ICS 13.020.40
CCS Z18

T/SDEPI

团　　体　　标　　准

T/SDEPI 043—2024

土壤有机污染物来源解析 主成分分析法技术指南

Technical guide for source apportionment of soil organic pollutants by Principal Component Analysis

2024- 04 - 19 发布　　2024 - 04 - 19 实施

山东省环境保护产业协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由山东省环境保护产业协会提出并归口。

本文件起草单位：山东大学、山东建筑大学、新汶矿业集团有限责任公司、济南实华工程咨询有限公司、山东鲁金环境工程有限公司、香山红叶集团有限公司。

本文件主要起草人：崔兆杰、蔺琨、任丽军、何竞宇、薛文秀、崔晓玮、刘雷、潘哲、单绍磊、田伟、张程伟、刘佳、潘汝东、张琦、贺波、赵世刚、张琰苹。

土壤有机污染物来源解析 主成分分析法技术指南

1 范围

本文件规定了土壤有机污染物来源解析主成分分析模型方法的样品采集、主成分分析模型方法和结果表达分析等内容。

本文件适用于农业用地、建设用地等开展土壤有机污染物来源解析工作。

2 规范性引用文件

本文件内容引用了下列文件或其中的条款，凡是注明日期的引用文件，仅注日期的版本适用于本标准，凡是未注明日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 26411 海水中 16 种多环芳烃的测定 气相色谱-质谱法

GB/T 28189 纺织品 多环芳烃的测定

GB/T 36197 土壤质量 土壤采样技术指南

GB/T 39107 消费品中可挥发性有机物含量的测定 静态顶空进样法

HJ 25.1 建设用地土壤污染状况调查技术导则

HJ 25.2 建设用地土壤污染风险管控和修复监测技术导则

HJ/T 166 土壤环境监测技术规范

HJ 605 土壤和沉积物 挥发性有机物的测定 吹扫捕集气相色谱/质谱法

HJ 642 土壤和沉积物 挥发性有机物的测定 顶空/气相色谱/质谱法

HJ 703 土壤和沉积物 酚类化合物的测定 气相色谱法

HJ 743 土壤和沉积物 多氯联苯的测定 气相色谱-质谱法

HJ 784 土壤和沉积物 多环芳烃的测定 高效液相色谱法

HJ 805 土壤和沉积物 多环芳烃的测定 气相色谱-质谱法

HJ 834 土壤和沉积物 半挥发性有机物的测定气相色谱-质谱法

HJ 1019 地块土壤和地下水中挥发性有机物采样技术导则

HJ 1021 土壤和沉积物 石油烃（C_{10}-C_{40}）的测定 气相色谱法术语和定义

3 术语及定义

下列术语和定义适用于本文件。

3.1

主成分分析 Principal Component Analysis（PCA）

是一种数据降维的方法，通常用于通过将数量很多的变量转换为仍包含集合中大部分信息的较少变量来降低数据集的维数。

3.2

方差贡献率 Variance Contribution Rate

是指一个主成分所能够解释的方差占全部方差的比例，这个值越大，说明主成分综合原始变量信息的能力越强。

3.3

Bartlett 球形检验 Bartlett's Test of Sphericity

是一种检验各个变量之间相关性程度的检验方法。一般在做因子分析之前都要进行 Bartlett 球形检验，用于判断变量是否适用于做因子分析。

3.4

碎石图 Scree Graph

是分析主成分特征值的线形图，用于确定主成分分析中保留的主成分数量。根据碎石检验，找到图中出现平台期的点，选择该点左边的因子为主要成分。

4 土壤有机污染物来源解析工作流程

土壤有机污染物来源解析工作流程见下图。

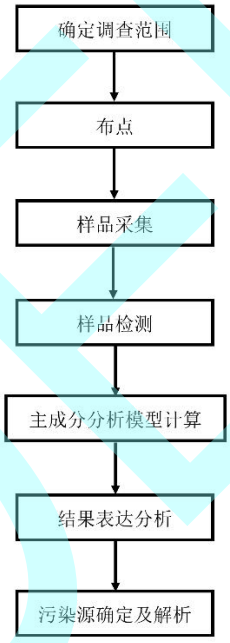

图 1 土壤有机污染物来源解析工作流程图

5 土壤有机污染物样品采集

5.1 布点

宜采用 HJ/T 166 进行布点，点位数应不少于 16 个。

5.2 样品采集

宜采用 GB/T 36197 和 HJ 1019 等技术导则进行样品采集和贮存，土壤样品的保存和流转执行 HJ 25.1、HJ 25.2 和 HJ/T 166 的相关规定。

