DOI: 10.11758/yykxjz.20140503 http://www.yykxjz.cn/

偏最小二乘回归法在海洋初级生产力影响因子分析中的应用*

关炯晖 杨振杰

(江门市海洋与渔业环境监测站 江门 529000)

摘要 叶绿素 a 浓度是浮游植物生物量的重要指标，能反映出海洋初级生产力的状况。从海洋监测数据中，分析影响叶绿素 a 浓度的理化环境因子，尝试找出海洋初级生产力的主导因子。将数据标准化，运用变量投影重要性分析和 T^2 椭圆图等辅助分析技术进行成分提取，进行偏最小二乘回归分析。实验结果表明，在贫营养化的川岛外水域，对叶绿素 a 影响能力最大的因子为磷酸盐和无机氮，其次为酸碱度、盐度和水温，溶解氧影响能力最小。但在富营养化的沙堤口水域，影响能力最大的因子为化学需氧量和溶解氧，水温也成为影响海洋初级生产力的重要因子，而无机氮却成为非重要因子。从以上结果可见，运用偏最小二乘回归法对叶绿素 a 影响因子进行探讨，既能在监测数据较少的条件下建立数学模型，又能解决多元回归分析时各理化环境因子之间存在多重相关性的问题，具有较高的相关分析精度。

关键词 偏最小二乘回归法；海洋初级生产力；理化环境因子；富营养化；贫营养化

中图分类号 S932 文献标识码 A 文章编号 1000-7075(2014)05-0019-07

海洋初级生产力是海洋生态系统能量活动的起点，主要的初级生产是光合作用，植物细胞内叶绿素含量和光合作用产量之间存在一定的相关性，制造的有机物可以用叶绿素含量来表示(沈国英等，2002)。利用海洋环境监测数据，分析海洋中各理化因子对叶绿素的影响，尝试找出海洋初级生产力的主导因子，可以为海洋生态环境和渔业资源的可持续发展提供理论依据，对赤潮预警也有一定的指导意义。

已有学者对海洋初级生产力的空间分布特征(柯志新等，2013)、环境影响因子进行研究(孙耀等，1996；高会旺等，2001；孙雪梅等，2013)，但都以简单的相关分析法进行讨论，忽略了各因子之间存在多重相关性。海洋浮游植物的生长与理化因子存在密切联系，各理化因子之间又存在相互的影响，如海水水温不仅影响水中溶解氧的含量，还通过海水的层化现象影响表层水中营养盐含量，所以单独讨论某因子对海洋初级生产力的贡献也是不够全面、客观的。

在分析环境变量的影响因素时，常用的数学分析方法如相关分析法、灰色关联度分析法、主成分分析法、逐步回归法和影响因子分析法(张丽旭等，2004)等，都存在一定的不足：前两种分析法没有考虑因子之间的多重相关性，第三种和第四种分析法虽然考虑了多重相关性，但程序在运算时需要样本量大，具有典型分布和计算量大等限制(Johnson，2008；王黎明等，2008)，影响因子分析法有考虑多重相关性，还能对因子进行影响大小的排序，但不能给出变量要素影响因子的数学模型，缺乏直观性。偏最小二乘回归法在遥感监测(李小斌等，2007；张正建等，2009)、水环境非点源负荷的估计(陈馨等，2013)等有较多应用，本研究将该法运用于样本容量小、影响因子多并且存在多重相关的海洋环境变量要素分析中，建立回归数学模型。此模型可以提取影响解释性最强、最能概括自变量集合中信息的综合变量，提高分析精度，更全面地探讨影响海洋初级生产力的主导因子。

* 关炯晖，E-mail: guanjh1018@163.com

收稿日期: 2014-03-04，收修改稿日期: 2014-05-21

1 材料

选取广东省江门市川岛海域两个监测站位。站位 A: 川岛外(21°30′0.0″N；112°30′0.0″E)；站位 B: 沙堤口(21°36′0.0″N；112°43′0.0″E)，采样站位见图 1。从 2010 年 3 月起，至 11 月，每月 1 次水环境监测。监测指标包括：水温(T)、盐度(S)、酸碱度(pH)、溶解氧(DO)、化学需氧量(COD)、磷酸盐(PO_4^{3-})、无机氮、石油类和叶绿素 a(Chl-a) 9 个环境因子。取深度为 0.5 m 的表层海水样品，按《海洋监测规范》(GB17378.4、7-2007)中分析方法进行测定，共得到 162 个数据(两个站位×9 个指标×9 个月份)。

