

DOI :10.3876/j.issn.1000-1980.2009.01.006

基于偏最小二乘回归法的河流病因诊断

蒲前超¹ 夏继红² 韩玉玲³ 胡 玲³ 罗 帷⁴

(1.河海大学水利水电工程学院,江苏 南京 210098 ;2.河海大学农业工程学院,江苏 南京 210098 ;
3.浙江省河道管理总站,浙江 杭州 310009 ;4.海宁市水利局,浙江 海宁 314400)

摘要 :在分析河流健康外部影响因子的基础上,指出各因素之间存在严重的多重相关性,并应用一种能消除这种不良影响的多元统计方法——偏最小二乘回归法,建立了河流病因诊断模型.结合海宁市辛江塘的具体条件进行了河流病因诊断实例分析.结果表明,影响辛江塘河流健康的最主要的原因工业污水排放量,其次为 GDP,再依次为生活污水排放量、人口、农业污水排放量、环境保护投资指数、降雨量.提出了严格控制污染源,进一步完善污水管网建设,加大河道综合整治力度,完善河道管理体制的治理对策.

关键词 :河流健康;病因诊断;偏最小二乘回归;辛江塘

中图分类号 :X171 **文献标识码 :**A **文章编号 :**1000-1980(2009)01-0027-05

河流作为生态系统的重要组成部分,是具有生命的有机体,其健康状况的好坏,直接影响其功能的发挥^[1].“河流健康”概念的提出,充分体现了“人与河流和谐发展”的河流建设与管理的新理念^[2-4].河流健康业已受到社会各界的广泛关注,许多学者也围绕“复兴河流生命、保护河流生态、维护河流健康”主题进行了大量的研究和探讨^[5-9].但目前对河流健康的研究大多集中于对河流健康现状的综合评价,而对引起河流健康问题的原因却研究不多.科学有效的河流治理或修复措施是建立在充分掌握致病原因基础上的,因此河流健康研究不仅需要评价河流健康现状,还应加强河流病因诊断,掌握河流的致病原因.本文在分析影响河流健康外部因素的基础上,研究基于偏最小二乘法的河流病因诊断方法和诊断过程,并结合海宁辛江塘的具体条件,诊断该河道存在的问题及其成因.

1 河流健康的外部影响因子

河流系统产生健康问题的主要原因在于系统受到外界的干扰或胁迫.当这些扰动在河流系统承载力范围内时,河流系统可以保持基本健康状态,但当这些扰动超出河流系统的承载力时,河流系统就会病变,处于亚健康状态或病态,从而严重影响河流各项功能的发挥.影响河流健康的外界因子包括自然因子和社会因子 2 大类.

a. 自然因子又可分为突发性因素和常规性因素.前者主要包括地震、洪涝灾害、异常旱情和河流袭夺等,这些剧烈的环境变化会极大地破坏河流系统,造成河流系统结构和功能不可逆转的改变;后者主要包括降雨量、蒸发量、气温、湿度、风向等水文、气象因子和下垫面的地形地貌条件,这些因素都长期持续作用于河流系统,维持或改变河流系统的结构和功能.

b. 社会因子包括社会环境和人类活动.社会环境指河流系统相应区域内的人口状况和经济发展水平;人类活动是损害河流健康的主要因素,包括对河流系统资源的过度开发利用、河道周边区域的各种工程建设和涉水工程中不合理的物理重建、工农业和生活污水的不达标排放等.

以上各外部因子均会从不同方面对河流健康产生一定程度的影响,但对各具体河流病因诊断而言,并非每个因子都需要考虑,而是要根据具体情况,因时而异、因地制宜地选取较为重要的影响因子作为病因诊断指标.

收稿日期 :2008-05-12

基金项目 :浙江省水利科技重点项目(RB0811)

作者简介 :蒲前超(1984 —)男,湖北天门人,硕士研究生,主要从事河流健康研究.

