基于主成分综合模型的矿区农田重金属污染评价

王从陆[1]，吴　超[2]，段　瑜[3]

（1. 湖南科技大学能源与安全工程学院，湖南湘潭 411201；2. 中南大学资源与安全工程学院，湖南长沙 410083；
3. 湖南理工职业技术学院，湖南湘潭 411104）

[摘要] 文章尝试利用变量聚类分析方法对矿区附近农田土壤重金属污染的主要污染物进行辨识，并采用综合主成分分析法对矿区附近农田土壤重金属污染情况进行评价和分级。分析结果表明：利用变量聚类分析法可以有效地辨识矿区附近农田土壤重金属污染中的主要成分；运用综合主成分分析法，确定样本的综合主成分，并对其排序和聚类，可以有效揭示矿区附近农田土壤重金属污染物的数据结构、相互关系和不同样品点的污染程度。采用主成分分析方法对矿区附近农田土壤重金属污染情况的评价结果，反映了矿区主要重金属污染物的影响，同时又定量化了土壤复合重金属污染研究。辨识和评价结果可为矿区附近农田土壤重金属污染治理对策的提出和重点治理区域的确定提供参考和指导。

[关键词] 主成分综合模型；矿区农田土壤；重金属污染；评价

[中图分类号] X502 [文献标识码] A [文章编号] 1009－1742(2008)07－0084－04

据统计，我国耕地因重金属而造成污染的面积接近133万 hm^2，约占耕地面积的1/5。我国每年因重金属污染导致的粮食减产超过1 000万t，被重金属污染的粮食多达1 200万t，合计经济损失至少200亿元。重金属污染对粮食安全的影响受到关注。

有色金属矿山生产过程中的许多环节，包括凿岩、爆破、运输、通风、排水、选矿和尾矿等，都会产生重金属污染物，且含量较高。当把它们从地下搬到地表后，由于物理、化学条件的改变，重金属元素的释放、迁移，对附近土壤等产生严重的重金属污染。因此，有色金属矿山是重金属污染的重要来源，首当其冲的就是矿区附近的农田。近年来，国内外学者对部分铅锌尾矿、铜尾矿污染区重金属污染现状，包括重金属含量、形态特征以及对矿区植被的影响等方面进行了大量研究，但对矿区农田重金属污染评价的研究较少[1~4]。笔者参照GB5618—1995有关标准，利用统计分析软件，应用聚类分析方法，分析主要污染源；采用综合主成分分析法，分析、评价不同采样点的农田的重金属污染程度，并进行排序，为矿区农田重金属污染治理提供参考。

1　主成分分析的统计依据与步骤

主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标。主成分分析的基本思想，在实证问题研究中为了全面、系统地分析问题，笔者必须考虑众多影响因素。每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关性，因而所得的统计数据反映的信息在一定程度上有重叠。采用统计方法研究多变量问题时，变量太多会增加计算量和增加分析问题的复杂性，人们希望在进行定量分析的过程中，涉及的变量较少，得到的信息量较多。主成分分析正是适应这一要求产生的，是解决这类题的理想工具[5~8]。

1.1　主成分分析的统计依据

设有 p 个指标 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$，这 p 个指标反映了客观对象的各个特征，因此每个对象观察到的 p 个指

[收稿日期] 2007－07－08；修回日期 2007－09－20

[基金项目] 湖南省自然科学基金资助科研项目(06JJ50079)

[作者简介] 王从陆(1972－)，男，博士，江西万年县人，湖南科技大学能源与安全工程学院，主要为研究方向矿山安全、环保研究

标值就是一个样本值，它是一个 p 维向量。如果观察了 n 个对象，就有 n 个 p 维向量，可用矩阵表示如下：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p)$$

每一行就是一个样本的观察值，可用数据矩阵 $\boldsymbol{X}$ 的 p 个向量作线性组合为：

$$F_1 = a_{11}\boldsymbol{x}_1 + a_{21}\boldsymbol{x}_2 + \cdots + a_{p1}\boldsymbol{x}_p$$
$$F_2 = a_{12}\boldsymbol{x}_1 + a_{22}\boldsymbol{x}_2 + \cdots + a_{p2}\boldsymbol{x}_p$$
$$\cdots\cdots\cdots\cdots$$
$$F_n = a_{1n}\boldsymbol{x}_1 + a_{2n}\boldsymbol{x}_2 + \cdots + a_{pn}\boldsymbol{x}_p$$

