第27卷,第10期
2007年10月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 27, No. 10, pp2107-2110
October, 2007

不同污染源海水样本的分类模型研究

林 燕[1], 朱尔一[1*], 徐晓琴[1], LEE Frank S C[2], 王小如[2]

1. 厦门大学化学化工学院, 现代分析科学教育部重点实验室, 福建 厦门 361005
2. 国家海洋局第一海洋研究所, 山东 青岛 266061

摘 要 简要介绍了偏最小二乘建立分类模型方法的原理及优点, 将该方法应用于处理胶州湾和莱州湾的几个主要污染源附近海域各站点取得的海水样本的气-质联用全谱数据, 建立海水样品的分类模型, 以判别海水中有机污染物的来源区域。结果表明: 由于PLS法适合于处理变量数多样本数少、具有严重多重共线性数据的问题, 应用于从两类及多类海水样品气-质联用全谱数据中提取海水污染来源区域的分类信息, 得到的分类模型交叉检验相关系数达0.91以上, 结果较为理想, 可为正确判别污染源提供一个可靠的基础。另外文章采用所得模型的拟合值等一些信息作分类图的方法, 与传统PLS作图方法比较, 所得分类图更为清晰、直观, 能较好地表达回归模型的分类效果。

关键词 气-质联用; 偏最小二乘(PLS); 分类模型
中图分类号: O657.6, O657.7 文献标识码: A 文章编号: 1000-0593(2007)10-2107-04

引 言

近年来, 由于工业及经济发展, 大量废物和废水被排放入海, 导致沿岸海域污染日益严重, 因此, 建立海洋环境中各污染源的指纹图谱数据库、对污染海域进行来源识别和示踪分析, 从而对污染源进行控制是十分必要的。色谱-质谱联用技术(GC-MS)由于同时具有GC的高分离能力和MS的高鉴别能力, 在复杂混合物的分析中具有独特的优势, 目前已成为混合物分离鉴定最常用的手段[1]。GC-MS在海水污染的研究情况已有报道[2,3]。然而海水成分复杂, 得到的GC-MS数据量非常大, 采用一般数据处理方法难以得到满意的结果。目前常采用化学计量学方法中的PLS法来提取成分复杂的图谱信息, 该法可解决谱图共线问题, 有效地提取谱图信息[4]; 具有预测能力强和模型相对简单等优点[5]; 尤其当解释变量个数多、样本量少的分析体系时, 该方法很有效[6,7]。PLS由于具有较强的提取信息的能力而成为化学计量学中倍受推崇的多变量校正法, 在分析化学中得到了广泛的应用[6-10]。

目前, 采用化学计量学技术, 结合GC-MS技术, 来获取海水的分类信息等方面, 国内外的相关报道较少。本文以胶州湾和莱州湾两大海域5个站点所取的海水样本的GC-MS全谱数据为研究对象, 用PLS建立了不同来源海域海水样品的分类模型。找到了一种可以将污染情况不同的海水GC-MS数据样本很好的进行分类的方法。

1 原 理

样品首先进行GC-MS分析, 在程序升温条件下, 使样品的各组分在色谱柱中分离, 依次进入质谱仪, 可得到样品中有机物的总离子流色谱图。将GC-MS数据与样品来源或类别直接关联, 将各馏分对应的谱数据组成自变量矩阵 $\boldsymbol{X}$, 来源或类别信息组成因变量矩阵 $\boldsymbol{Y}$, 根据这两个数据矩阵建立分类模型:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E} \quad (1)$$

其中 $\boldsymbol{B}$ 为系数矩阵, $\boldsymbol{E}$ 为残差矩阵, 根据所得到的模型来判别样品的来源或类别。

本文中处理建模问题的特点是矩阵 $\boldsymbol{X}$ 中的变量多(近3 000个)、样本少(几十个)、数据具有严重多重共线性, 因此采用了一种适合处理变量多样本少的PLS方法, 具体PLS正交变换算法见文献[7]。在建模过程中采用PRESS(prediction sum of squares)判据来检验所建立的数学模型的有效性[6], PRESS定义如下:

$$\mathrm{PRESS} = \sum (y_i - \hat{y}_{i,-i})^2 \quad (2)$$

其中 $\hat{y}_{i,-i}$ 为第 i 个样本不参加建模时, 得到的模型对该样本的预报值。PRESS越小, 表示模型的预报能力越强。为便于

收稿日期: 2006-07-21, 修订日期: 2006-11-06
基金项目: 福建省自然科学基金项目(Z0513003), 海洋环境污染物被动示踪研究和国家"863"计划(2003AA635180)资助
作者简介: 林 燕, 女, 1982年生, 厦门大学化学化工学院硕士研究生 * 通讯联系人 e-mail: ryzhu@xmu.edu.cn

