# 基于非线性 PLSR 模型的地下水水质预测

郝 健,刘俊民,张殷钦
(西北农林科技大学 水利与建筑工程学院,陕西 杨凌 712100)

[摘 要] 【目的】针对地下水水质预测中影响因素的非线性关系,采用非线性偏最小二乘回归技术(PLSR)模型进行地下水水质预测研究,为地下水水质的准确预测提供支持。【方法】运用拟线性方法建立非线性 PLSR 模型,选用核函数对原自变量进行非线性变换,以陕西咸阳市某观测井 2001－2009 年地下水资料为研究对象,进行地下水硬度预测,并将其与 BP 网络模型的预测结果进行比较。【结果】利用咸阳市地下水前 8 年(2001－2008)的水质资料建立非线性 PLSR 模型,采用该模型对咸阳市地下水 2009 年硬度进行预测,与实测值相比,非线性 PLSR 模型、BP 网络模型预测结果的平均相对误差分别为 0.944%和 1.354%,可知非线性 PLSR 模型具有更高的预测精度和实用性。【结论】基于核函数变化的非线性 PLSR 模型,将复杂的非线性问题转化为简单的线性问题,简化了计算过程,提高了预测精度,为地下水水质的预测提供了一种新思路。

[关键词] 地下水;水质预测;PLSR;核函数;高斯函数
[中图分类号] P641.12 [文献标识码] A [文章编号] 1671-9387(2011)07-0212-05

# Prediction of groundwater quality based on nonlinear PLSR model

HAO Jian, LIU Jun-min, ZHANG Yin-qin
(*College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling, Shaanxi* 712100, *China*)

**Abstract**: 【Objective】 In light of the nonlinear relationship of the factors of the groundwater quality prediction, the nonlinear partial least-squares regression (PLSR) model is used to predict the groundwater quality, supporting the accurate prediction of groundwater quality. 【Method】 Quasilinear approach is used to build nonlinear PLSR model, the kernel function is chosen to transform each dimension of the original independent variables into new variables. Using the data Shaanxi Province, from 2001 to 2009 of groundwater of Xianyang City, groundwater hardness is predicted, and the result is compared with the result of BP network model prediction. 【Result】 The first 8 year data of water quality of Xianyang City are used to build nonlinear PLSR model to predict the 2009 groundwater hardness of Xianyang City. Compared with the measured data, the average relative error of the prediction of nonlinear PLSR model and BP network model is 0.944% and 1.354%, so nonlinear PLSR model has higher prediction accuracy and strong practicality. 【Conclusion】 The nonlinear PLSR model based on kernel function can transform the complex nonlinear problems to simple linear ones, and effectively simplify the calculation and improve the prediction accuracy, and provide a new way to predict groundwater quality.

**Key words**: groundwater; water quality; PLSR; kernel function; Gaussian function

水质预测是水资源规划、评价和管理工作的基础,也是水资源保护和利用的依据。地下水水质的预测是根据已有地下水质观测资料,通过分析处理,利用已知量寻求未知量的过程。目前,水质预测的

* [收稿日期] 2010-12-13
[基金项目] 国家科技支撑计划项目(2006BAD11B05)
[作者简介] 郝 健(1987－),男,山东枣庄人,在读硕士,主要从事水文水资源研究。E-mail:xiaoguai625@163.com
[通信作者] 刘俊民(1953－),男,陕西咸阳人,教授,博士生导师,主要从事水文水资源研究。E-mail:jmlslx@yahoo.com.cn

方法很多，例如灰色系统预测法、人工神经网络预测法、多元线性回归分析法等。王开章等[1]运用灰色理论建立 $GM(1,1)$ 模型，并用该模型对淄博大武水源地的水质指标矿化度、总硬度和 $NO_3^-$ 含量变化趋势进行了预测，预测结果与实际状况较为吻合，其特点是数据需要量少，预测精度随时间的延长而降低；Maier[2]为 Adelaide 市建立了预测河水碱度的 BP 网络，预测结果合理可靠，其特点是建模容易，预测精度高，但数学理论基础不够完善；颜剑波等[3]用建立的多元线性回归模型，对三门峡断面 COD 浓度进行了预测，预测结果有效，但是需要的数据样本容量太大。因此，为了避免需要的数据样本过多或过少、精度受限、数学理论基础不够完善等局限，作者拟引入非线性偏最小二乘回归技术（PLSR）模型进行地下水水质预测。