上述方程组要求 $a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2 = 1$，且系数 a_{ij} 由下列原则决定：

1) F_i 与 F_j($i \neq j$) 不相关；

2) F_1 是 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 的一切线性组合(系数满足上述方程组)中方差最大的；F_2 是与 F_1 不相关的 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 的一切线性组合(系数满足上述方程组)中方差最大的；依次类推，F_p 是与 F_1，F_2，…，F_{p-1} 都不相关的 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 的一切线性组合(系数满足上述方程组)中方差最大的。F_1，F_2，…，F_{p-1} 为第 1,2,…,p 主成分。

在解决实际问题时，一般不是取 p 个主成分，而是根据累计贡献率的大小取前 k 个。称第 1 主成分的贡献率为 $\lambda_1/(\lambda_1+\lambda_2+\cdots+\lambda_p)$，由于 $\mathrm{Var}(F_1)=\lambda_1$，因此第 1 主成分的贡献率就是第 1 主成分的方差与全部方差的比值。这个值越大，表明第 1 主成分综合 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 信息的能力越强。前个主成分的累计贡献率定义为$(\lambda_1+\lambda_2+\cdots+\lambda_k)/(\lambda_1+\lambda_2+\cdots+\lambda_p)$。如果前 k 个主成分的累计贡献率达到某个给定值，则表明取前 k 个主成分基本包含了全部测量指标所具有的信息，这样既减少了变量的个数，又便于对实际问题进行分析和研究。

1.2 主成分分析的步骤

1)将原始数据标准化；

2)建立变量的相关系数阵 $\boldsymbol{R}=(r_{ij})=\boldsymbol{x}'\boldsymbol{x}$；

3)求 $\boldsymbol{R}$ 的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$；

4)写出主成分 $F_i = a_{1i}\boldsymbol{x}_1 + a_{2i}\boldsymbol{x}_2 + \cdots + a_{pi}\boldsymbol{x}_p$，$i=1,2,3,\cdots,p$。

5)每个主成分所对应的特征值占所提取主成分总的特征值之和的比例作为权重，计算主成分综合模型。

2 数据来源

利用湖南省某典型金属矿山区农田土壤重金属污染物调查数据作为主成分分析的土壤重金属污染评价素材，矿区农田土壤样品来自 15 个不同的片区，样本总数为 15 个，分析指标主要考虑国家颁布的土壤环境质量标准中的几种重点识别重金属污染物(汞 Hg、锌 Zn、铅 Pb、镉 Cd、砷 As、镍 Ni、铬 Cr、铜 Cu)，样品采集和分析严格按照 GB5618—1995 的相关要求进行，具体样品数据见表 1。

表 1 各土壤样品中的重金属含量

Table 1 Heavy meta content of different soil samples

样品编号	重金属含量/(mg·kg^{-1})							
	汞 Hg(x_1)	锌 Zn(x_2)	铅 Pb(x_3)	镉 Cd(x_4)	砷 As(x_5)	镍 Ni(x_6)	铬 Cr(x_7)	铜 Cu(x_8)
1	0.161	3 267	363.2	12.75	85.11	4.90	13.4	357.5
2	0.199	2 239	496.8	6.00	89.76	11.60	17.1	351.5
3	0.148	8 784	7 932	19.50	1 083.30	14.20	16.4	694
4	0.196	5 152	1 625	19.50	207.40	3.10	11.9	417.9
5	0.123	2 704	224.8	2.75	86.63	2.10	21.4	344.2
6	0.097	1 521	568.4	5.25	108.90	7.60	5.6	369.9
7	0.097	96	234.9	13.25	94.90	7.80	8.4	339.4
8	0.085	152	220.0	8.00	47.96	1.90	16.4	337.0
9	0.116	888	81.6	0.5	80.47	1.80	17.1	327.3
10	0.011	78.4	167.5	1.00	66.54	6.2	21.6	330.9
11	0.016	59.7	46.9	0.60	42.50	2.3	1.6	410.1
12	0.012	198.2	110.4	0.89	24	10.20	2.1	402.4
13	0.021	215.6	104.5	2.3	18.90	0.8	1.0	345.6
14	0.014	498.3	85.0	0.45	21.25	7.5	1.3	362.1
15	0.214	1 123	46.8	6.3	56.10	4.6	2.3	561.0