模型之间的比较，采用模型交叉检验(Cross validation)相关系数 CR 来衡量模型的预报准确率，CR 定义如下:

$$CR = \sqrt{1 - \mathrm{PRESS}/S_y} \quad (3)$$

其中 S_y 为变量 y 的总方差。CR 越接近于 1，说明模型越可靠。

2 结果与讨论

2.1 谱图数据处理

本实验取胶州湾海泊河入海口海域与胶州湾其它海域这两大块海域，及莱州湾龙口电厂附近、东营养殖区、龙口纸厂这 3 个海域不同站点水样各为 10，14，14，11，16 个样本，进行 GC-MS 程序升温全扫描。所得一张典型的谱图如图 1 所示，比较每张色谱图，很难发现各海域水样之间的规律性，欲根据不同海域水样中的有机物信息将各个海域水样区分开来，可采用化学计量学方法。由于 GC-MS 所得谱图的漂移和扭曲都不大，可采用不经处理的原始谱图数据进行数据预处理[11]，选取包含信息较为全面的合适谱图区间，本实验对每个样本选取质量范围 50~ 300、扫描号 1~ 2890 的谱图数据为基础数据。

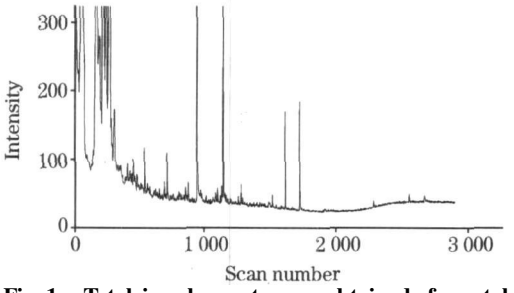

Fig 1 Total ion chromatogram obtained of coastal water of Laizhou Bay near power plant

2.2 两类样本的分类模型

胶州湾的污染源有很多，海泊河入海口是其主要污染源之一，本研究欲将海泊河入海口海域水样与胶州湾其它海域水样区分开来。首先将谱数据转化成一个以 24 个样品的平均质量数为横坐标，以 2890 个扫描数即变量数为纵坐标的多因素数据矩阵 $\boldsymbol{X}$。而目标变量矩阵 $\boldsymbol{Y}$ 对于两类样本的分类只需取一列矢量，其中对应于海泊河入海口水样 y 值取 0，胶州湾其它海域的水样 y 值取 1。

首先采用聚类分析方法处理各类别样本的 GC-MS 数据，期望通过各样本间的相似度将同一海域不同站点的水样聚在一起，而不同海域间的水样可以很好地分开，但聚类分析结果说明此法得不到满意的分类结果。采用主成分分析法对 $\boldsymbol{X}$ 矩阵中的数据处理得到的第一主成分和第二主成分构成分类图的方法，也得不到较理想的分类。

采用 PLS 法对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 矩阵的数据处理，对变量矩阵 $\boldsymbol{X}$ 进行 PLS 正交分解，选取使 PRESS 值接近最低的 7 个隐变量建模，得到回归分类模型 CR 值为 0.91，结果较为理想。采用传统 PLS 利用正交分解得到的第一和第二隐变量 $\boldsymbol{t}_1 - \boldsymbol{t}_2$ 作分类图的结果如图 2，该图能说明一定的分类效果，但此法只用了两个隐变量，不能完全反映分类模型的分类结果。因此本文用模型的拟合值 $\hat{\boldsymbol{y}}$ 和与之正交的第一主成分 $\boldsymbol{t}_1$ 作分类图，分类结果如图 3 所示($\boldsymbol{t}_1$ 的计算详见附录)，该图说明得到的模型有较好的分类效果。

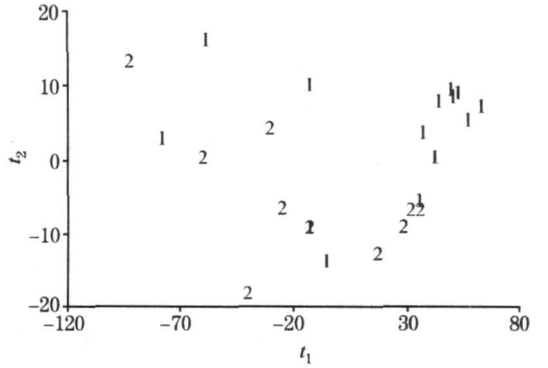

Fig 2 The PLS discrimination classification of different areas in Jiaozhou Bay ($t_1 - t_2$)

