



偏线性最小二乘法预测 SBR 工艺出水浓度的方法研究

张成甫,付永胜(西南交通大学环境科学与工程学院,成都 610031)

摘要:SBR 工艺出水浓度受进水浓度(COD、 $\text{NH}_3 - \text{N}$)及曝气时间等因素的影响,这些影响之间存在着一定的多层相关性。尝试从统计学角度用偏线性最小二乘法对 SBR 系统进行描述性分析,建立数学模型,对 SBR 系统进行解释性描述。在对 SBR 工艺工况进行充分的观察和认识,对其要素内在联系进行辨识和深入分析的基础上,对 SBR 出水浓度作出预测。

关键词:偏最小二乘法;SBR;污水

中图分类号:X703 **文献标识码:**A **文章编号:**1008 - 2301(2006)02 - 0006 - 05

The Study on Effluent Concentration Forecast from SBR by Partial Least - squares Regression. ZHANG Cheng - fu, FU Yong - sheng (School of Environmental Science and Engineering, Southwest Traffic University, Chengdu 610031, China). Environmental Protection of Xinjiang 2006, 28(2): 06 ~ 10

Abstract: Many factors, such as the inflow COD, $\text{NH}_3 - \text{N}$ concentration, oxygen charge time, work on the effluent concentration from SBR. There are complex correlations among those factors. This paper tries to describe and analyze SBR by Partial Least - squares Regression (PLS) in Multivariate Statistical Analysis, and establishing math model to explain SBR. Through carefully observation and cognition on SBR, and analyzing its inner correlations, this paper forecasts its effluent concentration.

Key words: Partial Least - squares Regression; SBR; wastewater

SBR 是序列间歇式活性污泥法 (Sequencing Batch Reactor Activated Sludge Process) 的简称,是一种按间歇曝气方式来运行的活性污泥污水处理技术,又称序批式活性污泥法。与传统污水处理工艺不同,SBR 技术采用时间分割的操作方式替代空间分割的操作方式,非稳定生化反应替代稳态生化反应,静置理想沉淀替代传统的动态沉淀。它的主要特征是在运行上的有序和间歇操作,SBR 技术的核心是 SBR 反应池,该池集均化、初沉、生物降解、二沉等功能于一池,无污泥回流系统。工艺流程为:进水 - 反应 - 沉降 - 排水 - 闲置。反应阶段主要的控制因数是曝气时间。污水中耗氧物质有有机物、 $\text{NH}_3 - \text{N}$ 等,因此进水 COD、 $\text{NH}_3 - \text{N}$ 浓度及出水要求是曝气时间的重要影响因素^[1]。一般沉降期、排水期的操作变化不大,因此,本文主要分析进水 COD 及 $\text{NH}_3 - \text{N}$ 浓度、曝气时间对出水的影响。

偏最小二乘回归 (Partial Least - squares Regres-

sion,简称 PLS) 是一种新型的多元统计数据分析方法。偏最小二乘回归又被称为第二代回归方法,这是由于它可以实现多种数据分析方法的综合应用。在一个算法下,将多元线性回归分析、典型相关性分析、主成分分析有机结合起来,可以使数据系统的分析内容更加丰富,同时又可以对所建立的回归模型给予许多更详细深入的实际解释。

本文尝试用偏线性最小二乘法建立进水 COD 及 $\text{NH}_3 - \text{N}$ 浓度、曝气时间三因素对出水浓度的数学模型,探讨之间的关联影响。

1 偏最小二乘回归方法

偏最小二乘回归方法在统计应用主要用来建立多因变量与多自变量的统计关系,可以较好地解决许多以往用普通多元回归无法解决的问题。在普通多

元线性回归的应用中,最典型的问题就是自变量之间的多重相关性。如果采用普通的最小二乘法,这种变量多重相关性就会严重危害参数估计,扩大模型误差,并破坏模型的稳定性。变量多重相关问题十分复杂,长期以来在理论和方法上都未给出满意的答案,这一困扰着从事实际系统分析的工作人员。在偏最小二乘回归中开辟了一种有效的技术途径,它利用对系统中的数据信息进行分解和筛选的方式,提取对因变量的解释性最强的综合变量,辨识系统中的信息与噪声,从而更好地克服变量多重相关性在系统建模