以 Chl-a 的含量值构成因变量数据表 $y=[y]_{9\times1}$，以 T、S、pH、DO、COD、PO_4^{3-}、无机氮、石油类等含量值构成自变量数据表 $X=[x_1, x_2, \ldots, x_8]_{9\times8}$，样本数 $n=9$，进行偏最小二乘回归分析。

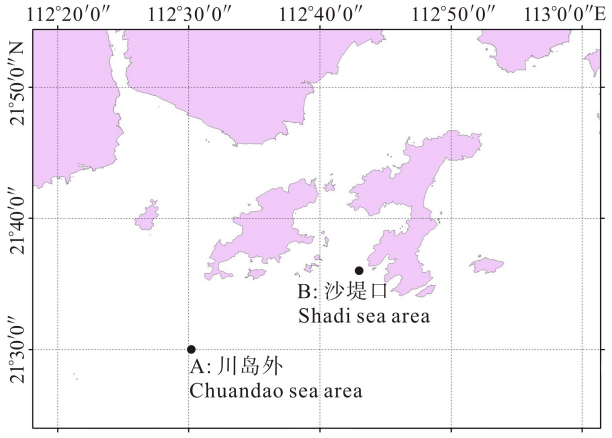

图 1 采样站位

Fig. 1 Sampling stations

2 方法

2.1 建模原理

设已知因变量 y 和 k 个自变量 $x_1, x_2, \ldots, x_k$，样本数为 n，构成数据表 $X=[x_1, x_2, \ldots, x_k]_{n\times k}$ 和 $y=[y]_{n\times1}$。在 X 中提取成分 t_1，t_1 是 $x_1, x_2, \ldots, x_k$ 的线性组合，要求 t_1 尽可能大地携带 X 中的变异信息，且与 y 的相关程度最大。实施 y 和 X 对 t_1 的回归，如果此时回归方程已经达到满意的精度，则算法停止；否则，将利用 X 被 t_1 解释后的残余信息以及 y 被 t_1 解释后的残余信息进行第二主成分 t_2 的提取，继续实施 y 和 X 对 t_1，t_2 的回归。如此反复，直到能达到一个较满意的精度为止(可用交叉有效性确定，即预测误差 PRESS 最小的原则)。若最终对 X 共提取了 h 个成分 $t_1, t_2, \ldots, t_h(h\leq n)$，偏最小二乘回归将施行 y 对 $t_1, t_2, \ldots, t_h$ 的回归，由于 $t_1, t_2, \ldots, t_h$ 都是 $x_1, x_2, \ldots, x_k$ 的线性组合，最后可表达成 y 对原变量 X 的回归方程。

2.2 辅助分析技术

2.2.1 自变量与因变量的相关关系分析　判断自变量集合 X 与因变量集合 y 之间是否存在较强的相关关系是检验是否可以建立 y 对 X 的线性回归的基本条件。在自变量 X 中提取主成分 t 与因变量 y 中提取的主成分 u，以 t/u 绘制平面图，如果图中明显观察到 t_h 与 u_h 之间存在线性关系，则说明 X 与 y 有显著的相关关系，这时采用偏最小二乘回归方法建立 y 对 X 的线性模型才会是比较合理的。

2.2.2 特异样本的判别　定义第 i 个样本点对第 h 主成分 t_h 的贡献率为 T_{hi}^2，它是一个椭圆，如果样本点都落在椭圆内，则认为没有特异点；反之落在椭圆外的点认为是特异点。

2.2.3 变量投影重要性分析　在偏最小二乘回归分析中，自变量对因变量的解释能力是以变量投影重要性指标 VIP 值来测度的。

$$\mathrm{VIP}_j=\sqrt{\frac{k}{Rd\left(y;t_1,\cdots,t_m\right)}\sum_{h=1}^{m}Rd\left(y;t_h\right)w_{hj}^2}$$

式中，W_{hj} 是轴 W_h 的第 j 个分量，用于衡量 x_j 对构造 t_h 主成分的贡献大小。$Rd(y; t_h)$和 $Rd(y; t_1, t_2, \ldots, t_m)$分别是 y 由 t_h 和 $t_1, t_2, \ldots, t_m$ 所解释的变异精度，分别代表了 t_h 对 y 的解释能力和 $t_1, t_2, \ldots, t_m$ 对 y 的累计解释能力。

