

改进多元回归法与神经网络应用于水质预测

袁 健¹, 树 锦²

(1. 河海大学水利水电工程学院, 江苏 南京 210098; 2. 河海大学环境科学与工程学院, 江苏 南京 210098)

摘要: 为提高水质预测模型的精度, 在传统算法的基础上, 提出了改进多元非线性回归算法, 并以黄河干流龙门至潼关监测断面的水质预测为例, 与神经网络算法进行了比较计算。结果表明, 两种模型均可用于黄河水质预测, 改进多元非线性回归模型预测效果略优于人工神经网络模型。

关键词: 多元非线性回归模型; Monte Carlo 法; 神经网络模型; Levenberg-Marquardt 算法; 水质预测

中图分类号: X3 **文献标识码**: A **文章编号**: 1004-693X(2008)03-0046-03

Application of modified multivariate nonlinear regression analysis model and neural network model to water quality prediction

YUAN Jian¹, SHU Jin²

(1. College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China; 2. College of Environmental Science and Engineering, Hohai University, Nanjing 210098, China)

Abstract: To improve the precision of traditional water quality models, a modified multivariate nonlinear regression analysis model was developed. The model was applied to the prediction of water quality in the Longmen to Tongguan section of the Yellow River and compared with values that were generated using neural network technology. Results show that both models are suitable for water quality prediction, but the modified multivariate nonlinear regression analysis model turns out to be better.

Key words: multivariate nonlinear regression analysis model; Monte Carlo method; neural network model; Levenberg-Marquardt algorithm; water quality prediction

水质预测是在水污染控制单元内建立水域功能区水质指标与陆域相应污染源之间的输入响应关系, 以便为水质目标责任管理提供科学依据^[1]。基于物质质量、能量和动量守恒原理以及物理生化规律所建立的分析模型的物理概念清晰, 但限于计算量及参数识别等困难, 实用中常对模型结构进行简化与假定, 从而影响了模型的精度及实用性。为此, 需引入新的方法来建立高精度及实用性强的水质预测模型, 以适应水环境管理的迫切要求。

1 改进多元非线性回归模型

1.1 回归模型的建立

依据数学中的多元非线性函数逼近理论, 工程中的大多复杂多元函数, 理论上都可以用多元幂级

数形式充分逼近, 幂级数形式是自变量的各阶乘幂与两两、三三、……交叉的线性和。

这类多元非线性回归模型对于待求参数而言都是线性的, 通过变量变换可转化为线性形式, 就其本质而言还是线性的回归模型。目前处理这类问题多采用最小二乘法确定待求参数, 再用逐步回归算法, 逐步剔除弱影响因子, 得到回归模型的最终模型。

本文采用的改进多元非线性回归方法, 在前期数据处理中对自变量进行累进变换和充分筛选, 有效地消除模型误差。同时, 引进合理的非线性项, 刻画物理过程的非线性扰动。最后, 依据各项两两间散点图, 确定各线性项、非线性项与因变量的关系, 选取与因变量线性趋势显著的各项, 按线性组合加上常数项组成广义线性拟合模型。这种模型只选取

趋势显著的各项,有助于从理论上分析客观存在的主要(非线性)关系。

1.2 回归模型的模拟检验

多元非线性回归模型建立后,还需对该回归模型预测效果进行检验。通常的方法是把已知的观测数据分成两部分,用较多的一部分建立回归模型,用较少的部分来检验预测效果。这就对已知数据的样本空间提出了较高要求,当样本空间容量不大,即观测数据量不足时,用这种检验方法得出的检验结果极易失真。在这种情况下,可采用 Monte Carlo 法检验回归模型的有效性^[2]。具体过程如下:

a. 设随机变量 y 与多个自变量 $x_1, x_2, \dots, x_m$ ($p > 1$) 有统计关系,这里 $x_1, x_2, \dots, x_m$ 只是记号,可能是前面选择得来的非线性项,假设自变量 $x_1, x_2, \dots, x_m$ 之间不存在完全的或近似完全的线性相关关系,且假设

$$y_n = \beta_0 + \sum_{i=1}^m \beta_i x_{ni} + \varepsilon_n \tag{1}$$

式中 $\beta_0, \beta_1, \dots, \beta_m$ 都是与 $x_1, x_2, \dots, x_m$ 无关的未知参数, ε_n 为误差。