中的不良作用。

PLS 的思路^[2]首先是,从自变量集合 $X = [x_1, x_2, \dots, x_p]_{n \times p}$ (p 为自变量个数)中提取主成分 t_h ($h = 1, 2, \dots$),各成分相互独立;与此同时,在因变量 $Y = [y_1, y_2, \dots, y_q]_{n \times q}$ (q 为因变量个数)中提取主成分 u_h ($h = 1, 2, \dots$)。要求 t_h, u_h 尽可能大的提取它们各自变量系统中的变异信息,且二者相关程度能达到最大。这样自变量主成分对因变量主成分的解释性能达到最大。然后,实施 X, Y 对 t_h 的回归并分析结果。

表 1 样本数据及预测数据

Tab.1 Sample data and predica data

序号	C ₁ (mg/L)	C ₂ (mg/L)	t (h)	COD 去除率 Y ₁			NH ₃ - N 去除率 Y ₂		
				观察值	预测值	相对误差 (%)	观察值	预测值	相对误差 (%)
1	547.2	50.4	2	0.740 0	0.757 5	2.36	0.349 0	0.357 3	2.37
2	547.2	50.4	4	0.855 0	0.812 2	- 5.01	0.434 0	0.585 8	34.99
3	547.2	50.4	6	0.861 0	0.864 0	0.35	0.804 0	0.771 8	- 4.01
4	531	55.2	2	0.687 0	0.746 6	8.68	0.454 0	0.364 6	- 19.68
5	531	55.2	4	0.861 0	0.801 9	- 6.86	0.628 0	0.591 7	- 5.78
6	531	55.2	6	0.866 0	0.854 3	- 1.35	0.744 0	0.776 7	4.39
7	459	54.1	2	0.696 0	0.736 6	5.84	0.482 0	0.392 7	- 18.52
8	459	54.1	6	0.852 0	0.845 3	- 0.79	0.768 0	0.795 3	3.56
9	459	54.1	8	0.863 0	0.895 5	3.77	0.972 0	0.948 7	- 2.40
10	386.19	69	2	0.721 6	0.697 9	- 3.29	0.281 2	0.423 1	50.47
11	386.19	69	4	0.772 5	0.755 9	- 2.15	0.704 1	0.638 6	- 9.30
12	386.19	69	8	0.848 3	0.862 6	1.69	0.964 9	0.966 2	0.14

2 偏线性最小二乘法用于 SBR 系统建模

表 2 相关系数矩阵

Tab.2 Matrix of related coefficient

	C ₁	C ₂	t	Y ₁	Y ₂
C ₁	1	- 0.459	0	0.067 8	- 0.077 2
C ₂		1	0	- 0.064 8	0.129 7
t			1	0.779 8	0.963 8
Y ₁				1	0.659 3
Y ₂					1

分析所用样本数据^[3]及预测结果见表 1。表中数据得到条件为 SBR 系统中 DO 维持在 2mg/L, MLSS 控制在 3 500mg/L 左右,沉淀 1h,出水 0.5h,进水为限制性曝气。C₁、C₂ 分别为进水 COD、NH₃ - N 浓度;t

为曝气时间;因变量 Y₁、Y₂ 分别为 COD、NH₃ - N 去除率。表 2 为相关系数矩阵。可见进水 COD 与 NH₃ - N 浓度负相关,污染物质去除率 Y₁、Y₂ 之间正相关,曝气时间与 COD、NH₃ - N 去除率强正相关。可见变量之间存在着多重相关性,用普通多元线性回归进行分析将存在较大误差^[4]。

2.1 系统建模

假设因变量与自变量之间为线性关系^[5,6],作 $t_1 - u_1$ 平面图(第一主成分平面图,见图 1)。由图 1 可知自变量与因变量之间线性关系比较弱。曲线类似于 $y = \ln x$ 曲线图,变量之间存在半对数关系。对因变量 Y 作 $Y' = \exp(Y)$ 处理后进行回归,其 $t_1 - u_1$ 平面图见图 2。