1: Sea areas in Jiaozhou Bay other than Haibo river estuary;
2: Haibo river estuary

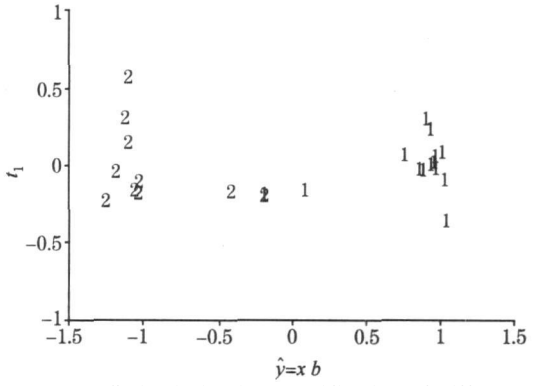

Fig 3 The PLS discrimination classification of different areas in Jiaozhou Bay ($t_1 - \hat{y}$)

Table 1 Construction of dependent variable Y for the three classes of samples

样品	站点	目标变量		
		$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{y}_3$
1	电厂附近海域	1	0	0
…	…	…	0	0
14	电厂附近海域	1	0	0
15	造纸厂附近海域	0	1	0
…	…	0	…	0
30	造纸厂附近海域	0	1	0
31	养殖区海域	0	0	1
…	…	0	0	…
41	养殖区海域	0	0	1

2.3 三类样本的分类模型

取莱州湾龙口电厂附近、东营养殖区、龙口纸厂 3 个海域不同站点各 14，11，16 个水样，与 2.2 节同法对样品进行数据预处理，生成一个维数为 41×2890 的多维数据矩阵 $\boldsymbol{X}$。

对于莱州湾三类海水的分类，目标变量矩阵 $\boldsymbol{Y}$ 有 3 列矢量(如表 1)。

数据矩阵 $\boldsymbol{X}$ 进行 PLS 回归分析得到预报模型时，选取使 PRESS 最小或接近最小的隐变量建模。其中 $\boldsymbol{y}_1$ 模型中所取隐变量数为 11，模型 CR 达 0 976; $\boldsymbol{y}_2$ 模型中所取隐变量数为 12，模型 CR 达 0 973; $\boldsymbol{y}_3$ 模型中所取隐变量数为 8，模型 CR 达 0 98。采用 3 列矢量模型的拟合值 $\hat{\boldsymbol{y}}_1$，$\hat{\boldsymbol{y}}_2$ 和 $\hat{\boldsymbol{y}}_3$ 的不同组合可构成不同区域海水样品的判别分类图，其中将 $\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_2$ 作图如图 4 所示，图中显示出较好的分类效果。

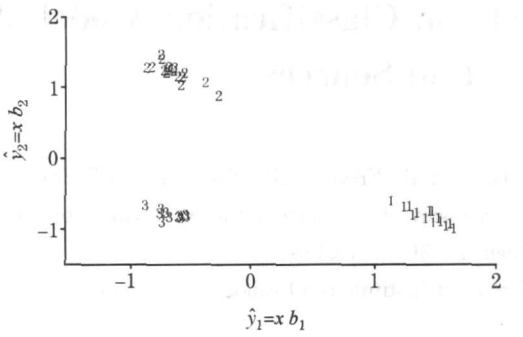

Fig 4　The PLS discrimination classification of three different areas in Laizhou Bay

1: Power plant; 2: Paper mill; 3: Breed aquatics district

3　结　论

本文中 GC-MS 海水数据具有变量数多，样本数少和严重多重共线性的特点，采用主成分分析、聚类分析及一般的多元线性回归数据处理方法都不能得到较好的分类结果。而采用 PLS 建立分类模型的方法可达到较理想的分类效果。与传统 PLS 法利用 $\boldsymbol{t}_1 - \boldsymbol{t}_2$ 作分类图比较，本文采用模型的拟合值 $\hat{\boldsymbol{y}}_i$ 之间或与之正交的第一主成分 $\boldsymbol{t}_1$ 等作分类图的方法，能较好的表达分类模型的分类效果。

从本文建立 2 类和 3 类不同来源海水样品的分类模型来看，模型 CR 值均达 0 91 以上，可见用该方法建立不同区域海水样品的 GC-MS 全谱数据的分类模型，预报稳定性较高。研究表明，利用 PLS 建立分类模型的方法，可为正确判别污染源提供一个可靠的基础，是处理 GC-MS 等获取的变量数很大的数据的一种有效方法。

