主成分分析在农田土壤环境评价中的应用

高吉喜[1]，段飞舟[2]，香 宝[1]
（1. 中国环境科学研究院生态所，北京 100012； 2. 清华大学环境科学与工程系，北京 100084）

摘要：本文尝试利用主成分分析方法对农田土壤污染物进行识别，并对土壤环境质量进行分级。结果表明，利用主成分分析可以有效地识别土壤污染物中的主要成分，揭示土壤污染物的数据结构和相互间的关系。主成分分析方法可用于定量化的土壤复合污染研究或对历史数据较为缺乏的地区进行土壤环境质量评价。在污染物检测指标数量较大时，可以在一定程度上简化农田土壤重点污染物的定量化识别过程。

关 键 词：主成分分析；农田土壤；污染物；评价
文章编号：1000-0585(2006)05-0836-07

主成分分析是在一组变量中找出其方差和协方差矩阵的特征量，将多个变量通过降维转化为少数几个综合变量的统计分析方法。由于其在对高维变量系统进行最佳的综合与简化、客观地确定各个指标的权数和避免主观随意性方面的突出特点，已经被引入土地资源的开发与保护[1~3]、环境脆弱性与环境退化研究[4,5]等诸多研究领域。与模糊综合评判法、灰色聚类法、综合指数法、神经网络等环境质量的定量评价方法相比，主成分分析方法具有能够减少原始数据信息损失、简化数据结构、避免主观随意性等优点[6,7]，在水、土壤等环境介质中的污染物评价研究中均有应用[8~10]。

近年来，土壤的污染源确认及污染物分布特征的定量评价开始受到关注，不同来源污染物在土壤中分布的特点及其相互间关系的定量化研究成为环境研究的重要课题[11~13]。传统的统计学和空间分析方法虽然已经被成功地用于污染物含量和分布的评价研究，但是这些方法都需要大量历史记录和监测数据的支持。主成分分析方法的突出特点可以揭示土壤污染物数据的结构和污染物的内在联系，描述土壤污染物分布的主要过程[14]。对于土壤污染物成分复杂或监测数据不足的地区的土壤污染定量评价，可弥补现有方法的不足。本文利用该方法对几种不同类型农田土壤环境质量进行评价，以期为土壤污染物的定量化评价研究进行有益的尝试。

1 材料和方法

1.1 数据来源

利用中意合作典型区生态调查项目鞍山市农田土壤污染物调查数据作为主成分分析的土壤环境质量评价素材，土壤样品包括清灌稻田、污灌稻田、旱田、菜田、温室及对照区 6 个不同类型农田，样本总数为 45 个。分析指标主要考虑国家颁布的土壤环境质量标准

收稿日期：2005-10-07；**修订日期**：2006-03-06
基金项目：国家环保总局“中意合作典型区生态环境调查”项目
作者简介：高吉喜（1964-），男，内蒙古呼和浩特人，研究员，博士。主要从事区域生态学、可持续发展等领域的研究。E-mail：gaojx@craes.org.cn。

中的几种重点识别重金属和有机农药污染物（As、Hg、Cu、Pb、Cd、Cr、DDT、BHC），样品采集和分析严格按照 GB5618－1995 的相关要求进行（表 1）。

表 1　不同耕作类型农田土壤污染物含量（mg/kg）

Tab. 1　Concentrations of soil contaminants in different agricultural lands（mg/kg）

农田类型	样本数	As	Hg	Cu	Pb	Cd	Cr	DDT
菜地	10	6.5	0.22	20.34	28.16	0.205	41.06	0.102
大棚	8	6.24	0.22	27.24	26.61	0.404	49.3	0.165
对照	6	7.22	0.21	20.3	26.48	0.176	36.95	0.118
旱田	3	7.6	0.23	18.43	27.93	0.148	38.17	0.038
清灌	3	9.91	0.08	22.67	25.27	0.168	40.67	0.004
污灌	15	8.38	0.42	30.54	39.63	0.347	47.49	0.008