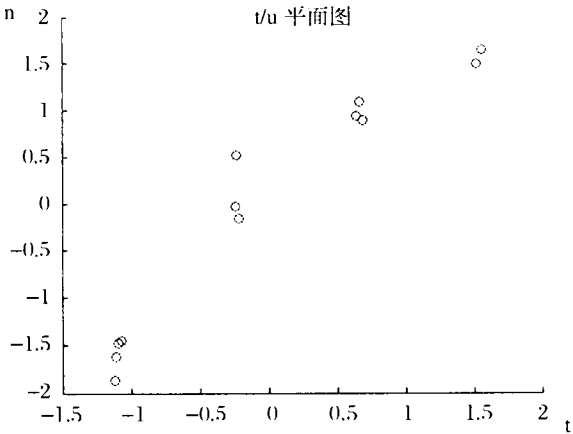


图 1 第一主成分散点(线性关系)

Fig.1 Scatter diagram of the first main component(linear relation)

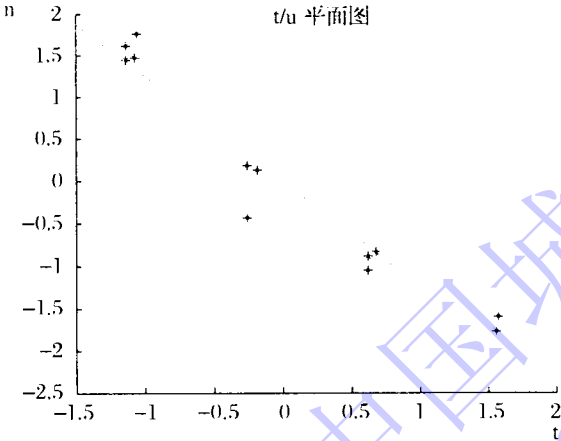


图 2 第一主成分散点(半对数关系)

Fig.2 Scatter diagram of the first main component(semilogarithmic system)

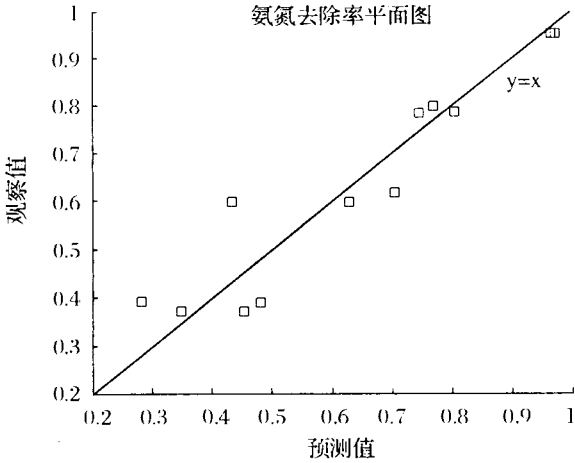
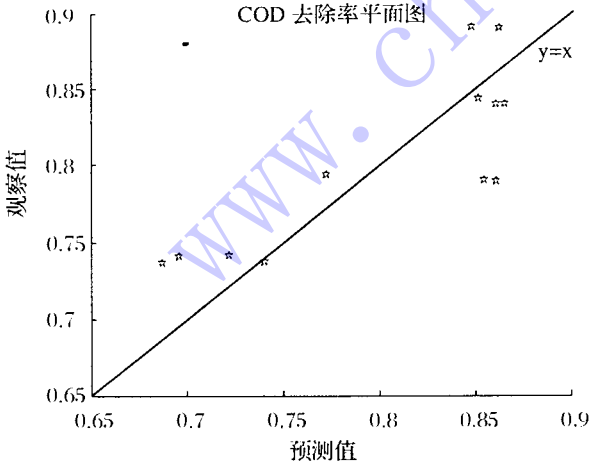