5.3 样品检测

土壤样品有机污染物的检测方法见表 1。

表 1 土壤样品有机污染物的检测方法

测试项目	检测方法
甲苯-d8	GB/T 39107 HJ 605 HJ 642
4-溴氟苯	
二溴一氟甲烷	
苯	
甲苯	
乙苯	
间&对-二甲苯	
邻-二甲苯	
二溴氯甲烷	
1,2-二氯丙烷	
四氯乙烯	
三氯乙烯	
氯仿	
氯乙烯	
二氯甲烷	
1,2,3-三氯丙烷	
1,1,2-三氯乙烷	
1,1,2-三氯丙烷	
1,1-二氯乙烯	
氯甲烷	
反-1,2-二氯乙烯	
二溴甲烷	
1,1,1-三氯乙烷	
1,1,1,2-四氯乙烷	
1,2-二氯乙烷	
氯苯	
三氯氟甲烷	
2-甲基萘	HJ 834
2-氟苯酚	

邻苯二甲酸二甲酯	
3,3'-二氯联苯胺	
邻苯二甲酸二(2-乙基己)酯	
邻苯二甲酸二丁酯	
邻苯二甲酸二乙酯	
苯胺	
苯酚-d6	
邻苯二甲酸二正辛酯	
硝基苯-d5	
2-氟联苯	
2,4,6-三溴苯酚	
对-三联苯-d14	
六氯苯	
1,4-二氯苯	
萘	GB/T 28189 GB/T 26411 HJ 805 HJ 784
苊	
芴	
菲	
蒽	
荧蒽	
芘	
䓛	
苯并(a)蒽	
苯并(a)芘	
苯并(b)荧蒽	
苯并(k)荧蒽	
苯并(g,h,i)苝	
茚并(1,2,3-cd)芘	
二苯并(a,h)蒽	
苯酚	HJ 703
3&4-甲基苯酚	
2,4-二氯苯酚	

2-氯苯酚	
2,4,6-三氯苯酚	
2,4,5-三氯苯酚	
2,4,5,6-四氯-间-二甲苯	HJ 743
十氯联苯	
2,3',4,4',5-五氯联苯	
2',3,4,4',5-五氯联苯	
2,3,4,4',5-五氯联苯	
2,3,3',4,4'-五氯联苯	
3,3',4,4',5-五氯联苯	
2,3,3',4,4',5-六氯联苯	
2,3,3',4,4',5'-六氯联苯	
3,3',4,4',5,5'-六氯联苯	
石油烃（C_{10}-C_{40}）	HJ 1021

6 主成分分析模型（SPSS 软件）方法

6.1 模型原理

主成分分析（PCA）原理是设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量信息的统计方法。

主成分分析法是一种数学变换的方法，把给定的一组相关变量通过线性变换转化为另一组不相关的变量，这些新的变量按照方差依次递减的顺序排列。假设对 p 个变量进行了 n 次测量/观测，那么原始测量数据的矩阵表示如下：

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \text{.....（1）}$$

其中，x_i表示如下：

$$x_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{pmatrix}, i = 1,2,\ldots,p \quad \text{.....（2）}$$

引入一个新变量 F，是原来 P 个变量的线性组合，表示如下：

$$F_1 = \alpha_{11}x_1 + \alpha_{21}x_2 + \ldots + \alpha_{p1}x_p \quad \text{.....（3）}$$

$$F_2 = \alpha_{12}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{p2}x_p \quad \text{.....（4）}$$

……

$$F_p = \alpha_{1p}x_1 + \alpha_{2p}x_2 + \ldots + \alpha_{pp}x_p \quad \text{.....(5)}$$

$$\text{或者} F_i = \alpha_{1i}x_1 + \alpha_{2i}x_2 + \ldots + \alpha_{pi}x_p (i = 1,2,\ldots,p) \quad \text{.....(6)}$$

$$\text{其中，} \alpha_{1k}^2 + \alpha_{2k}^2 + \alpha_{3k}^2 + \ldots + \alpha_{pk}^2 = 1 (k = 1,2,\ldots,p) \quad \text{.....(7)}$$

新变量 F 还需满足以下要求：（1）F_i 和 F_j（$i \neq j$）不相关；（2）F_1 的方差大于 F_2，以此类推。这样 F_1 中就包含了原数据库中最多的信息，F_2 中包含的信息是除 F_1 外最多的，以此类推。经过以上计算，得到的 F_i 就是原数据信息中包含的第 i 个主成分。