1.2　数据处理

主成分分析的基本原理是：设有 n 个相关变量 X_i（i=1，2，…，n）组合成 n 个独立变量 Y_y（i=1，2，…，n），使得独立变量 Y_i 的方差之和等于原来 n 个相关变量 X_i 的方差之和，并按方差大小由小到大排列。把 n 个相关变量的作用看作主要由为首的几个独立变量 Y_i（i=1，2，…m）（m＜n）所决定，于是 n 个相关变量就缩减成 m 个独立变量 Y_i，Y_i（i=1，2，…m），也就是主成分（principle component）。通过降维产生的新变量能够在不损失原有信息的情况下，使原有变量所代表的信息更集中、更典型的体现出来。在对土壤环境质量进行评价时可以利用主成分分析的这些特点提取主要的污染因子，并利用主成分得分进行土壤质量评级。

数据处理主要包括数据标准化，由标准化后的数据求协方差矩阵，计算特征方程中所有特征值并根据特征值累计比例确定主成分的数量，计算主成分载荷值和主成分得分，以及进行主成分评分等。本文主成分分析过程采用 SPSS 软件的相关分析模块进行处理，具体步骤参见文献[15,16]。

2　结果和分析

2.1　主成分识别

主成分识别是以土壤污染物含量作为原变量，通过计算变量方差和协方差矩阵的特征量，将多个变量通过降维转化为少数几个综合变量，即将土壤污染物的信息进行了集中和提取，使我们能够从众多土壤污染物中识别出起主导作用的成分。由于数据中各污染物的量纲不同，各变量的作用难以直接比较，在计算时需对变量数据进行标准化处理。

表 2 是各污染物含量的总方差分解表，可以看出第一、第二主成分特征值占总方差的

表 2　观测指标总方差分解表

Tab. 2　Total variance explained of contaminant index

主成分	初始特征值及贡献率		
	特征值	贡献率%	累计贡献率 %
1	59.065	76.584	76.584
2	14.951	19.385	95.969
3	2.531	3.281	99.25
4	0.578	0.75	100

百分比已经大于 95%，即前两个主成分已经对 8 个监测指标所涵盖的大部分污染物信息进行了概括，其中第一主成分携带的信息最多，达到 76%以上，第一、第二主成分的累计贡献率达到 95.969%。主成分 3 和 4 对总方差的贡献很小，为了以尽可能少的指标反映尽量多的信息，选取前 2 个因子作为主成分，代表主要的土壤污染物指标。

2.2 主要污染物识别分析

主要污染物识别是通过土壤污染物对主成分的贡献率即主成分载荷进行分析，载荷大的即可认为是重要污染因子。表 3 是各变量对应于两个主成分的荷载值，载荷值反映的是主成分与变量的相关系数，可以据此写出主成分载荷表达式：

$$\text{第一主成分} = 0.967\text{Cu} + 0.888\text{Cr} + 0.860\text{Cd} + 0.815\text{Pb} + 0.748\text{Hg} - 0.108\text{BHC} - 0.132\text{DDT} + 0.023\text{As}$$

$$\text{第二主成分} = 0.157\text{Cu} + 0.443\text{Cr} + 0.435\text{Cd} - 0.576\text{Pb} - 0.502\text{Hg} + 0.787\text{BHC} + 0.612\text{DDT} - 0.284\text{As}$$

表 3 主成分载荷矩阵

Tab. 3 Component matrix

	第一主成分	第二主成分
Cu	0.967	0.157
Cr	0.888	0.443
Cd	0.86	0.435
Pb	0.815	−0.576
Hg	0.748	−0.502
BHC	−0.108	0.787
DDT	−0.132	0.612
As	0.023	−0.284

表 4 旋转后主成分载荷矩阵

Tab. 4 Rotated component matrix

	第一主成分	第二主成分
Cr	0.976	−0.176
Cu	0.972	0.121
Cd	0.948	−0.176
BHC	0.117	−0.786
Pb	0.621	0.781
Hg	0.578	0.691
DDT	0.045	−0.625
As	−0.058	0.279

用这 2 个因子代替 8 个原始变量，已经概括了绝大多数土壤污染物的信息。但由于每个因子中各原始变量的系数差别不明显，需要利用方差最大旋转对因子荷载矩阵进行旋转，将因子中各变量的系数向最大和最小转化，使每个因子上具有最高载荷的变量数最少，以使得对因子的解释变得容易。