图 3 预测值——观察值平面

Fig.3 Forecasting value - plane of observe value

经过三轮计算,得到主轴见表 3。由此推出三个主成分,如第一主成分:

t₁ = - 0.054 7C₁^{*} - 0.023 5C₂^{*} + 0.998 2t^{*} (1)

式中: C₁^{*}、C₂^{*}、t^{*} 一分别为中心压缩化后的数据。

表 3 主轴
Tab.3 Principal axis

w ₁	w ₂	w ₃
- 0.054 7	0.792 2	0.607 8
- 0.023 5	- 0.609 6	0.792 4
0.998 2	0.029 1	0.052

根据交叉有效性及精度要求^[5]确定提取成分个数为一个,分别作 Y₁^{*}、Y₂^{*} 关于 t₁ 的回归,则有 Y['] 与 t₁ 的回归方程为:

Y₁^{*} = t₁ × r₁ (1) = 0.810 2t₁ (2)

Y₂^{*} = t₁ × r₁ (2) = 0.949 8t₁ (3)

将式(1)代入式(2)、式(3)后得关于标准化数据的回归方程,然后作相关变换得关于原始变量的回归方程如下:

Y₁ = ln(2.055 4 - 0.000 1C₁ - 0.000 4C₂ + 0.056 4t) (4)

Y₂ = ln(1.337 7 - 0.000 3C₁ - 0.001 3C₂ + 0.185 3t) (5)

面图见图 3。用同一实验中的数据进行检验,结果见表 4,可见预测效果较好。由表 5 可知,第一主成分 t_1 解释了原自变量系统中约 35% 的变异信息,同时,解释了原因变量系统中约 79% 的变异信息。即原自变量系统的 35% 的变异信息即可解释原因变量系统中高达 79% 的变异信息。第二主成分 t_2 虽能解释原自变量系统中 61.53% 的变异信息,却只能解释因变量系统中 5.53% 的变异信息,噪音现象明显。由式 4 ~ 5 及回归系数直方图 4,模型和实际情况接近。去除

率 Y 与进水浓度 C_1 、 C_2 负相关,与曝气时间 t 正相关,即进水浓度越高,要达到相同的去除率所需的时间越长。从图 4 可知,曝气时间对去除率的影响明显大于进水浓度,而 COD 的抑制作用较 $\text{NH}_3 - \text{N}$ 大。这可解释为污染物质在好氧菌的作用下降解,由于传质阻力的影响,SBR 系统需长时间曝气维持好氧菌的最佳生境。而 COD 主要为异养菌降解, $\text{NH}_3 - \text{N}$ 主要为自养菌降解,异养菌的比增长速率高于自养菌,故 COD 耗氧速率高于 $\text{NH}_3 - \text{N}$ 。

表 4 模型检验
Tab.4 Mould inspection

C_1 (mg/L)	C_2 (mg/L)	t (h)	COD 去除率			$\text{NH}_3 - \text{N}$ 去除率		
			观察值	预测值	相对误差 (%)	观察值	预测值	相对误差 (%)
547.2	50.4	8	0.855 0	0.887 3	3.77	0.974 0	0.941 1	- 3.37
531	55.2	8	0.865 0	0.887 2	2.57	0.986 0	0.940 8	- 4.59
459	54.1	4	0.850 0	0.793 2	- 6.68	0.571 0	0.613 7	7.47
386.19	69	6	0.826 7	0.843 7	2.06	0.819 7	0.798 9	- 2.54

表 5 提取成分的解释作用
Tab.5 Explanatory action of extracting component

	t_1	t_2	t_3
C_1	0.039 1	0.920 7	0.040 2
C_2	0.007 4	0.924 3	0.068 3
t	0.998 8	0.000 9	0.000 3
X	0.348 4	0.615 3	0.036 3
Y_1	0.664 9	0.107 2	0.031
Y_2	0.913 7	0.003 4	0.000 9
Y	0.789 3	0.055 3	0.015 9