3 结果与讨论

3.1 主要污染物辨识分析

把表 1 中的原始数据导入 SPSS 统计软件，为了辨识出主要的重金属污染物，把所有评估的污染物变量聚为两类，分类结果见表 2。

表 2 重金属污染物辨识聚类

Table 2 Identification cluster of heavy metal pollutants

聚类 1	聚类 2
x_2，x_3	x_1，x_4，x_5，x_6，x_7，x_8

从表 2 中看出，锌和铅被划分为同一类，其他重金属元素为另一类。说明锌和铅是该地区的主要重金属污染物，需要引起高度重视。事实上，土壤样本来源于一个铅锌矿附近的农田。辨识结果表明，变量聚类的方法，可以用于辨识矿区土壤重金属污染的主要污染源。

3.2 综合主成分分析模型

同样，采用表 1 中的原始数据，导入 SPSS 统计软件，通过软件计算，提取主成分，得到初始因子载荷矩阵，见表 3。特征值在某种程度上可以被看成是表示主成分影响力度大小的指标，如果特征值小于 1，说明该主成分的解释力度还不如直接引入一个原变量的平均解释力度大，因此一般可以用特征值大于 1 作为纳入标准。因此，主成分个数提取原则为主成分对应的特征值大于 1 的前 m 个主成分。

表 3 主成分载荷矩阵

Table 3 Principal component matrix

因子	主成分	
	1	2
x_1	0.550	0.593
x_2	0.943	0.143
x_3	0.944	−0.201
x_4	0.792	0.278
x_5	0.944	−0.179
x_6	0.592	−0.409
x_7	0.305	0.714
x_8	0.814	−0.367
λ(特征值)	4.708	1.334

用图表 3 中的数据除以主成分相对应的特征值开平方根便得到两个主成分中每个指标所对应的系数，对照主成分系数可以写出主成分表达式：

$$F_1 = 0.253\,5x_1 + 0.434\,6x_2 + 0.435\,1x_3 + 0.365\,0x_4 + 0.435\,1x_5 + 0.272\,8x_6 + 0.140\,6x_7 + 0.375\,1x_8$$

$$F_2 = 0.513\,4x_1 + 0.123\,8x_2 - 0.174\,0x_3 + 0.240\,7x_4 - 0.155\,0x_5 - 0.354\,2x_6 + 0.618\,2x_7 - 0.317\,8x_8$$

对应的主成分综合模型为：

$$F = 0.310\,9x_1 + 0.366\,0x_2 + 0.300\,6x_3 + 0.337\,6x_4 + 0.304\,8x_5 + 0.134\,4x_6 + 0.246\,0x_7 + 0.222\,2x_8$$

根据主成分表达式和主成分综合模型表达式，结合样品数据，计算出各样品的主成分和综合主成分的具体值，并进行排序，见表 4。

表 4 主成分和综合主成分的取值

Table 4 Value of principal component and synthetic principal component

样品	第一主成分 F_1	排名	第二主成分 F_2	排名	综合主成分 F	排名
1	0.462 819	4	1.222 882	3	0.630 631	4
2	0.579 778	3	0.904 296	6	0.651 43	3
3	6.988 308	1	−1.223 12	12	5.175 325	1
4	1.837 323	2	1.473 693	2	1.757 036	2
5	−0.435 34	8	1.511 361	1	−0.005 53	6
6	−0.308 23	7	−0.518 68	10	−0.354 7	8
7	−0.270 32	6	0.036 732	8	−0.202 53	7
8	−0.932 24	9	0.966 247	5	−0.513 08	9
9	−1.109 32	11	1.039 406	4	−0.634 91	10
10	−1.198 88	12	0.244 277	7	−0.880 25	11
11	−1.608 65	14	−1.222 83	13	−1.523 47	14
12	−1.080 99	10	−1.860 65	15	−1.253 13	12
13	−1.849 57	15	−0.822 97	11	−1.622 91	15
14	−1.402 3	13	−1.542 68	14	−1.433 3	13
15	0.327 609	5	−0.207 97	9	0.209 36	5