对于 VIP_j 很大(>1)的 x_j，它在解释 y 时就有更加重要的作用，因此 VIP 值可用于自变量的选取。

3 结果与分析

3.1 线性相关分析

对监测数据进行变量间的相关分析，见表 1。可知不仅一些自变量与因变量间存在高度线性相关，而且某些自变量与自变量之间还存在多重显著线性相关，甚至高度线性相关。

3.2 数据标准化

运用偏最小二乘回归法建模，其步骤参照文献(张恒喜等, 2002)。为了公式表达上的方便和减少运算误差，首先对样本数据进行标准化处理，得到标准化后的自变量矩阵 E_0 和因变量矩阵 F_0，见表 2。

表 1　变量间相关系数

Tab. 1　Correlation coefficients of the variables

站位 Station	r	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	y
A	X_1	1	0.430	0.956	0.577	0.022	–0.275	–0.466	0.186	–0.423
	X_2		1	0.483	0.321	–0.057	–0.326	–0.610	0.027	–0.563
	X_3			1	0.644	0.132	–0.387	–0.581	0.202	–0.548
	X_4				1	–0.307	0.043	–0.040	–0.422	–0.008
	X_5					1	–0.206	–0.189	0.180	–0.291
	X_6						1	0.728	–0.364	0.929
	X_7							1	–0.491	0.899
	X_8								1	–0.436
	y									1
B	X_1	1	0.456	0.504	0.673	–0.708	–0.623	0.457	0.148	0.714
	X_2			0.037	–0.083	0.075	0.077	–0.067	0.516	–0.114
	X_3				0.108	–0.293	–0.387	0.227	–0.214	0.275
	X_4					–0.873	–0.842	0.186	–0.116	0.919
	X_5						0.697	0.432	–0.117	–0.929
	X_6							0.124	0.443	–0.864
	X_7								0.380	0.292
	X_8									–0.141
	y									1

注：表中 X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8 分别代表 T、S、pH、DO、COD、PO_4^{3-}、无机氮、石油类 8 个自变量，y 代表因变量 Chl-a。相关系数 r：$0.5 \leq |r| < 0.8$ 为显著线性相关；$0.8 \leq |r| \leq 1.0$ 为高度线性相关。表 1–表 5 同

Note: X_1, X_2, X_3, X_4, X_5, X_6, X_7, and X_8 represent independent variables of T, S, pH, DO, COD, PO_4^{3-}, nutrient N, and oil, respectively. y represents the dependent variable of chlorophyll-a. Correlation coefficient r: $0.5 \leq |r| < 0.8$ significant linear correlation; $0.8 \leq |r| \leq 1.0$ highly significant linear correlation

表 2　样本数据标准化

Tab. 2　Normalization of the sample data

站位 Station	样本 Sample No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	y
A	1	–2.065	0.388	–1.814	–0.891	–0.914	–0.430	0.112	–0.218	–0.152
	2	–1.318	–2.241	–1.119	–1.076	1.699	0.384	1.159	–0.021	0.834
	3	–0.362	–0.097	–0.598	0.307	–0.687	2.625	1.286	–0.662	2.090
	4	0.176	–0.391	–0.250	0.399	–1.149	0.181	0.747	–0.662	0.727
	5	0.803	–0.707	0.617	0.922	–1.180	–0.226	0.541	–0.416	0.116
	6	0.923	1.509	1.138	1.137	1.150	–0.430	–0.045	–0.711	–0.681
	7	0.953	0.564	1.486	1.322	–0.119	–0.837	–1.663	0.963	–1.106
	8	0.564	0.281	0.270	–1.629	0.487	–0.633	–0.870	2.441	–0.870
	9	0.325	0.693	0.270	–0.492	0.714	–0.633	–1.267	–0.711	–0.959
B	1	–2.054	–1.525	–0.767	–1.386	1.859	0.633	–1.202	–1.596	–1.236
	2	–1.361	0.0228	–2.002	–0.371	0.089	1.215	–0.509	1.407	–0.629
	3	–0.349	–0.353	1.590	–0.481	0.185	0.168	–0.325	0.155	–0.637
	4	0.199	1.806	–0.261	–0.920	1.440	0.866	–0.579	0.005	–1.404
	5	0.805	1.323	0.074	–0.358	0.025	0.400	0.944	1.707	0.046
	6	0.892	0.570	0.973	0.834	–1.004	–1.810	–1.183	–0.945	1.366
	7	0.921	–0.471	–0.598	2.219	–1.197	–1.694	–0.102	–0.545	1.520
	8	0.574	–0.542	0.467	0.231	–0.731	0.051	1.182	0.305	0.345
	9	0.372	–0.830	0.523	0.231	–0.666	0.168	1.775	–0.495	0.629