2 河流病因诊断模型

2.1 病因诊断方法的选择

河流病因诊断^[10]是通过分析各外部影响因子与病症状态之间的定量关系来揭示产生河流健康的主要外部原因的.通常,河流健康问题往往表现为多种症状同时出现,各病症因子受多种外部因子影响,而每个影响因子同时作用于一个或多个病症因子.可见,病症因子和外部影响因子之间是一种多对多的映射关系.同时,影响河流健康的众多外部因素也掺杂了大量的噪声,而且各因素相互间的多重相关性关系是不可避免的.因此,河流病因诊断是对具有多重相关性的多维变量统计分析的过程.

偏最小二乘回归法^[11-12]是一种多因变量对多自变量的回归建模方法,易于辨识系统中的信息与噪声.特别是当各变量间存在较高度度的相关性或样本容量小于解释变量个数时,用偏最小二乘回归法建模分析,可以有效地提取对因变量解释最强的综合变量,克服变量间的多重相关性,而且每一个自变量的回归系数更容易解释.同时,在该方法的辅助分析技术中,利用变量投影重要性分析可以确定每一个自变量在解释因变量时的重要程度.所以,偏最小二乘回归法可以较好地建立病症因子和外部因子间的定量关系,有效确定影响河流健康的最主要原因.所以,本文选择该方法来研究河流病因诊断模型和诊断过程.

2.2 病因诊断思路和模型

2.2.1 病因诊断思路

在病因诊断过程中,选取河流某一病症因子作为因变量 y ,以其 p 个影响因子 $x_1, x_2, \dots, x_p$ 组成自变量集合 X ,搜集各变量的现时和历史资料作为分析样本,以偏最小二乘回归法为基础,建立因变量 y 与自变量集合 X 的回归方程.根据计算所得的回归系数,分析回归方程的拟合精度,解释各影响因子对河流健康状态的影响性质,并利用变量投影重要性指标计算病因影响程度,从而确定影响河流健康的各因子的主次顺序,最后在此基础上有针对性地提出治理对策.

2.2.2 病因诊断模型

2.2.2.1 回归方程的建立

根据病因诊断思路和单因变量偏最小二乘回归算法,河流病因诊断的建模步骤如下:

a. 将自变量集合数据矩阵 $X = (x_{ij})_{n \times p}$ 与因变量数据表 $y = (y_i)_{n \times 1}$ 进行标准化处理,得到标准化变量矩阵 E_0 和 F_0 .

$$E_0 = (x_{ij}^*)_{n \times p} \quad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \quad F_0 = (y_i^*)_{n \times 1} \quad y_i^* = \frac{y_i - \bar{y}}{s_y}$$

b. 从 E_0 中提取第一个成分 $t_1 = E_0 w_1$,其中 $w_1 = \frac{E_0^T \cdot F_0}{\|E_0^T \cdot F_0\|}$,且 $\|w_1\| = 1$.实施 E_0 和 F_0 在 t_1 上的回归:

$$E_0 = t_1 p_1^T + E_1, F_0 = t_1 r_1 + F_1. \text{ 其中 } p_1 \text{ 和 } r_1 \text{ 为回归系数, } p_1 = \frac{E_0^T t_1}{\|t_1\|^2}, r_1 = \frac{F_0^T t_1}{\|t_1\|^2}. \text{ 记残差矩阵为}$$

$$E_1 = E_0 - t_1 p_1^T \quad F_1 = F_0 - t_1 r_1$$

利用交叉有效性检验其收敛性.若 y 对 t_1 的回归方程达到满意的精度要求,则进行下一步,否则令 $E_0 = E_1$, $F_0 = F_1$,回到步骤b,对残差矩阵进行新一轮的成分提取和回归分析.