近年来，非线性 PLSR 方法的理论已较为完善，在水资源领域的应用范围也不断扩大，但是在地下水水质预测方面还未见涉及。本研究将 PLSR 与非线性元素结合起来，建立非线性 PLSR 模型对地下水水质进行动态预测，并采用该模型对陕西咸阳市地下水硬度进行预测，对预测结果进行研究分析，旨在简化地下水水质预测过程，提高预测精度，为有效地解决具有复杂非平稳动态特性的地下水水质预测问题提供参考。

## 1　非线性 PLSR 模型的原理及建立步骤

### 1.1　模型的原理

目前，非线性问题的建模方法很多，其中拟线性化方法是最为常用的一种，其实质是将线性方法中一些发展较为完善的技术拓展到非线性建模中，通过对原变量的适当变换，将原变量间的非线性关系转化为线性关系，再利用线性方法求解。常用于变换的基函数有对数函数[4]、样条函数[5-6]、模糊推理[7]、核函数[8]等。其中核函数是一种十分适用的曲线光滑拟合技术，而且采用高斯变换[9]后的拟合精度高，曲线光滑度好，定义清楚，适宜变换。

PLSR 方法是研究多元要素之间线性关系的常用方法，在解决自变量集合间存在多重相关性以及样本点个数少于变量个数问题方面有显著优势[10-11]，其缺点是不能有效解决非线性问题。为此，本研究将 PLSR 与非线性元素结合起来，以期解决地下水水质预测问题。

### 1.2　模型的建立步骤

在模型的建立过程中，针对多因变量 PLSR 中需要多次求解矩阵特征值和特征向量的弊端，采用单因变量 PLSR 简化算法，建立采用基于高斯函数变化的非线性 PLSR 模型[12-14]。

非线性 PLSR 模型建立步骤[15-17]如下：

第 1 步，对原自变量的每一维进行非线性变换，即对自变量空间的每一维 $x_j(j=1,2,\cdots,p)$ 进行高斯函数变换 $x_j \to z_j$。高斯函数一般形式为：

$$K(u)=\frac{1}{2\pi}\exp\left(-\frac{u^2}{2}\right)。\tag{1}$$

式中：$u$ 为任一未知数，$-\infty<u<+\infty$。

具体计算过程如下：

1）设 $\zeta_{j,l-1}$、$h_j$、$M_j$ 分别为变量区间 $x_j$ 上划分的区间分点、分段长度以及分段个数，记变量 $x_j$ 上的最小观测值为 $\min(x_j)$，最大观测值为 $\max(x_j)$，则有

$$\zeta_{j,l-1}=\min(x_j)+(l-1)h_j,\quad (l=0,1,\cdots,M_j+2)。\tag{2}$$

式中：$h_j=\dfrac{\max(x_j)-\min(x_j)}{M_j}$，$l$ 为高斯函数转换后每一维自变量的长度。

2）对 $x_j$ 作如下的高斯函数变换，有：

$$z_{j,0}=K\left(\frac{x_j-\zeta_{j,-1}}{h_j}\right),$$

$$z_{j,1}=K\left(\frac{x_j-\zeta_{j,0}}{h_j}\right),\cdots,$$

$$z_{j,M_j+2}=K\left(\frac{x_j-\zeta_{j,M_j+1}}{h_j}\right),\text{记}$$

$$z_j=(z_{j,0},z_{j,1},\cdots,z_{j,M_j+2})=K\left(\frac{x_j-\zeta_{j,l-1}}{h_j}\right)。\tag{3}$$