将已知的 N 组观测数据 ($x_{n1}, x_{n2}, \dots, x_{nm}, y_n$) 代入式 (1) 进行回归分析,并求得复相关系数

$$R = \sqrt{1 - \frac{\sum_{n=1}^N (y_n^* - y_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}} \quad (n = 1, 2, \dots, N) \tag{2}$$

式中 y_n^* 为因变量 y 由回归模型得出的预测值。

b. 用 Monte Carlo 法产生一组服从 $N(\bar{z}, \sigma_z^2)$ 分布的正态随机数 $z_1, z_2, \dots, z_N$,

$$\bar{z} = \frac{1}{N} \sum_{n=1}^N y_n \quad \sigma_z^2 = \frac{1}{N-1} \sum_{n=1}^N (y_n - \bar{y})^2$$

c. 用 $z_1, z_2, \dots, z_N$ 替换式 (1) 中的 $y_1, y_2, \dots, y_N$ 进行回归分析,得到新因变量 z 的回归模型,并代入式 (2) 求得相应的复相关系数 R 。

d. 重复上述 (b) (c) 步骤,得到一组复相关系数 R 的数值,统计假设检验可获知 R 的分布密度函数。对给定的显著性水平 α ,若 $P(R > R^*) > \alpha$ 成立,式中 R^* 为步骤 (a) 中由原始数据求得的复相关系数,则认为所得回归预测的效果和随机预测的效果相比差异不显著,进而可以拒绝接受该回归模型;反之,接受回归模型。

1.3 基本步骤

a. 分析各自变量与因变量间的关系,线性化其中的非线性关系。以变换后的新自变量与未变换的原自变量(线性项)组成一阶项。

将各一阶项分别自身相乘及交叉相乘组成二阶

项,视需要可用各一阶项与各二阶项分别相乘组成三阶项,二、三阶及以上项均为非线性项。

b. 分析各线性项、非线性项与因变量的关系,选取与因变量线性趋势显著的各项,按线性组合加上常数项组成广义线性拟合模型。

c. 用最小二乘法确定拟合模型系数,并用 Monte Carlo 法模拟检验回归模型的有效性。

d. 依据最终回归模型,进行水质预测。

2 人工神经网络模型

2.1 人工神经网络的 LM 算法

人工神经网络模型摒弃了任何人为因素,只根据水质标准本身的特点进行学习,从而产生连接权值矩阵和阈值矩阵,这是与传统水质预测方法的根本区别。当将测试样本输入模型后,网络将自动地调用权值与阈值,进行回想或联想,从而得到测试样本的输出结果(预测结果)。就此而言,其输出结果与其他需人为确定权值矩阵的评价方法的输出结果相比具有更真实的客观性。

工程中常用 Levenberg-Marquardt 优化法(以下简称 LM 法),它比 trainbp 和 trainbpx 函数使用的梯度下降法要快得多,但它需要更多的内存。更新参数的 LM 学习规则,即:

$$\Delta w = (J^T J + \mu I)^{-1} J^T e$$

其中变量 μ 确定了学习算法是根据牛顿法还是梯度法来完成,随着 μ 的增大,式中 $J^T J$ 可以忽略。因此学习过程主要根据梯度下降,即 $\mu^{-1} J^T e$ 项。只要迭代使误差增加, μ 也就会增加,直到误差不再增加为止。调用 trainlm 算法后,可以得到每个网络误差对网络层每个权值的导数的雅可比(jacobian)矩阵^[3]。

2.2 基本步骤

神经网络水质预测分为学习与工作两个阶段^[4]。前者精选已知的实例作为样本来训练网络,后者则依据训练好的网络输出相应于某影响因子的水质指标值。其步骤如下:①通过物理成因及统计分析,确定预测对象及其影响因子;②利用样本集进行网络的训练与测试,直至满意为止;③依据既定的网络,进行具体的水质预测。

3 实例分析

3.1 基本资料

如图 1 所示,对黄河干流重点监测断面潼关断面进行水质建模预测。根据排入龙门至潼关河段内主要污染物浓度及各污染物在 GB 8978—1996 标准中Ⅰ类水质标准,通过各污染源及各污染物的等标