注：同表 1

The same as Tab.1

3.3 模型建立合理性分析

在自变量中提取的主成分 t 与因变量中提取的主成分 u 绘制平面图，得 A、B 站位的 $t1$ 与 $u1$ 关系图，见图 2 和图 3。从图 2、图 3 中可知，自变量 x 和因变量 y 有显著的相关关系，说明采用偏最小二乘回归法建立 y 对 x 的线性模型是较合理的。从图 4 和图 5 特异样本判别的 T^2 椭圆图可知，所有样本点都落在椭圆内，认为各监测站位中使用的样本没有特异点。

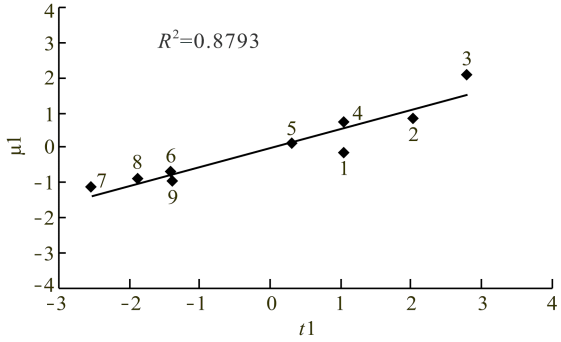

图 2 站位 A 的 $t1/u1$ 关系

Fig. 2 $t1/u1$ diagram of Station A

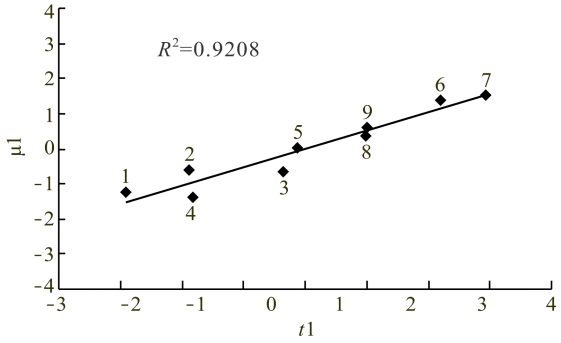

图 3 站位 B 的 $t1/u1$ 关系

Fig. 3 $t1/u1$ diagram of Station B

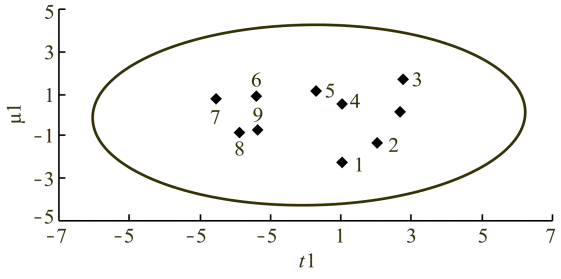

图 4 站位 A 的 T^2 椭圆图

Fig. 4 T^2 ellipse fitting to Station A

3.4 提取主成分数量和变量解释能力分析

用交叉有效性分析提取主成分数量，得表 3。

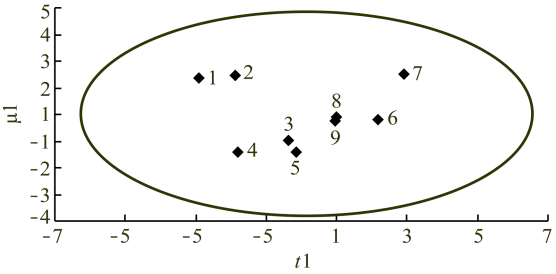

图 5 站位 B 的 T^2 椭圆图

Fig. 5 T^2 ellipse fitting to Station B

表 3 预测误差平方和(*PRESS*)、拟合误差平方和(*SS*)

Tab. 3 Predicted residual sum of squares (*PRESS*) and sum of square error (*SS*) of two stations