式中：$K(u)$ 为高斯函数，$z_j$ 为高斯函数变换后的第 $j$ 维自变量。

第 2 步，对因变量及新的自变量进行标准化处理，即：

$$\tilde{z}_{j,l}^{i}=\frac{z_{j,l}^{i}-\bar{z}_{j,l}}{s_{j,l}},\tilde{y}=\frac{y_i-\bar{y}}{s_y},(i=1,2,\cdots,n)。\tag{4}$$

式中：$\tilde{z}_{j,l}^{i}$、$\tilde{y}$ 分别是经标准化处理后的自变量和因变量，$z_{j,l}^{i}$、$y_i$ 分别为高斯函数转换后第 $j$ 维中第 $i$ 个自变量和第 $i$ 个因变量，$\bar{z}_{j,l}$、$\bar{y}$ 分别是 $z_{j,l}$、$y$ 的样本均值，$s_{j,l}$、$s_y$ 分别是样本 $z_{j,l}$、$y$ 的标准差，$i$ 为每一维自变量的个数。

记经过标准化处理后的数据为 $\tilde{z}=(\tilde{z_1},\tilde{z_2},\cdots,\tilde{z_p})$，因变量为 $\tilde{y}$，则原变量空间 $(x,y)=(x_1,x_2,\cdots,x_p,y)$ 可变换为 $(\tilde{z},\tilde{y})=(\tilde{z_1},\tilde{z_2},\cdots,\tilde{z_p},\tilde{y})$，从而得到的新数据符合线性关系，即：

$$\tilde{y}=\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\alpha_{j,l}\tilde{z}_{j,l}+\varepsilon。\tag{5}$$

式中:$p$ 为自变量的维数,$\varepsilon$ 为计算误差。

第 3 步,用单因变量 PLSR 简化算法对式(5)进行回归,求得回归系数 $\alpha_{j,l}$。

第 4 步,将式(4)代入式(5),得:

$$\frac{y_i-\bar{y}}{s_y}=\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\alpha_{j,l}\frac{z_{j,l}^{i}-\bar{z}_{j,l}}{s_{j,l}}+\varepsilon。\tag{6}$$

继而得到

$$y=\beta_0+\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}\,z_{j,l}+\varepsilon。\tag{7}$$

式中:$\beta_0=\bar{y}-\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}\bar{z}_{j,l}$,(8)

$$\beta_{j,l}=s_y\frac{\alpha_{j,l}}{s_{j,l}}。\tag{9}$$

第 5 步,将回归系数及核函数变换式(3)代入到式(7)中,可得到 $y$ 关于 $x$ 的非线性回归预测模型为:

$$\hat{y}=\beta_0+\sum_{j=1}^{p}f_j(x_j)=\beta_0+\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}K\left(\frac{x_j-\zeta_{j,l-1}}{h_j}\right)+\varepsilon。\tag{10}$$

## 2 实例分析

为了检验非线性 PLSR 模型的实际预测效果,本研究根据陕西咸阳市某观测井 2001—2009 年(每年 3 月和 9 月监测 2 次)的地下水水质监测资料(表 1),对咸阳市地下水的硬度进行分析。按照中国地下水质量国家标准对咸阳市地下水水质分类可知,咸阳市地下水为Ⅲ类水,属于硬水。硬度过高的水对锅炉威胁很大,形成水垢后既浪费燃料还会引发爆炸,同时饮用硬度过高的水还会引起肠胃不适等,因此对水硬度的预测与分析是水质研究中的重要内容之一。

**表 1 陕西咸阳市地下水水质的监测资料**

Table 1 Monitoring data of groundwater quality of Xianyang City, Shaanxi Province