4　附　录

$\boldsymbol{t}_1$ 的迭代计算公式

一般主成分分析中的 $\boldsymbol{t}_1$ 可用下列计算公式，

$$\boldsymbol{r}_1 = \frac{1}{\lambda}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{r}_1 \text{(迭代计算)} \qquad (1)$$

$$\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{r}_1 \qquad (2)$$

本文中要求 $\boldsymbol{t}_1$ 与 $\hat{\boldsymbol{y}}$ 正交，即 $\boldsymbol{t}_1^{\mathrm{T}}\hat{\boldsymbol{y}} = 0$。在该约束条件下，以上(1)式可改为，

$$\boldsymbol{r}_1 = \frac{1}{\lambda}\left(\boldsymbol{I} - \frac{\boldsymbol{p}\boldsymbol{p}^{\mathrm{T}}}{\boldsymbol{p}^{\mathrm{T}} - \boldsymbol{p}}\right)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{r}_1 \text{(迭代计算)} \qquad (3)$$

其中 $\boldsymbol{I}$ 为单位矩阵，$\boldsymbol{p} = \boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{y}}$ 为一矢量，由(3)式和(2)式可求得本文中的 $\boldsymbol{t}_1$ 的迭代计算公式。

参　考　文　献

[1] LIN Ping, SANG Wen-qiang, LI Jun, et al(林　平, 桑文强, 李　军, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(10): 1599.

[2] YAO Zi-wei, JIANG Gui-bin, CAI Ya-qi, et al(姚子伟, 江桂斌, 蔡亚岐, 等). Chinese Science Bulletin(科学通报), 2002, 47(15): 1196.

[3] YANG Qing-xiao, XU Jun-ying, ZHANG Hai-yun, et al(杨庆霄, 徐俊英, 张海云, 等). Marine Science Bulletin(海洋通报), 1987, 6(1): 88.

[4] ZHENG Yong-mei, ZHANG Jun, CHEN Xing-dan, et al(郑咏梅, 张　军, 陈星旦, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(9): 1047.

[5] ZHANG Lin, ZHANG Li-ming, LI Yan, et al(张　琳, 张黎明, 李　燕, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(10): 1610.

[6] ZHU Er-yi, YANG Peng-yuan(朱尔一, 杨芃原). Technology and Application of Chemometrics (化学计量学技术及应用). Beijing: Science Press(北京: 科学出版社), 2001. 92

[7] ZHU Er-yi(朱尔一). Computers and Applied Chemistry(计算机与应用化学), 2005, 22(8): 639.

[8] ZHUO Lin, ZHU Er-yi, WANG Qiao-er, et al(卓　林, 朱尔一, 王巧娥, 等). Chinese Traditional and Herbal Drugs(中草药), 2005, 36(supplement): 204.

[9] ZHU Er-yi(朱尔一), Barnes RM. Journal of Chemometrics, 1995, 9: 363.

[10] ZHU Er-yi, et al(朱尔一, 等). Chemical Journal of Chinese Universities(高等学校化学学报), 1997, 18: 212.

[11] ZHAO Li-na, LIU Ze-long, TIAN Song-bo(赵丽娜, 刘泽龙, 田松柏). Chinese Journal of Analytical Chemistry(分析化学), 2005, 33(1): 90.

Study on Classification Model of Seawater Samples with Different Pollution Sources

LIN Yan[1], ZHU Er-yi[1*], XU Xiao-qin[1], LEE Frank S C[2], WANG Xiao-ru[2]

1. Department of Chemistry, the Key Laboratory of Analytical Sciences of the Ministry of Education, Xiamen University, Xiamen 361005, China
2. The First Institute of Oceanography, State Oceanic Administration, Qindao 266061, China

Abstract In the present article the principle and advantages of the method to build classification model by partial least squares are briefly introduced. The method was applied to deal with the seawater data obtained from the primary polluted sea area of Jiaozhou bay and Laizhou bay by GC-MS. The classification models have been built for seawater samples from different contaminated areas. The results indicate that PLS is very suitable for dealing with the problems with the data sets that contain many variables and few samples and have serious co-linearity. Accurate classification models can be built by use of PLS to get the classification information of pollution sources from two or many kinds of polluted seawaters data sets from GC-MS. The cross validation relativities of the model comes to over 0.91. This result is approving, which can provide a reliable foundation for distinguishing pollution sources correctly. Moreover, compared with the traditional method, the classification figures constructed by model's $\hat{y}_i$ in the article are more clear and intuitive, and can express the model's discrimination effect better.

Keywords GC-MS; Partial least squares(PLS); Classification model

(Received Jul. 21, 2006; accepted Nov. 6, 2006)

* Corresponding author