6.2 主要计算流程

应用主成分分析（PCA）方法对有机污染物的来源进行解析，主要计算流程包括数据标准化、计算相关系数矩阵、计算特征值和特征向量、特征值排序、选择主成分、构建变换矩阵和数据投影等环节，见图 2。

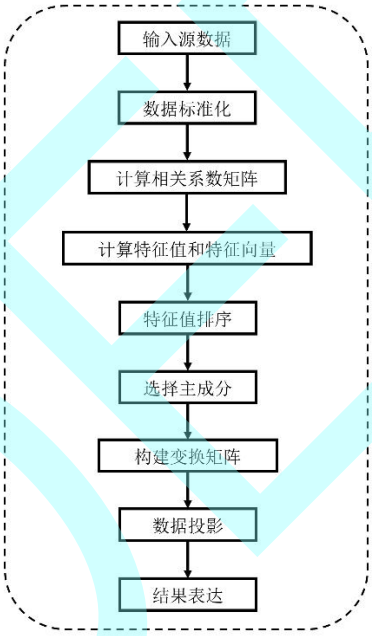

图 2 PCA 主要计算流程

6.3 数据标准化

对土壤有机污染物原始数据进行标准化处理，使得每个特征的平均值为 0，方差为 1，以消除不同特征的量纲差异。

对于一组样本资料 X，其中有 m 个观测值 x_1，x_2，…，x_m，共 n 个样。

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \quad \text{.....(8)}$$

（1）计算每一列的平均值 $\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$

（2）每一列的方差 $\sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_{ji} - \mu_i)^2$

（3）对数据进行标准化处理，$Z_{ji}=\frac{x_{ji}-\mu_i}{\sigma_i}$

得到协方差矩阵 Z：

$$Z=\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} \quad \text{.....(9)}$$

6.4 计算相关系数矩阵

根据标准化后的数据，计算相关系数矩阵。计算相关系数，$r_{ij}=\frac{\sum_{k=1}^{n} z_{ki}*z_{kj}}{n-1}$，（i, j=1, 2, …, m），得到相关系数矩阵 R：

$$R=\begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix} \quad \text{.....(10)}$$

r_{ij}即 X 矩阵的第 i 列的样本序列和第 j 列的样本序列之间的相关关系，其值为-1 到 1 之间，且 R 矩阵应该为对称矩阵，即$r_{ij}=r_{ji}$。

相关系数程度区分如下表 2 和表 3 所示：

表 2 正负相关性

相关系数 r	相关性
$r>0$	正相关
$r=0$	无关
r<0	负相关

表 3 相关性程度大小

相关系数绝对值|r|	相关性程度
1	完全相关
[0.8, 1)	高度无关
[0.5, 0.8)	中度相关
[0.3, 0.5)	低度相关
[0, 0.3)	不相关

6.5 计算特征值和特征向量

协方差矩阵Σ是实对称阵，知其特征值为非负，不妨设其特征值$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$，它们对应的正交化后的单位特征向量如下：

$$a_1=\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix};\ a_2=\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix};\ \cdots;\ a_p=\begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} \quad \text{.....(11)}$$

若原先 X 的各个列代表的指标变量，合成向量，记为$Var = [Var_1, Var_2, ..., Var_p]^T$，则有 X 的第 i 个主成分为$F_i = (a_i)^T Var = \alpha_{1i} * Var_1 + \alpha_{2i} * Var_2 + ... + \alpha_{pi} * Var_p$。

特征值是衡量主成分影响力的重要指标，指的是每个主成分坐标轴所对应的主成分变量能解释多少原始数据中的变异（即方差）。特征值表示每个特征向量的重要程度，特征向量表示数据中的主要方向。

6.6 特征值排序

将特征值按照从大到小的顺序进行排序，同时对应的特征向量也进行相应的排序。求出特征值后要按大小予以排列：$\lambda_1 \geq \lambda_2 > ... \geq \lambda_p \geq 0$。如果特征值小于 1，表示该主成分的解释力非常低，一般以特征值大于 1 为筛选主成分的标准。

6.7 选择主成分

方差贡献率越大，主成分综合原始变量信息的能力越强。方差贡献率的计算公式为：

$$\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}(i = 1,2,...,p) \quad .....(12)$$

相应的，主成分筛选中所确定的前 m 个主成分所能解释的全部方差占总方差的比例称为累计方差贡献率。其公式为：

$$\frac{\sum_{k=1}^{i}\lambda_i}{\sum_{k=1}^{p}\lambda_k}(i = 1,2,...,p) \quad .....(13)$$