提取主成分数量/个 The number of principal components extraction	站位 A Station A		站位 B Station B	
	PRESS	*SS*	*PRESS*	*SS*
1	2.716	1.349	3.162	1.210
2	2.258	0.397	4.234	0.725
3	1.805	0.097	6.016	0.661
4	1.987	0.072	10.304	0.266
5	3.330	0.068	10.793	0.143
6	8.459	0.065	9.230	0.099
7	34.574	0.054	9.352	0.052
8	1.5×10^{10}	7.6×10^{-26}	1.0×10^{4}	7.4×10^{-29}

由于偏最小二乘法将多元校正的目标直接定位在预测上，所以确定主成分数量的原则就是使预测误差最小。但如果提取的主成分之间预测误差平方和(*PRESS*)相差不大，应参考拟合误差平方和(*SS*)的大小选取主成分。从表 3 可得站位 A 确定提取 3 个主成分，站位 B 确定提取 1 个主成分。

用变量投影重要性分析进行自变量对因变量解释能力的分析，得表 4。

从表 4 中变量投影重要性指标 VIP 值排序可知：在站位 A 点，各自变量对因变量(Chl-a)影响能力排序为 PO_4^{3-}＞无机氮＞pH＞S＞T＞石油类＞COD＞DO；在站位 B 点，各自变量对因变量(Chl-a)影响能力排序为 COD＞DO＞PO_4^{3-}＞T＞无机氮＞pH＞石油类＞S。但在站位 B 点上，自变量中 S 和石油类的 VIP 值明显偏小，这与表 2 变量间简单相关关系分析结果一致：S 和石油类与因变量(Chl-a)的相关系数很小，也与其他自变量的相关系数很小，说明 S 和石油类对 Chl-a 的解释作用很差，同时也不能通过其他自变量的传递来解释因变量。因此，认为在站位 B 上 S 和石油类对 Chl-a 的解释不重要而应删去。

表 4　变量投影重要性指标 **VIP** 值
Tab. 4　Variable importance in projection of two stations

自变量 Independent variables	VIP 站位 A Station A	站位 B Station B
X_1	0.822	1.137
X_2	0.910	0.182
X_3	0.962	0.438
X_4	0.421	1.462
X_5	0.496	1.478
X_6	1.564	1.376
X_7	1.449	0.464
X_8	0.775	0.224

经主成分数量提取和变量解释能力的分析后，对站位 A 提取 3 个主成分和对站位 B 自变量修改后提取 1 个主成分分别进行成分解释能力的累计，得表 5。

从表 5 成分累计解释能力可知：在站位 A 点上，$Rd(X) = 0.772$，$Rd(Y) = 0.991$，说明经过 3 个主成分的提取，已概括了自变量集合中 77.2%、因变量中 99.1%的变异信息。在站位 B 点上，$Rd(X) = 0.580$，$Rd(Y) = 0.909$，说明经过一个主成分的提取，已概括了自变量集合中 58.0%、因变量中 90.9%的变异信息。因此，所建立的回归方程精度满足要求。

表 5　成分 ***t*** 的累计解释能力
Tab. 5　$Rd(x)$ and $Rd(y)$ of two stations

累计 Rd	站位 A Station A			站位 B Station B
	t_1	t_2	t_3	t_1
x_1	0.447	0.909	0.916	0.730
x_2	0.404	0.437	0.515	—
x_3	0.585	0.959	0.959	0.184
x_4	0.030	0.684	0.912	0.850
x_5	0.071	0.163	0.282	0.863
x_6	0.626	0.803	0.889	0.730
x_7	0.855	0.879	0.879	0.122
x_8	0.252	0.321	0.823	—
X	0.409	0.644	0.772	0.580
Y	0.879	0.964	0.991	0.909

X 代表自变量集合，Y 代表因变量集合

X represents independent variables, Y represents dependent variables

3.5　建立回归方程

在站位 A 点上，实施 F_0 在 t_1，t_2，t_3 上的回归，得标准化方程 $\hat{F}_0(\hat{y}^*)$。再经标准化的逆过程得原自变量的回归方程：$y = 0.0076x_1 - 0.0806x_2 - 1.2869x_3 + 0.1498x_4 - 0.1905x_5 + 126.6x_6 + 5.985x_7 - 0.3538x_8 + 10.84$。

在站位 B 点上，实施 F_0 在 t_1 上的回归，得标准化方程 $\hat{F}_0(\hat{y}^*)$。再经标准化的逆过程得原自变量的回归方程：$y = 0.0781x_1 + 0.5852x_2 + 0.4773x_3 - 0.5661x_4 - 38.11x_5 + 0.8501x_6 - 4.429$。