| 年份 Year | 月份 Month | pH | 阳离子/(mg·L$^{-1}$) Cation | | 阴离子 $HCO_3^-$/(mg·L$^{-1}$) Anion $HCO_3^-$ | 总硬度(德国度)/°G Total hardness (German degrees) |
|---|---|---|---|---|---|---|
| | | | $Ca^{2+}$ | $Mg^{2+}$ | | |
| 2001 | 3 | 7.10 | 106.0 | 15.8 | 439.0 | 18.5 |
| | 9 | 7.30 | 53.9 | 24.6 | 559.0 | 13.2 |
| 2002 | 3 | 6.00 | 72.3 | 14.3 | 314.0 | 13.4 |
| | 9 | 7.30 | 72.4 | 19.4 | 324.0 | 14.6 |
| 2003 | 3 | 7.50 | 74.1 | 12.3 | 306.0 | 13.2 |
| | 9 | 7.80 | 53.9 | 40.9 | 359.0 | 17.0 |
| 2004 | 3 | 7.52 | 94.2 | 24.6 | 372.0 | 18.8 |
| | 9 | 7.30 | 87.6 | 22.5 | 407.0 | 17.4 |
| 2005 | 3 | 7.40 | 88.8 | 143.6 | 342.9 | 45.5 |
| | 9 | 7.30 | 92.2 | 26.8 | 341.7 | 19.1 |
| 2006 | 3 | 7.50 | 104.2 | 21.9 | 341.7 | 19.6 |
| | 9 | 7.40 | 124.2 | 141.1 | 488.2 | 49.9 |
| 2007 | 3 | 7.45 | 133.0 | 40.0 | 349.0 | 27.9 |
| | 9 | 7.30 | 147.0 | 42.2 | 438.0 | 30.3 |
| 2008 | 3 | 7.70 | 147.0 | 34.8 | 427.0 | 28.6 |
| | 9 | 7.70 | 129.0 | 32.2 | 409.0 | 25.5 |
| 2009 | 3 | 7.70 | 131.0 | 23.6 | 401.0 | 23.7 |
| | 9 | 7.30 | 106.0 | 27.2 | 322.0 | 21.1 |

注(Note):1 °G=10 mg/L CaO。

### 2.1 非线性 PLSR 模型的建立

水的硬度主要是指水中含有的可溶性钙盐和镁盐,可分为碳酸盐硬度(即通过加热能以碳酸盐形式沉淀下来的钙、镁离子,故又叫暂时硬度)和非碳酸盐硬度(即加热后不能沉淀下来的那部分钙、镁离子,又称永久硬度)两大类。上述暂时硬度和永久硬度合称为总硬度[18]。根据 pH、$Ca^{2+}$、$Mg^{2+}$、$HCO_3^-$ 对水硬度的非线性影响,将 pH、$Ca^{2+}$、$Mg^{2+}$、$HCO_3^-$ 作为自变量,水硬度作为因变量,根据咸阳市某观测井地下水水质前 8 年(2001—2008 年)数据建立模型。取 $M_j(j=1,2,3,4)=4$,根据公式(1)~(10),得到如下计算结果:

$$\alpha_{j,l}=\begin{pmatrix} 0.012\,2 & -0.018\,0 & 0.051\,7 & -0.020\,4 & 0.031\,2 & 0.034\,1 & 0.025\,9\\ -0.186\,2 & -0.042\,4 & -0.044\,7 & -0.001\,6 & 0.029\,3 & 0.059\,4 & 0.091\,7\\ -0.149\,3 & -0.162\,9 & -0.108\,6 & 0.067\,2 & 0.156\,8 & 0.148\,3 & 0.147\,7\\ -0.009\,2 & 0.001\,6 & 0.008\,1 & -0.034\,4 & 0.029\,7 & 0.017\,3 & -0.008\,8 \end{pmatrix},$$

$$\beta_{j,l}=\begin{pmatrix} 5.599\,3 & -5.041\,5 & 24.095\,3 & -6.216\,8 & 8.508\,0 & 9.050\,1 & 10.616\,5 \\ -63.858\,0 & -8.041\,5 & -8.203\,5 & -0.342\,8 & 5.857\,7 & 10.618\,9 & 30.521\,8 \\ -55.129\,1 & -35.177\,8 & -24.810\,1 & 32.197\,9 & 53.930\,8 & 30.274\,8 & 51.412\,5 \\ -3.170\,2 & -0.283\,1 & -1.979\,3 & -7.261\,4 & 6.619\,3 & 4.495\,5 & -4.042\,2 \end{pmatrix},$$