第一主成分的方差贡献率最大，它能解释原始变量 X_1，X_2，…，X_p 的能力最强，第 2，第 3，…，第 p 个主成分的解释能力依次递减。

主成分数量的选取则是根据累积贡献率确定，一般要求累积贡献率达到 85%以上，这样能保证新变量能包括原始变量的绝大多数信息。

此外，土壤有机污染物主成分的个数选取有 3 个主要的衡量标准：保留的主成分使得方差贡献率达到 85%以上；保留的主成分的方差（特征值）大于 1；碎石图绘制了关于各主成分及其特征值的图形，只需要保留图形中变化最大之处以上的主成分即可。衡量标准的选择无先后顺序，达到某一个标准即可确定相应的主成分。

6.8 构建变换矩阵

将选择的 k 个特征向量按列形成一个新的矩阵，称为变换矩阵。该矩阵可以对原始数据进行线性变换，将其映射到新的 k 维特征空间。

6.9 数据投影

将原始数据通过变换矩阵进行投影，得到降维后的新数据。投影的计算方法是将原始数据与变换矩阵相乘。

7 结果表达分析

（1）KMO 取样适合度检验和 Bartlett 球形检验

这两项检验是用来判断是否可以进行主成分分析。对于 KMO 值：0.8 以上非常合适做主成分分析，0.7-0.8 之间一般适合，0.6-0.7 之间不太适合，0.5-0.6 之间表示差，0.5 以下表示极不适合；对于 Bartlett 球形检验（$p < 0.05$，严格来说 $p < 0.01$），若显著性小于 0.05 或 0.01，说明可以做土壤有机污染物主成分

分析，结果可信。

（2）方差解释表格、成分矩阵表

方差解释表格给出了提取的主成分方差解释量，每个主成分能解释的方差比例不同。特征根的值按照从大到小进行排序，一般选取方差（特征值）大于 1 的主成分，可以有效保留主成分。成分矩阵表体现了每个主成分和原始变量之间的相关系数，由成分矩阵表可得到主成分分析降维后的原始变量的线性组合。

（3）碎石图

一般选取碎石图曲线上由陡峭变为舒缓的结点前的碎石为土壤有机污染物的主成分。例如：选取图 3 中虚线以上的部分，即左侧 3 个因子作为主成分。

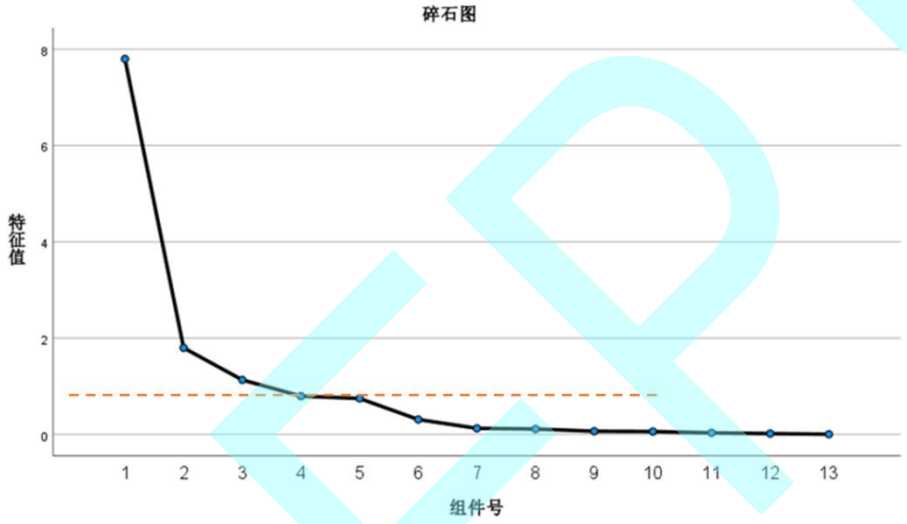

图 3 碎石图

（4）因子载荷系数表、因子载荷矩阵热力图和因子载荷象限分析

因子载荷系数表和因子载荷矩阵热力图可以分析到每个主成分中隐变量的重要性。若因子载荷系数出现负值，表示污染来源与该因子呈负相关，可忽略。热力图颜色越深说明相关性越大。因子载荷象限分析图通过将多因子降维成双主成分或者三主成分，通过象限图的方式呈现土壤有机污染物主成分的空间分布。