4　讨论

根据《GB3097-1997》海水水质标准，站位 A 海水中 COD、PO_4^{3-}和无机氮含量基本优于第一类海水水质标准；站位 B 海水中 COD 和 PO_4^{3-}含量处于第二类至第三类海水水质标准之间，无机氮含量劣于第四类海水水质标准。参考国家海洋局《海水增养殖区监测技术规程》的评价准则，运用营养指数法 E = COD×无机氮×活性磷酸盐×10^6/4500，计算出站位 A 的平均营养指数 E_A=0.041<1，站位 B 的平均营养指数 E_B=1.005>1。由此可将两站位的水质状况作简单划分：川岛外站位 A 属于贫营养化状态，沙堤口站位 B 属于富营养化状态。

经偏最小二乘回归法分析后可得，川岛海域 8 个理化环境因子对 Chl-a 含量影响具有如下特点：1) 在离岸的川岛外海域，对 Chl-a 含量的影响最重要因子为 PO_4^{3-}和无机氮，其 VIP 值分别为 1.564 和 1.449，比次要因子 pH 的 VIP 值高出约 63%；DO 的 VIP 值最小，是在 8 个因子中对 Chl-a 含量影响最低。这是由于浮游植物光合作用的同时也需要吸收无机盐以合成氨基酸、蛋白质、核酸等生命活动所需的物质，在贫营养化水域，营养盐被植物所吸收，容易形成缺营养盐的状态，因此，营养盐的补充情况就成为决定海洋初级生产量的重要条件。同时，离岸海域表层海水容易处于氧饱和状态，DO 对 Chl-a 含量的影响作用小。2) 在近岸的沙堤口海域，对 Chl-a 含量的影响最重要因子为 COD 和 DO，其 VIP 值分别为 1.478 和 1.462。同时，T 的 VIP 值为 1.137，亦成为影响 Chl-a 含量的重要因子。而营养盐中无机氮的 VIP 值较低，成为对 Chl-a 影响的非重要因子。这一特点与酶动力学的米氏方程：$v = (Vm \cdot I)/(K_I + I)$描述相一致。式中，$v$ 为营养盐被吸收的速率；Vm 为最大吸收速率；K_I为吸收半饱和常数；I 为介质中的营养盐浓度。在营养盐低浓度条件下，植物对营养盐的吸收率随着浓度提高而迅速增大，达到一个平衡状态后，吸收率不再随营养盐浓度提高而加快。因此，在已富营养化的沙堤口海域，营养盐成为非重要因子。COD 和 DO 成为影响 Chl-a 含量的重要因子，这是由于浮游植物虽能通过光合作用产生 DO，但浮游植物不论在什么时候都在

进行呼吸作用，消耗大量的 DO 以维持自身新陈代谢活动的需要，正常情况下，海水中 DO 含量增加对浮游植物的新陈代谢活动有促进作用，有利于浮游植物的增殖，但海水中有机残体的分解也需要消耗大量的 DO，因此，DO 和 COD 通过影响浮游植物的新陈代谢活动对海洋初级生产力产生影响。T 对 Chl-a 的影响主要表现在：当光照强度达到光饱和值后，植物的光合作用速率随 T 的升高而增加；另一方面，T 通过影响营养盐吸收半饱和常数 K_1，影响营养盐被吸收的速率，从而影响 Chl-a 含量。

5 结论

用偏最小二乘回归法，在川岛海域不同水域中，选取 8 个理化环境因子对海洋初级生产力的影响情况进行分析。结果表明，1) 在贫营养化的川岛外水域，各理化因子对 Chl-a 含量的回归数学模型：$y = 0.0076x_1 - 0.0806x_2 - 1.2869x_3 + 0.1498x_4 - 0.1905x_5 + 126.6x_6 + 5.985x_7 - 0.3538x_8 + 10.84$。对 Chl-a 含量影响能力最大为 PO_4^{3-}和无机氮，其次为 pH、S 和 T，DO 影响能力最小。2) 在已富营养化的沙堤口水域，各理化因子对 Chl-a 含量的回归数学模型：$y = 0.0781x_1 + 0.5852x_2 + 0.4773x_3 - 0.5661x_4 - 38.11x_5 + 0.8501x_6 - 4.429$。对 Chl-a 含量影响能力最大为 COD 和 DO，T 也成为影响海洋初级生产力的重要因子，而营养盐中的无机氮却变成为非重要因子。