c. 假设在第 h 步($h = 1, 2, \dots, m$)回归方程满足精度要求,这时得到 m 个成分 $t_1, t_2, \dots, t_m$,实施 F_0 对 $t_1, t_2, \dots, t_m$ 的回归,得 $\hat{F}_0 = r_1 t_1 + r_2 t_2 + \dots + r_m t_m$.由于 $t_1, t_2, \dots, t_m$ 均是 E_0 的线性组合, $\hat{F}_0$ 可写成 E_0 的线性组合形式,即

$$\hat{F}_0 = r_1 E_0 w_1^* + r_2 E_0 w_2^* + \dots + r_m E_0 w_m^*$$

式中 $w_h^* = [\prod_{j=1}^{h-1} (I - w_j p_j^T)] w_h$, I 为单位矩阵.最后可得

$$\hat{y}^* = \alpha_1 x_1^* + \alpha_2 x_2^* + \dots + \alpha_p x_p^*$$

式中 x_j^* ——所选取的主要外部影响因子; α_j —— x_j^* 的回归系数,其计算式为

$$\alpha_j = \sum_{h=1}^m r_h w_{hj}^*$$

w_{hj}^* 是 w_h^* 的第 j 个分量.

d. 按照标准化的逆过程 将 $\hat{y}^*$ 的回归方程还原成 y 对 X 的回归方程.

2.2.2.2 交叉有效性判别

在偏最小二乘回归建模中,究竟应该选取多少个成分为宜,一般采用交叉有效性进行判别.记 y_i 为原始数据 $t_1, t_2, \dots, t_m$ 是在偏最小二乘回归过程中提取的成分. $\hat{y}_{hi}$ 是使用全部样本点并提取 h 个成分回归建模后,第 i 个样本点的拟合值. $\hat{y}_{h(-i)}$ 是在建模时删去样本点 i 取 h 个成分回归建模后,再用此模型计算 y_i 的拟合值.记

$$S_h = \sum_{i=1}^n (y_i - \hat{y}_{hi})^2 \qquad S'_h = \sum_{i=1}^n (y_i - \hat{y}_{h(-i)})^2 \qquad Q_h^2 = 1 - \frac{S'_h}{S_{h-1}}$$

当 $Q_h^2 \geq 0.0975$ (即 $\sqrt{S'_h} \leq 0.95 \sqrt{S_{h-1}}$) 时,表明引进新的主成分 t_h 会改善模型质量;反之,则不能.

3 海宁市辛江塘病因诊断

3.1 辛江塘概况

辛江塘是海宁市“六纵九横”河道网络框架中西引东泄的主干河道,亦是横贯东西的一条水上运输通道.其河道西起盐官下河,东至袁硖港,中连斜郭塘、绵长港、平阳堰港、唐家堰港、麻泾港、木场桥等诸多海宁市级河流和镇主要河流,全长 22.528 km.辛江塘存在着许多浙江省平原河流共有的健康问题^[13],而且由于海宁市城镇规模的扩大,工农业的不断发展和城市污水处理能力的不足,造成辛江塘水质污染日益加剧,使得河道水环境问题较为突出.

3.2 变量分析与数据选取

影响辛江塘健康的外部因素众多,根据辛江塘健康状态的实际情况和现有数据资料的完备性,本算例选择表征河道水质状态的水质综合污染指数作为因变量 y .而从影响水质状态的众多外部因素中,选取年降雨量(mm)、人口数量(万人)、GDP(亿元)、工业污水排放量(万 t)、农业污水排放量(万 t)、生活污水排放量(万 t)、环境保护投资指数(%)等 7 个指标作为自变量.其中,年降雨量 x_1 表示自然因素对水质的影响作用;人口数量 x_2 和 GDP x_3 反映城市规模和经济发展水平对水环境的影响;工业污水排放量 x_4 、农业污水排放量 x_5 和生活污水排放量 x_6 表示人类生产生活对水环境的破坏;环境保护投资指数 x_7 表示河道建设与管理活动对水环境的影响.以辛江塘 1996~2007 年水环境污染状况及其主要影响因子的历史数据作为分析样本.数据主要来自海宁市统计年鉴、地表水资源质量评价报告和嘉兴市水资源公报等,标准化后的数据见表 1.