表 4 是旋转后的主成分载荷矩阵，由于不同主成分对应的各变量的系数向最大和最小转化，使每个主成分上具有最高载荷的变量数最少，旋转后的荷载系数矩阵中各变量对两个主成分的荷载系数差别比较明显。可以看出，第一主成分以 Cr、Cu、Cd 为主的重金属贡献最大，第二主成分中 BHC、DDT 和重金属 Pb、Hg 的贡献较大。表中主成份载荷的正负可以反映出污染物的复合性，在主成分载荷图中表现为对斥因子（图 1）。如 As 与 Cd、Cr 在载荷图中的位置反映出它们是不同的污染因子，另外该地区农田土壤不受它们的共同污染，即某种土壤中 Cd 和 Cr 含量较高，但其对斥元素的含量却相对较低[17]。

利用旋转后的因子载荷生成的载荷散点图可以直观地看出决定因子的变量（图 1）。图中横坐标和纵坐标分别代表提取出的第一主成分和第二主成分，变量与原点的距离反映其因子载荷，位于坐标轴原点远端的变量具有较大的因子载荷，位于原点近端的变量具有较小的因子载荷。

从图 1 和表 4 可以看出，第一主成分中重金属污染物 Cd、Cr、Cu 载荷最高，第二主

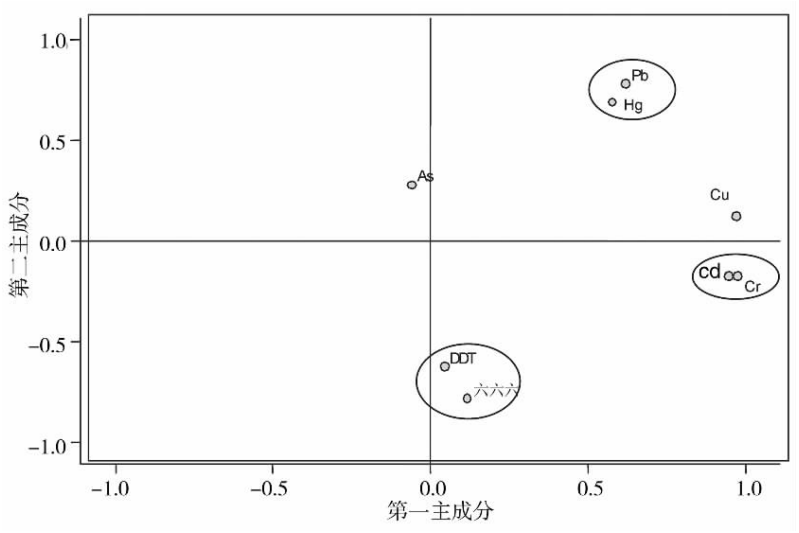

图 1　因子载荷散点图

Fig. 1　Component load plots in rotated space

成分主要受到有机农药 DDT、BHC 的影响。Pb 和 Hg 虽然不是主要的污染因子，但对第一、第二主成分均有一定的贡献。通过主成分分析，农田土壤污染物的组成结构和贡献率被确认出来。可以看出，外源重金属输入对当地农田土壤环境质量的影响高于有机农药残留，是当地农田土壤环境质量的主要影响因子。

对变量进行的相关关系检验，可以进一步反映土壤污染物间的相关关系（表 5）。

表 5　污染物相关系数矩阵

Tab. 5　Correlation coefficient matrix of soil contaminants

	As	Hg	Cu	Pb	Cd	Cr	DDT	BHC
As	1. 00							
Hg	−0. 24	1						
Cu	0. 10	0. 595	1					
Pb	0. 13	0. 919**	0. 685	1				
Cd	−0. 34	0. 537	0. 882*	0. 46	1			
Cr	−0. 17	0. 469	0. 902*	0. 477	0. 964**	1		
DDT	−0. 865*	−0. 106	−0. 077	−0. 435	0. 35	0. 172	1	
BHC	−0. 41	−0. 348	−0. 063	−0. 517	0. 294	0. 305	0. 508	1

注：* 置信区间为 95%时相关性显著； 置信区间为 99%时相关性显著**

可以看出，重金属 Hg−Pb、Cd−Cu、Cr−Cu、Cr−Cd 间具有很强的相关性，这在一定程度上反映了几种土壤重金属污染物的同源性、差异性以及在农田土壤中的组合情况。尽管 BHC 和 DDT 类农药在我国已停用多年，但在土壤中仍有一定残留，二者的相关系数达到 0. 508，反映出该区域过去的农药使用情况。这两种化学物质属于人工合成物