从成分累计解释能力分析可知，在贫营养化的川岛外水域中各理化环境因子集合的累计解释能力大于在富营养化的沙堤口水域。因此，偏最小二乘回归法在贫营养化的川岛外水域海洋初级生产力影响因子分析应用的精度更高。

参 考 文 献

王黎明, 陈颖, 杨楠. 应用回归分析. 上海: 复旦大学出版社, 2008, 44–100

孙耀, 宋云利, 崔毅, 等. 桑沟湾养殖水域的初级生产力及其影响因素的研究. 海洋水产研究, 1996, 17(2): 32–40

孙雪梅, 夏斌, 过锋. 2013. 青岛崂山近岸海域浮游植物群落结构及其与环境因子的关系. 渔业科学进展, 24(2): 46–53

李小斌, 陈楚群, 施平, 等. 珠江口海域总无机氮的遥感提取研究. 环境科学学报, 2007, 27(2): 313–318

沈国英, 施并章. 海洋生态学(第二版). 北京: 科学出版社, 2002

张正健, 刘志红, 郭艳芬, 等. 偏最小二乘在遥感监测西藏草地生物量上的应用. 草地学报, 2009, 17(6): 735–739

张丽旭, 张小伟. 用于海洋环境科学的一种新方法—影响因子分析法. 海洋科学进展, 2004, 22(1): 55–61

张恒喜, 郭基联, 朱家元, 等. 小样本多元数据分析方法及应用. 西安: 西北工业大学出版社, 2002, 24–33

陈馨, 楚宪法. 将偏最小二乘回归模型应用于非点源负荷预测. 水利科技与经济, 2013, 19(11): 7–9

柯志新, 黄良民, 谭烨辉, 等. 2008 年夏末南海北部叶绿素 a 的空间分布特征及其影响因素. 热带海洋学报, 2013, 32(4): 51–57

高会旺, 杨华, 张英娟, 等. 渤海初级生产力的若干理化影响因子初步分析. 青岛海洋大学学报, 2001, 31(4): 487–494

Johnson RA, Wichern DW. 实用多元统计分析(第六版英文). 北京: 清华大学出版社, 2008, 430

(编辑 江润林)

The Application of Partial Least Squares (PLS) Regression to the Factors Affecting the Ocean Primary Productivity

GUAN Jionghui, YANG Zhenjie
(*Jiangmen Marine & Fishery Environment Monitoring Station, Jiangmen* 529000)

Abstract The concentration of chlorophyll-a is an important index of phytoplankton biomass that can reflect the status of ocean primary productivity. According to the analysis on the physicochemical environmental factors that affect the concentration of chlorophyll-a, we may clarify the dominant factors of ocean primary productivity. Two monitoring stations were situated on the sea area in Jiangmen, Guangdong Province, China. The survey was conducted from March to November 2010 to determine chlorophyll-a, water temperature, salinity, pH, dissolved oxygen (DO), chemical oxygen demand (COD), phosphate, nutrient N and oil in two stations. According to the principal component analysis (PCA), the application of the partial least squares (PLS) regression solved the multiple regression analysis including the multiple correlations between the environmental factors with a high precision; it also established the mathematical model with fewer monitoring data. In the PLS regression, the monitoring data were normalized, and then used the variable importance in projection and T^2 ellipse fitting-aided analysis technique for extraction. The regression equation of station A was $y = 0.0076x_1 - 0.0806x_2 - 1.2869x_3 + 0.1498x_4 - 0.1905x_5 + 126.6x_6 + 5.985x_7 - 0.3538x_8 + 10.84$. The regression equation of station B was $y = 0.0781x_1 + 0.5852x_2 + 0.4773x_3 - 0.5661x_4 - 38.11x_5 + 0.8501x_6 - 4.429$. The eutrophic index of station A was 0.041, suggesting that the water of Chuandao sea area was oligotrophic. The eutrophic index of station B was 1.005, suggesting that the water of Shadi sea area was eutrophic. The results showed that the main effective factors of chlorophyll-a were nutrients N and phosphate in oligotrophic waters of Chuandao sea area, and that COD, DO and water temperature were the effective factors of the ocean primary productivity in Shadi sea area.

Key words Partial least squares (PLS) regression; Ocean primary productivity; Physicochemical environment factors; Eutrophic; Oligotrophic

First author: GUAN Jionghui, E-mail: guanjh1018@163.com