表 1 标准化数据

Table 1 Standardized data

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	-0.179 6	2.029 3	1.963 3	2.006 4	2.248 7	2.028 9	0.397 1	1.366 3
2	-0.574 3	1.437 6	1.441 4	1.485 1	1.165 7	1.438 1	-1.101 9	2.175 1
3	-0.969 8	0.752 5	0.921 2	0.895 3	0.906 3	0.752 3	1.518 3	0.667 8
4	-1.255 3	0.472 3	0.515 4	0.471 8	0.553 8	0.472 6	0.555 5	0.704 6
5	-1.419 8	-0.057 1	0.095 6	0.045 5	-0.263 6	-0.057 2	0.653 0	-0.251 2
6	1.176 4	-0.368 4	-0.245 0	-0.304 4	-0.640 5	-0.368 3	0.470 2	-0.398 2
7	0.039 1	-0.306 2	-0.449 1	-0.508 3	-0.527 9	-0.306 4	0.555 5	-0.435 0
8	-0.473 8	-0.368 4	-0.617 9	-0.660 7	-0.581 8	-0.368 3	0.470 2	-0.582 1
9	1.087 2	-0.773 3	-0.817 0	-0.776 5	-1.100 6	-0.773 5	0.263 0	-0.986 4
10	1.072 5	-0.617 6	-0.878 7	-0.837 5	-0.846 1	-0.617 5	-0.443 8	-0.876 1
11	1.470 7	-0.742 1	-0.943 7	-0.889 7	-0.371 3	-0.742 1	-1.699 1	-0.618 8
12	0.026 8	-1.458 4	-0.985 4	-0.926 9	-0.542 6	-1.458 4	-1.638 2	-0.765 9

3.3 回归方程的建立

按照病因诊断的建模思路和步骤,利用面向对象的可视化程序设计软件 VB6.0 进行编程计算,各变量之间的相关系数如表 2 所示,可见部分自变量因子之间存在较强的相关性.根据交叉有效性的计算结果(表 3)提取了前 2 个主成分进行模型的参数估计和预测.将标准化变量分别还原成原始变量后,得到最终的 PLS 回归方程为

$$y = -20.2 - 7.073 \times 10^{-5}x_1 + 0.16x_2 + 7.468 \times 10^{-4}x_3 + 2.459x_4 + 2.797x_5 + 4.478 \times 10^{-3}x_6 - 7.171 \times 10^{-2}x_7$$

表 2 相关系数矩阵

Table 2 Matrix of correlation coefficients

指标	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
x_1	1.000	-0.445	-0.510	-0.490	-0.453	-0.445	-0.410	-0.485
x_2	-0.445	1.000	0.896	0.895	0.854	0.917	0.316	0.844
x_3	-0.510	0.896	1.000	0.916	0.873	0.896	0.306	0.862
x_4	-0.490	0.895	0.916	1.000	0.879	0.895	0.273	0.865
x_5	-0.453	0.854	0.873	0.879	1.000	0.854	0.155	0.830
x_6	-0.445	0.917	0.896	0.895	0.854	1.000	0.316	0.844
x_7	-0.410	0.316	0.306	0.273	0.155	0.316	1.000	0.093
y	-0.485	0.844	0.862	0.865	0.830	0.844	0.093	1.000

根据公式

$$R_d(X; i_1, i_2, \dots, i_m) = \frac{1}{p} \sum_{h=1}^m \sum_{j=1}^p R_d(x_j; i_h) = \sum_{h=1}^m r^2(x_j; i_h)$$

$$R_d(y; i_1, i_2, \dots, i_m) = \sum_{h=1}^m R_d(y; i_h) = \sum_{h=1}^m r^2(y; i_h)$$

由回归系数计算可得： $R_d(X) = 0.904$ ， $R_d(y) = 0.928$ ，前者表示所提取的主成分对 X 的累计解释能力，后者表示所提取的主成分对 y 的累计解释能力。由此可见，所建立的模型比较满意，所提取的 2 个主成分已能概括自变量集合中 90.4% 和因变量 92.8% 的变异信息。