从第 1 和第 2 主成分值的排名情况来看，存在较大的差异，原因在于不同主成分之间反映样本信息的重点是不同的。第 1 主成分包含了样本的大多数信息，在综合主成分中占有很大的比重，因此，综合主成分的排名与第 1 主成分基本类似。考虑到主成分反映信息的侧重点不同，第 2 主成分修正了第 1 主成分在综合主成分中的排名。主成分的值表示在确定的信息提取规则下，各样品中待评价因素的综合贡献。从综合评价的排名情况来看，样品 1 至 4 及 15 的重金属污染情况较为严重。对应的采样区域就是重金属污染治理的重点区域。

3.3 样品污染等级划分

假设把样本采集区的农田污染等级划为3类，即污染程度严重、污染程度一般和污染程度轻微。根据综合主成分值，采用3层聚类的方法，可以把土壤样品污染情况进行聚类，见表5。

表5 样品污染程度聚类结果

Table 5 Cluster result of samples pollution degree

类别	严重污染	一般污染	轻微污染
样品编号	3	4	1,2,5,6,7,8,9,10,11,12,13,14,15

从聚类结果可以得到，样本3和4分别单独成类，其它的所有样本构成一类。结合综合主成分的排序，可以看出样本3是重金属污染最严重的样本，其次为样本4。参考样本3和4的原始数据，对于样本3来说，Pb和Zn的含量分别为7 932 $mg \cdot kg^{-1}$和8 784 $mg \cdot kg^{-1}$，对于样本4来说，Pb和Zn的含量分别为1 625 $mg \cdot kg^{-1}$和5 152 $mg \cdot kg^{-1}$。因此，在15个样品中，样品3和样品4的重金属污染的程度较高，主要的重金属污染源为Pb和Zn。分析结果同实际情况吻合较好，说明了主成分分析和聚类分析科学性、有效性，为矿井附近农田重金属污染治理指明了方向。

4 结语

1)评价和分析多因素影响的系统，主成分分析方法可以在保证样本大多数信息的前提下，有效的减少决策因子的个数，达到简化问题，突出重点。

2)利用主成分分析方法可以有效地揭示土壤重金属污染物的数据结构和各重金属污染物间的内在相关性及差异，并能很好地识别出矿区农田主要重金属污染物。

3)分析结果反映了不同区域的重金属污染物的组合情况和对污染负荷的贡献率，外源重金属输入对矿区农田土壤环境质量有很大的影响。

4)综合主成分值的聚类分析，可以对样品进行科学的分类，分类结果表明：Pb和Zn是样本集的主要重金属污染源，样品3和样品4的采样区是需要重点治理的区域。

参考文献

[1] 赵彦锋，史学正，于东升．工业型城乡交错区农业土壤Cu，Zn，Pb和Cd的空间分布及响因素研究[J]．土壤学报，2007，44(2)：227－234

[2] Bloemen M L，Markert B，Lieth H．The distribution of Cd，Cu，Pb and Zn in top soils of Osnabrck in relation to land use[J]．The Science of the Total Environment，1995，166：137－148

[3] 李娟娟，马金涛，楚秀娟．应用地积累指数法和富集因子法对铜矿区土壤重金属污染的安全评价[J]．中国安全科学学报，2006，16(12)：136－142

[4] 高吉喜，段飞舟，香 宝．主成分分析在农田土壤环境评价中的应用[J]．地理研究，2006，25(5)：836－842

[5] 宋桂杰，田小娟．基于主成分分析法的房地产投资环境分析[J]．扬州大学学报(自然科学版)，2006，9(4)：69－72

[6] Polat K，Güneş S．An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease[J]．Digital Signal Processing，2007，17(4)：702－710

[7] Ahn Hyunchul，Choi Eunsup，Han Ingoo．Extracting underlying meaningful features and canceling noise using independent component analysis for direct marketing[J]．Expert Systems with Applications，2007，33(1)：181－191

[8] Tan M H，Hammond J K．A non-parametric approach for linear system identification using principal component analysis [J]．Mechanical Systems and Signal Processing，2007，21(4)：1576－1600

(下转192页)