$$\beta_0=28.586\,6。$$

将上述参数代入所建立的预测模型，利用编写的程序即可进行预测分析。

## 2.2　非线性 PLSR 模型的预测结果

利用已建立的模型对咸阳市 2009-03 和 2009-09 地下水的硬度进行预测，并与 BP 网络预测模型的预测结果进行对比，结果见表 2。由表 2 可以看出，利用本研究建立的非线性 PLSR 模型，预测咸阳市 2009-03 和 2009-09 地下水的硬度值分别为 23.524和 20.858 °G，与实测值的相对误差分别为 0.743%和 1.145%，平均相对误差为 0.944%。而利用 BP 网络模型预测时，平均相对误差为 1.354%。可见，利用非线性 PLSR 模型进行预测具有较高的精度。

**表 2　基于非线性 PLSR 模型和 BP 网络模型的陕西咸阳市地下水硬度实测值及预测值的比较**

Table 2　Comparison of measured results and predictions of groundwater hardness of Xianyang City, Shaanxi Province based on nonlinear PLSR model and BP network model

| 年份 Year | 月份 Month | 实测值/°G Measured valane | 非线性 PLSR 模型 Non-linear PLSR model | | BP 网络模型 BP network model | |
|---|---|---|---|---|---|---|
| | | | 预测值/°G Predictive value | 相对误差/% Relative error | 预测值/°G Predictive value | 相对误差/% Relative error |
| 2009 | 3 | 23.7 | 23.524 | 0.743 | 24.207 | 2.138 |
| | 9 | 21.1 | 20.858 | 1.145 | 21.220 | 0.570 |

# 3　结　论

本研究建立的非线性 PLSR 模型，是将 PLSR 与非线性元素结合起来，通过核函数将原变量间的非线性关系转化为线性关系，从而进行求解。与 BP 网络模型相比，该模型的预测精度较高，具有一定的实用性。此外，基于核函数变化的非线性 PLSR 模型，不但适用于地下水水质的预测，可能还适用于地下水动态、大气环境质量预测等复杂的非平稳特性的非线性问题。

此外，本研究建立的非线性 PLSR 模型也有其不足之处。由于地下水硬度的诸多影响因素间没有明显的相关关系或相关关系很小，甚至有些因素与地下水硬度变化的关系还不清楚，因此增加了判断的难度，如果简单地将 pH、$Ca^{2+}$、$Mg^{2+}$、$HCO_3^-$ 4 种因素作为回归分析要素，会在一定程度上影响预测结果的客观性，但若分析要素过多，则又会增加计算量，而且还会引入一些次要或无关的信息，反而使分析误差加大。究竟选用哪些因素作为回归分析要素，需要根据具体情况进行分析。

## [参考文献]


[1]　王开章，刘福胜，孙　鸣. 灰色模型在大武水源地水质预测中的应用 [J]. 山东农业大学学报：自然科学版，2002，33(1)：66-71.
Wang K J，Liu F S，Sun M. The application of grey model in Dawu water quality predication water resource site [J]. Journal of Shandong Agricultural University：Natural Science Edtion，2002，33(1)：66-71. (in Chinese)

[2]　Maier H R. The use of artificial neural networks for the prediction of water quality parameters [J]. Water resources research，1996，32(4)：1013-1022.

[3]　颜剑波，阮晓红，孙　瀚. 多元回归分析在黄河水质预测中的应用 [J]. 人民黄河，2010，32(3)：35-36.
Yan J B，Ruan X H，Sun H. The application of Yellow River water quality prediction by multiple regression analysis [J]. Yellow River，2010，32(3)：35-36. (in Chinese)

[4]　朱坚民，宾鸿赞，王中宇，等. 测量数据粗大误差的非统计判别 [J]. 华中理工大学学报，2000，28(4)：17-19.
Zhu J M，Bin H Z，Wang Z Y，et al. Non-statistical distinguish of gross errors in measurement values [J]. Journal of Huazhong University of Science and Technology，2000，28(4)：17-19. (in Chinese)

[5]　Durand J F. Local polynomial additive regression through PLS：PLSS [J]. Chemometrics and Intelligent Laboratory Systems，2001，58：235-246.