3.4 病因重要程度计算

在偏最小二乘法的辅助分析技术中，自变量 x_j 在解释因变量时的重要性程度的大小可以通过变量投影重要性指标 V_{IP_j} 来测度（其定义式见式（1）），所得结果较回归方程式的结果更加直观。

$$V_{IP_j} = \sqrt{\frac{p}{R_d(y; i_1, i_2, \dots, i_m)} \sum_{h=1}^m R_d(y; i_h) w_{hj}^2} \tag{1}$$

对于 V_{IP_j} 越大的 x_j ，它在解释 y 时就具有更加重要的作用。因此，可以根据 V_{IP_j} 的大小来判断外部因子对河流健康的影响程度。辛江塘各病因的 V_{IP_j} 值计算结果为： $V_{IP_1} = 0.6409$ ， $V_{IP_2} = 1.0806$ ， $V_{IP_3} = 1.1765$ ， $V_{IP_4} = 1.2322$ ， $V_{IP_5} = 1.0167$ ， $V_{IP_6} = 1.1087$ ， $V_{IP_7} = 0.7475$ 。由此可见，在影响辛江塘水质的 7 个因子中，其影响程度的大小顺序为：工业污水排放量、GDP、生活污水排放量、人口、农业污水排放量、环境保护投资指数、降雨量。

3.5 病因分析和治理对策

由回归方程可知，所有自变量的系数符号均与实际情况十分吻合。年降雨量和环境保护投资指数与水质污染指数负相关，表明降雨对河道污染水体有一定的稀释作用，而环保资金的投入也对改善水质有积极作用；其余变量与水质污染指数正相关，表明城市规模、经济发展水平和污水排放量都会在一定程度上使水质恶化。由 V_{IP_j} 值计算结果可知：工业污水排放量是造成水质下降的最主要原因，这是因为海宁市有大批的能对水环境造成严重污染的皮革企业；其次 GDP 的影响也很明显，说明经济的发展有很大一部分是靠牺牲环境获得的；而随着人口的增长，生活污水排放量不断加大，也给水环境造成了很大的压力；在农业生产中，农药化肥使用量持续增高，造成的面源污染使得水体质量进一步恶化，而环保投资的增加、水环境综合整治措施的实施，虽然能在一定程度上保护水环境，但投入的力度太小，效果也只能是杯水车薪；对水环境影响最小的是降雨量，这说明靠自然调节来缓解人为对环境造成的污染和破坏是不太现实的。

以上分析结果表明，要有效遏制辛江塘水环境的进一步恶化，改善河流健康现状，还必须在以下几个方面采取措施：

- a. 控制工业污染源。对现有工业点源，要严格控制污水的排放和处理，实行达标排放或限停污染严重的企业；调整产业结构和工业布局，严格控制新建污染企业，从源头上遏制水环境的进一步恶化。
- b. 减轻农业污染源。严格控制化肥和农药的滥用，减少肥药对水环境的污染，积极推广生态农业模式。
- c. 完善城镇生活污水管网系统建设，逐步实行雨污分流，提高污水进网率。