质，其含量多少与其他元素没有内在的联系，它们与重金属元素的相关性亦不显著。BHC 和 DDT 的施用大多在旱田和菜地，而污灌区有大量的外源重金属污染物随着灌溉用水输入，因此农药与绝大多数重金属呈负相关关系可以在一定程度上反映出土地利用状况和耕作方式对土壤污染物含量的影响。

2.3 土壤质量分级结果

由于主成分得分可以反映观测量的情况，并根据主成分得分情况进行排序，得分较低观测值含有较少的信息，得分最多的观测值包含最多的信息。在进行土壤环境评价时可以利用主成分得分，利用不同类型农田土壤污染物的主成分得分来对农田土壤环境进行排序。表 6 为各主成分的得分系数，根据得分系数可以计算每个观测值在各个污染指标上的得分数，并据此进行进一步的分析。主成分得分（FAC）的表达式为：

表 6 主成分得分系数

Tab. 6 Score coefficients of component

	第一主成分	第二主成分
As	−0.009	0.033
Hg	0	0
Cu	0.409	−0.12
Pb	0.069	1.149
Cd	0	0
Cr	0.572	−0.609
DDT	0	0
BHC	0	0

$$FAC\text{-}1 = -0.009Cu + 0.000Cr + 0.000Cd + 0.069Pb + 0.000Hg - 0.000BHC - 0.000DDT - 0.009As$$

$$FAC\text{-}2 = -0.120Cu - 0.609Cr + 0.000Cd + 1.149Pb - 0.000Hg + 0.000BHC + 0.000DDT + 0.033As$$

将不同土壤污染物浓度值与相应主成分得分系数的乘积相加得到土壤污染物的主成分得分（表 7）。主成分得分可以反映污染物对土壤的综合作用和土壤环境质量状况，得分越低代表土壤污染物含量越低，即土壤环境质量越好。从表中 6 种类型的农田土壤污染物总得分排序结果可以看出，清灌稻田的土壤环境质量最高，以下依次是对照区、旱田、大棚和菜田，利用污水进行灌溉的污灌稻田土壤综合得分最高，反映出污灌稻田污染物含量较高，土壤环境质量较差。

表 7 不同类型农田污染物主成分得分排序

Tab. 7 Component score coefficient and range of soil contamination of agricultural lands

农田类型	FAC−1	FAC−2	FAC−1+ FAC−2	总得分排序
清灌	34.19	1.87	36.06	1
对照	31.20	5.73	36.93	2
旱地	31.23	6.89	38.12	3
大棚	41.12	−2.51	38.61	4
菜地	33.69	5.12	38.81	5
污灌	42.31	13.23	55.54	6

3 结论

利用主成分分析有效地揭示土壤污染物的数据结构和土壤污染物间的内在相关性及差异性，并很好地识别出土壤污染物的主要成分。分析结果基本上反映了不同耕作类型下土壤污染物的组合情况及对污染负荷的贡献率，可以看出外源重金属输入对当地农田土壤环

境质量的影响高于有机农药残留，是当地农田土壤环境质量的主要影响因子。

本文中土壤污染物指标种类数和样本量略显不足，可能会在一定程度上影响分析的精度，不能充分反映主成分分析在处理大量数据信息时的效果，有待在今后的研究中加以完善。另外，利用主成分分析法得到的土壤环境质量排序更多的是反映不同类型土壤在污染物含量上的差异性，在对土壤环境进行评价时不能完全取代以土壤环境质量标准为依据的评价方法。尽管如此，主成分分析方法在对历史监测数据不足或存在土壤复合污染的区域进行土壤环境评价和重点污染物定量识别方面仍具有一定的优越性，可以简化农田土壤重点污染物的定量化识别过程，是较为有效的土壤环境污染定量评价工具。

致谢：本文数据采集得到辽宁省环境监测中心站及鞍山市环境监测站协助，在此谨致谢意。

参考文献：

[1] 王秀红，何书金，张镱锂，等．基于因子分析的中国西部土地利用程度分区．地理研究，2001，20(6)：731～738.

[2] 王世岩，杨永兴，杨波．三江平原典型湿地土壤温度变化及其影响因子分析．地理研究，2003，22(3)：389～396.

