一些聚类树。在数值分析与判别过程中,带有指标与影响指标的各因子信息的各样本经由不同的路径与分割节点而凝聚于相应的枝端中。于是,各指标所相应的各聚类树均反

映了各因子的影响状况。只须知道未来时段各因子的分布状况,便可由这些聚类树而实现对指标的预测。

3. 预测结果以分类均值和凝聚半径综合表述,更符合客观实际,从而使之在应用于决策时具更大的可信度^[15,16]。

4. 本方法体系适于环境因子分布较为明确、环境污染现象较为显著的城市地带或工业区。事实证明,应用本方法进行预测,无论在预测质量(准确度)还是在预测周期(耗时)和预测耗资上均具较大的优越性,可为其他地区参考借鉴。

参 考 文 献

- [1] 黄国和,环境科学, 9(1),9(1988).
- [2] 万钟英,建筑学报, (4),39(1987).
- [3] 林文生、黄国和,环境科学与技术, (4),25(1987).
- [4] 沈铎,中国环境科学, 4(3),41(1984).
- [5] 张尧庭,应用数学学报, 4,312(1978).
- [6] Giri, N. C., *Multivariate Statistical Inference*, p. 16, Academic Press, New York, 1977.
- [7] Wilks, S. S., *Mathematical Statistics*, p. 5, John Wiley, New York, 1962.
- [8] Rao, C. R., *Linear Statistical Inference and Its Applications*, p. 10, John Wiley, New York, 1965.
- [9] 方开泰,数学的实践与认识, (2),54(1978).
- [10] 陈韵源等,中国科学, (5),817(1977).
- [11] 张尧庭等,多元统计分析引论,第 393 页,科学出版社,北京,1982 年.
- [12] Dystra, R. L., *Ann. Math. Statist.*, 41, 2153 (1970).
- [13] 黄国和,环境科学丛刊, 7(11),49(1986).
- [14] 黄国和,中国环境监测, 4(1),1(1988).
- [15] 黄国和,环境科学学报, 6(3),306(1986).
- [16] 曾光明、黄国和,水利学报, (10),54(1988).

(收稿日期: 1990 年 6 月 6 日)

• 环境信息 •

没有证据说明生活在核动力设备的人们癌症发病率高

据美国国家癌症研究所“核设施附近居民中的癌症”课题经二年研究的结果,对 34 个州、107 个县、靠近 62 个核设施生活的人们进行调查评定,其癌症死亡率绝不比没有核设施的 292 个县的高。事实上发现死于白血病的风险确实还低一些,之所以突出白血病是因为它潜伏期短,平均约五年。而且

已经表明,暴露在 10 雷姆 (rem, 剂量当量, $1\text{rem} = 10^{-2}\text{Sv}$) 以下对人体健康无不良影响。10 雷姆比美国核规章委员会假定的一个人站在核电厂防护隔离室每天 24 小时连续一年所受剂量还高出上千倍。

仲民译自 ES&T, 25(1),11(1991).

leached from red soil were much larger than those from yellow brown soil or black soil. Leaching of Cd was more sensitive than that of Cu.

(2) The effects of simulated acid rain on speciations of Cu or Cd in soils were relatively obvious. With pH values of the precipitation reducing, a partial Cu and Cd in yellow brown soil and black soil were transformed into exchangeable speciation, and those in red soil transformed into water-soluble form.

(3) Acid rain evidently affected biotoxicity of Cu or Cd in soil. The effect of Cu- or Cd-biotoxicity was high in red soil and low in black soil.

In short, the acid resistant capacity of soils in order is black soil > yellow brown soil > red soil.

Key Words: acid rain, copper and cadmium, soil, biotoxicity.

Application of the Theory of Multivariate Statistical Analysis to the Prediction of Air Pollution in Urban Environment. Huang Guo-he (Xiamen Research Institute of Environmental Protection, Fujian): *Chin. J. Environ. Sci.*, 12(2), 1991, pp. 29—34.

A stepwise cluster analysis method toward a multitude of environmental variables was advanced for the prediction of air pollution in urban environment. The method has improved monovariate AID(Automatic Interaction Detection) algorithm. It can effectively deal with continuous and discrete variables, as well as the nonlinear relations among variables. Compared with classical modelling theories, the method provides a new way for predicting air pollution.

In a case study, the relationship between three air pollutant concentrations and four impact factors in thirty-one grid squares in Xiamen was analyzed. Cluster trees concerning quantitative correlation between air quality and the distribution of relevant impact factors were derived from significant test, discriminant analysis and cluster analysis to realize the aim of predicting air quality.

Key Words: multivariate statistical analysis, air pollution, stepwise cluster.

Use of the Residue after Reclamation of Protein and Chrome Salt from Leather Shavings. Jiang Ting-da, Zhang Chun-ping (Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing): *Chin. J. Environ. Sci.*, 12(2), 1991, pp. 35—37.

This paper presents a part of the process which was developed for reclamation of protein and by-products from chrome leather shavings. After protein and chrome(III) are extracted from leather scraps, the remainder is neutralized with quick lime, and then processed into granu-

lated fertilizer. The results of tests showed that the organic fertilizer is available for flowers because of its long-term and high effectiveness.

Key Words: reclamation of leather shavings, residue, fertilizer.

Analysis and Prediction of Traffic Noise in the Mountain City. Chen Zi-ming, Lin Dong (Department of Physics, Qingdao University of Oceanography): *Chin. J. Environ. Sci.*, 12(2), 1991, pp.38—41.

This paper presents a method for analysis and prediction of traffic noise in the mountain city, emphasizing the characteristics of sloping roads. Taking Qingdao City as an instance, based on road structure, distribution of terraced buildings along both sides of the sloping roads and the conditions of motor vehicles shuttling to and fro on the roads with radiative noise, three different models have been set up for long-term monitoring, and the linear relations among varieties of vehicle noise, speeds and road slopes have been demonstrated. The results summed up correct range of noise levels caused by the road-side buildings and sloping road structure. The authors propose the method for predicting L_{eq} of traffic noise on the sloping road surface in practice.

Key Words: traffic noise, mountain city, prediction, model.

Influence of Acidic Leaching upon the Contents of Organic Matter and Clay Particle in Soil. Liu Hong-jie (Department of Geography, South China Normal University, Guangzhou): *Chin. J. Environ. Sci.*, 12(2), 1991, pp. 42—44.

This paper deals with variations of the contents of organic matter and clay particle in soil under influence of acidic precipitation. The results of simulated experiments showed that organic matter contents decreased in varying degree after acidic leaching, but acidity of leaching did not have marked influence upon the clay contents. If the contents of metals in soil solution increased to critical coagulation, efficiency of eluviation would decrease.

Key Words: acidic leaching, organic matter, clay particle.

Treatment of Furfural Wastewater by Membrane Separation Process. Chen Yu-lian, Zhou Guang-jun, Zhang He (Department of Chemical Engineering, Taiyuan University of Technology, Shanxi): *Chin. J. Environ. Sci.*, 12(2), 1991, pp. 45—49.

A new method for recovering acetic acid from furfural