表 3 各成分的交叉有效性结果

Table 3 Cross-validation results of components

h	Q^2_h	临界值
1	0.778 8	0.097 5
2	0.173 2	0.097 5
3	-0.161 9	0.097 5

d. 加大河道综合整治的投入和力度,在疏浚河道的基础上,采用生态环保的治理措施,改善水生态环境,提高水环境容量,尽快实现“水清、流畅、岸绿、景美”的整治目标。

e. 建立和完善河流管理体制,基于人与自然和谐共处的理念,加强河流管理,建立健全法律法规和政策体系,完善执法监督体系,建设全民共同参与机制。

总之,只有在加强水污染防治投入的同时,不断提高人民的环保意识,转变河流建设与管理的理念,才能改善或维持河流的健康状态。

4 结 论

本文在总结和借鉴现有河流健康相关研究成果的基础上,分析了河流健康的影响因素,提出一种能克服变量多重相关性在系统建模过程中不良作用的方法——偏最小二乘回归法。本文还研究了基于单因变量偏最小二乘回归方法的河流病因诊断过程,并结合海宁市辛江塘的具体条件,建立了辛江塘病因诊断模型,其结果与实际情况较为符合,通过变量重要性投影指标,分析得出各因素对辛江塘水环境问题的影响程度,确定了辛江塘健康的主要病因,提出了相应的治理措施,为辛江塘下一步的建设与管理提供了可靠的依据。由此可见,基于偏最小二乘回归法的河流病因诊断是河流健康研究的一种较为有效的方法。

参考文献：

[1] 夏继红,卢智灵.河流生命健康仿生学研究[J].水科学与工程技术,2006(5):44-46.
[2] 李国英.黄河治理的终极目标是“维持黄河健康生命”[J].人民黄河,2004,26(1):1-3.
[3] 蔡其华.维护健康长江 促进人水和谐 摘自蔡其华同志 2005 年长江水利委员会工作报告[J].人民长江,2005,36(3):1-3.
[4] 魏梓兴,蔡恭杰,吴阿娜,等.河流管理中引入河流健康状况评价的意义与思考[J].中国水利,2004(17):18-19.
[5] 杨文慧,严忠民,吴建华.河流健康评价的研究进展[J].河海大学学报:自然科学版,2005,33(6):607-611.
[6] 吴阿娜,杨凯,车越,等.河流健康状况的表征及其评价[J].水科学进展,2005,16(4):602-607.
[7] 杨文慧,杨宇.河流健康概念及诊断指标体系的构建[J].水资源保护,2006,22(6):28-30.
[8] 蔡守华,胡欣.河流健康的概念及指标体系和评价方法[J].水利水电科技进展,2008,28(1):23-27.
[9] 酆建强,杨晓华,陆桂华,等.河流健康复杂系统评价的 IFMMAAM[J].河海大学学报:自然科学版,2008,36(2):152-156.
[10] 杨文慧.河流健康的理论构架与诊断体系的研究[D].南京:河海大学,2007.
[11] WOLD S, RUHE A, WOLD H, et al. The collinearity problem in linear regression: the partial least square (PLS) approach to generalized inverses[J]. J Stat Comp, 1984, 5: 735-743.
[12] 王惠文.偏最小二乘回归方法及其应用[M].北京:国防工业出版社,1999.
[13] 胡玲,韩玉玲,夏继红.浙江省河流健康诊断思路及关键内容[J].河海大学学报:自然科学版,2008,36(3):322-325.

Reason diagnosis of river health based on partial least squares regression

PU Qian-chao¹, XIA Ji-hong², HAN Yu-ling³, HU Ling³, LUO Wei⁴

- (1. College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China ;
- 2. College of Agricultural Engineering, Hohai University, Nanjing 210098, China ;
- 3. River Management Station of Zhejiang Province, Hangzhou 310009, China ;
- 4. Haining Water Conservancy Bureau, Haining 314400, China)

Abstract : Based on the analysis of the external factors for river health, the severe multiple correlation among various factors was found. A multiple statistical method, partial least squares (PLS) regression method, was put forward to eliminate the adverse effect. A reason diagnosis model of river health was established. According to the existing conditions of Xinjiangtang River in Haining City, the reason diagnosis case was analyzed. The results indicate that the key factor to have effect on the health of Xinjiangtang is the discharge of industrial sewage and then GDP, and the order of other factors is the discharge of domestic sewage, the population, the discharge of agricultural sewage, the investment index of environmental protection and the rainfall. Therefore, the corresponding measures were proposed: to control pollution sources, to promote sewage pipe networks, to strengthen river regulations and to improve river management systems, etc.

Key words : river health ; reason diagnosis ; partial least squares regression ; Xinjiangtang River