[3] 张鹏飞，田长彦，卞卫国，等．克拉玛依农业开发区土壤质量评价指标的筛选．干旱区研究，2004，21(2)：166～170.

[4] 黄淑芳．主成分分析及 MAPINFO 在生态环境脆弱性评价中的应用．福建地理，2002，17(1)：47～49.

[5] 何尧启．主成分分析在喀斯特土壤环境退化研究中的初步运用——以贵州麻山地区紫云县宗地乡为例．贵州师范大学学报(自然科学版)，1999，17(1)：12～19.

[6] 傅湘，纪昌明．区域水资源承载能力综合评价—主成分分析方法的应用．长江流域资源与环境，1999，l8(5)：168～173.

[7] 冯利华．环境质量的主成分分析．数学的实践与认识，2003，33(8)：32～35.

[8] 蔡启铭，高锡芸，陈宇炜，等．太湖水质的动态变化及影响因子的多元分析．湖泊科学，1995，l7(2)：97～106.

[9] 卢瑛，龚子同．南京城市土壤重金属含量及其影响因素．应用生态学报，2004，15(1)：123～126.

[10] 林小苹，黄长江，林福荣，等．海水富营养化评价的主成分一聚类分析方法．数学的实践与认识，2004，34(12)：69～74.

[11] Korre A. Statistical and spatial assessment of soil heavy metal contamination in areas of poorly recorded, complex sources of pollution, Part 2: Canonical correlation analysis and GIS for the assessment of contamination sources. Stochastic Environmental Research and Risk Assessment, 1999, 13: 288～316.

[12] Critto Andrea, Carlon Claudio, Marcomini Antonio. Characterization of contaminated soil and groundwater surrounding an illegal landfill (S. Giuliano, Venice, Italy) by principal component analysis and kriging. Environmental Pollution, 2003, 122: 235～244.

[13] McGratha David, Zhang Chaosheng, Cartona Owen T. Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland. Environmental Pollution, 2004, 127: 239～248.

[14] Korre A. Statistical and spatial assessment of soil heavy metal contamination in areas of poorly recorded, complex sources of pollution, Part 1: factor analysis for contamination assessment. Stochastic Environmental Research and Risk Assessment, 1999, 13: 260～287.

[15] 王学民．应用多元分析．上海：上海财经大学出版社，2004. 232.

[16] 刘先勇，袁长迎，段宝福，等．SPSS10.0 统计分析软件与应用．北京：国防工业出版社，2002. 337.

[17] 吴聿铭．环境统计学．北京：中国环境科学出版社，1991. 426.

The application of principal component analysis to agriculture soil contamination assessment

GAO Ji-xi[1], DUAN Fei-zhou[2], XIANG Bao[1]

(1. Institute of Ecology, CRAES, Beijing 100012, China;
2. Department of Environment Science and Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Principal components analysis (PCA) is a statistical technique used to investigate the structure of a data set, in an effort to identify the procedures controlling the scores of the variables in the data. PCA produces several linear combinations of observed variables, each linear combination being a component or factor. The factors summarize the patterns of the correlations in the observed correlation matrix and can in fact be used to reproduce the observed correlation matrix. Since the number of factors is usually far fewer than the number of the observed variables, there is a considerable parsimony in factor analysis. Furthermore, when scores on factors are estimated for each subject, they are often more reliable than scores on individual observed variables. The advantages of PCA are particularly useful in soil complex contamination studies, especially in poorly recorded areas historically, and could be further used for the spatial assessment. Now PCA has been used in the fields of resource exploitation and protection, environmental degradation and quantitative soil contamination assessment.

In this paper, data structure of soil contaminations, relationships and differences of soil pollutions were discovered, and the major components of soil pollutions were identified. The result of agriculture field quality classified with component scores showed that paddy field irrigated with clean water was on the top of the six types of land, and soil environment of sewage irrigated paddy field had worst quality. The relationships with and contribution to contamination of soil pollutants were reflected well. The effect of heavy metals inputting was higher than organic pesticide, and is the major factor of soil contamination.

The study implied that PCA is advantageous in the assessment on complex soil contamination and classification of soil environmental quality, and could be used in soil pollutants identify ication and soil environment assessment as well. The method could simplify the process of major soil pollutants identification, especially in cases of complex or poorly recorded contamination.

Key words: PCA; agricultural soil; contaminations; assessment