[6]　吕慧刚，张凤鸣，钟　忠. 基于样条变换的非线性 PLSR 方法及应用 [J]. 系统工程与电子技术，2008，30(10)：1999-2002.
Lü H G，Zhang F M，Zhong Z. Nonlinear PLSR method and its application based on spline transform [J]. Systems Engine ering and Electronics，2008，30(10)：1999-2002. (in Chinese)

[7]　Bang Y H，Yoo C K，Lee I B. Nonlinear PLS modeling with fuzzy inference system [J]. Hemometrics and Intelligent Laboratory Systems，2003，64：137-155.

[8]　王　珏，石纯一. 机器学习研究 [J]. 广西师范大学学报：自然

科学版,2003,21(2):1-15.
Wang Y,Shi C Y. Investigations on machine learning [J]. Journal of Guangxi Normal University: Natural Science Edtion, 2003,21(2):1-15. (in Chinese)

[9] 王惠文,吴载斌,孟 浩. 偏最小二乘回归的线性与非线性方法[M]. 北京:国防工业出版社,2006.
Wang H W,Wu Z B,Meng H. Partial least-squares regression linear and nonlinear methods [M]. Beijing: National Defence Industry Press,2006. (in Chinese)

[10] 梁炳仁. 关联分析法在地下水动态分析中的应用 [J]. 地下水,1990(4):110-112.
Liang B R. Application of correlation analysis in groundwater analysis [J]. Groundwater,1990(4):110-112. (in Chinese)

[11] Srensen H A,Petersen M K,Jacobsen S,et al. Mass spectrometry and partial least-squares regression:a tool for identification of wheat variety and end-use quality [J]. Journal of Mass Spectrometry,2004,39(6):601-612.

[12] 王海燕,卢 山. 非线性时间序列分析及其应用 [M]. 北京:科学出版社,2006.
Wang H Y,Lu S. Nonlinear times series analysis and its applications [M]. Beijing:Science Press,2006. (in Chinese)

[13] [美]Douglas M Bates,[加拿大]Donald G Watts. 非线性回归分析及其应用 [M]. 韦博成,万方焕,朱宏图,译. 北京:中国统计出版社,1997.
[America]Douglas M B,[Canada]Donald G W. Nonlinear regression analysis and its applications [M]. Translated by Wei B C,Wan F H,Zhu H T. Beijing:China Statistics Press,1997. (in Chinese)

[14] 刘兴堂,梁炳成,刘 力,等. 复杂系统建模理论、方法与技术[M]. 北京:科学出版社,2008.
Liu X T,Liang B C,Liu L,et al. The theory,method & technique for complex system modeling [M]. Beijing: Science Press,2008. (in Chinese)

[15] 刘玉邦,梁 川. 基于核函数变换的非线性 PLSR 模型在叶水势预测中的应用 [J]. 水资源与水工程学报,2010,21(4):84-88.
Liu Y B,Liang C. Application of non-linear PLSR model to the prediction of leaf water potential based on kernel function transformation [J]. Journal of Water Resources and Water Engineering,2010,21(4):84-88. (in Chinese)

[16] 刘玉邦,梁 川. 地下水动态水位预测的非线性 PLSR 方法[J]. 武汉理工大学学报,2010,32(4):127-130.
Liu Y B,Liang C. Application of method of nonlinear PLSR in groundwater level prediction [J]. Journal of Wuhan University of Technology,2010,32(4):127-130. (in Chinese)

[17] 孟 洁,王惠文,黄海军,等. 基于核函数变换的 PLS 回归的非线性结构分析 [J]. 系统工程,2004,22(10):93-97.
Meng J,Wang H W,Huang H J,et al. Nonlinear structure analysis with partial least-squares regression based on kernel function transformation [J]. Systems Engineering, 2004, 22(10):93-97. (in Chinese)

[18] 谢协忠,张钰镭,于瑞生,等. 水分析化学 [M]. 南京:河海大学出版社,2007.
Xie X Z,Zhang Y L,Yu R S,et al. Water analytical chemistry [M]. Nanjing:HoHai University Press,2007. (in Chinese)