注: X 、 Y — 分别代表自变量、因变量系统

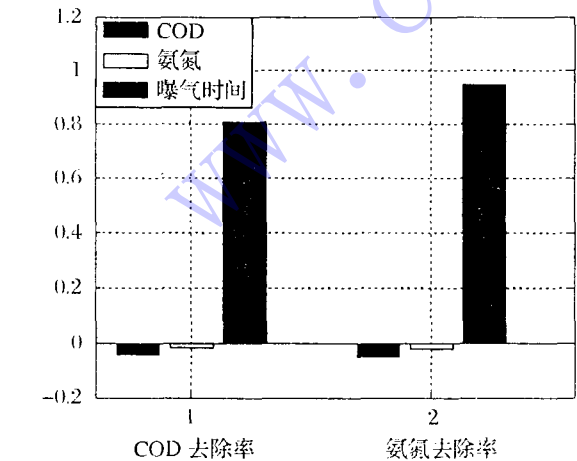


图 4 回归系数直方图
Fig.4 Histogram of regression coefficient

假设进水浓度无变化或变化很小,将式(4)对时间求导有:

$$\begin{aligned} dY/dt &= 0.056\ 4/(2.005\ 4 - 0.000\ 1C_1 - 0.000\ 4C_2 \\ &+ 0.056\ 4t) = B/(A + Bt) = 1/(\lambda + t) \end{aligned}$$

即: $\Delta t = (\lambda + t)\Delta Y$ (6)

式中: λ —浓度影响因子,常数。

将式(4)对某浓度求导,同理可得:

$$\Delta Y = \Delta C/(C - \lambda) \quad (7)$$

式中: C —进水 COD(或氨氮)浓度; λ —与进水氨氮(COD)浓度及曝气时间有关,常数。

式(6)即为进水浓度无变化或变化很小时,曝气时间变化量与去除率变化量关系,为非线性关系。式(7)即为进水浓度 COD(氨氮)浓度有变化,而氨氮(COD)进水浓度及曝气时间无变化或变化很小,去除率变化量与进水浓度变化量关系,为非线性关系。可见系数 λ 是预测曝气时间或去除率的关键因子。实际运用中可以考虑根据需要测出相应的系数 λ ,为 SBR 系统运行调控提供一定参考。

3 结论

通过对屠宰加工废水 SBR 处理系统进行进水 COD、 $\text{NH}_3 - \text{N}$ 浓度、曝气时间与 COD 及 $\text{NH}_3 - \text{N}$ 去除



率进行偏线性最小二乘法回归分析,得出以下结论:

- 1) 污染物质(COD、 $\text{NH}_3 - \text{N}$)的去除率与曝气时间强正相关,是污染物质去除的关键控制因子。
- 2) 污染物质去除率与其进水浓度负相关,其中 COD 的作用较 $\text{NH}_3 - \text{N}$ 的作用大。
- 3) 进水浓度无变化或变化很小时,曝气时间变化量与去除率变化量关系为 $\Delta t = (\lambda + t) \Delta Y$,非线性关系。
- 4) 进水 COD(氨氮)浓度有变化,而氨氮(COD)进水浓度及曝气时间无变化或变化很小,去除率变化量与进水浓度变化量关系,其关系为 $\Delta Y = \Delta C / (C - \lambda)$,为非线性关系。

参考文献

[1] SARI A. LUOSTARINEN, JUKKA A. RINTALA. Anaerobic on - site

treatment of black water and dairy parlour[J]. *Water Research*, 2005, 39: 436 - 448.

[2] 任若恩,王惠文,等.多元统计数据分析-理论、方法、实例[M].北京:国防工业出版社,1997.149 - 163.

[3] 付永胜,屠宰加工废水生物脱氮工艺过程及动力学研究[D].成都:西南交通大学,2005.

[4] 梅自量,王斌,王建平,等.宜宾市酸雨 pH 值预测的偏最小二乘法[J].环境技术,2005,(2):20 - 23.

[5] 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,1999.63 - 112.

[6] 张恒喜,郭基联,朱家元,等.小样本多元数据分析方法及应用[M].西安:西北工业大学出版,2002.67 - 135.

作者简介:张成甫(1982 -),男,重庆人,在读硕士研究生,从事环境工程。