污染指数、等标污染负荷以及各污染物污染负荷比的大小,对河段内的主要污染物进行综合评价,以氨氮为黄河潼关段主要污染物进行预测分析。

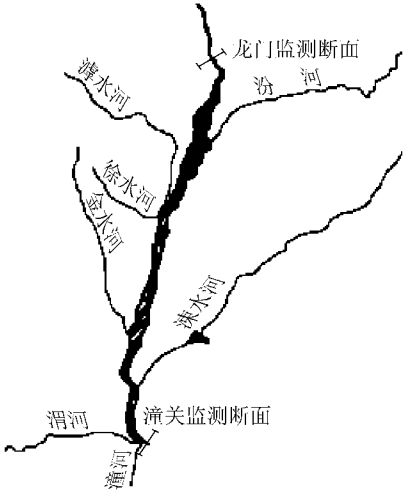


图 1 龙门至潼关水质监测断面分布

资料^[5]中,可能对氨氮浓度产生影响的因素有月平均流量、月平均输沙量、溶解氧、高锰酸盐指数、化学需氧量等。经分析,最终取龙门断面氨氮浓度、流量、溶解氧及区间汇入氨氮量 4 个变量作为潼关断面氨氮浓度影响因子建模预测。但 1999 年以前资料缺乏“溶解氧”等指标,为有效利用新资料及完整考虑各指标及其耦合对黄河水质的影响,本文选用 2000~2002 年的相关数据建模,并按月统计,共获得 36 组数据。

3.2 预测效果

分别采用改进多元非线性回归方法与神经网络 LM 算法,对 2003 年潼关断面氨氮浓度建模预测,并用 Monte Carlo 法对回归模型进行检验,复相关系数服从正态分布,在显著性水平 $\alpha = 0.1$ 条件下接受该回归模型,预测效果见表 1。

表 1 2003 年潼关断面氨氮质量浓度

月份	实测值/ ($\text{mg}\cdot\text{L}^{-1}$)	改进多元非线性回归方法		人工神经网络算法	
		预测值/ ($\text{mg}\cdot\text{L}^{-1}$)	相对 误差/%	预测值/ ($\text{mg}\cdot\text{L}^{-1}$)	相对 误差/%
1	8.14	4.263	47.63	5.152	36.71
2	4.59	6.593	43.64	6.308	37.43
3	2.48	2.574	3.79	2.163	12.78
4	3.58	3.026	15.47	3.887	8.58
5	4.33	4.310	0.46	5.007	15.64
6	5.28	4.120	21.97	6.360	20.45
7	5.91	6.785	14.81	6.882	16.45
8	2.88	3.336	15.83	2.544	11.67
9	2.64	1.704	35.45	3.236	22.58
10	1.43	1.905	33.22	2.556	78.74
11	2.31	1.942	15.93	2.600	12.55
12	2.32	1.701	26.68	2.905	25.22

改进多元非线性回归法预测相对误差绝对值的

均值为 22.91%,方差为 0.022,神经网络 LM 算法预测相对误差绝对值的均值为 24.89%,方差为 0.037。

结果表明,预测结果与实测值基本接近,两种模型均可用于黄河水质预测。预测效果改进多元非线性回归模型略优于人工神经网络模型,这主要是因为改进多元非线性回归法在建模过程中既已对自变量进行了充分筛选,并引进了一些非线性项,在一定程度上刻画了水质演变的物理过程。但如不考虑溶解氧这一影响因素,而采用较多年份相关变量数据对神经网络模型进行训练,LM 算法预测精度将有所改善。

4 结 论

a. 本文分别用多元非线性回归模型与人工神经网络模型对黄河龙门至潼关断面的水质响应关系进行了预测,结果表明两种模型均可用于黄河水质预测,改进多元非线性回归模型的预测效果略优于人工神经网络模型。

b. 改进多元非线性回归方法在建模过程中既已进行累进变换和充分筛选,有效消除模型误差。与传统的多元非线性回归方法相比较,该方法物理背景明确,计算简便。

c. Monte Carlo 法是一种普遍适用的数值模拟方法,将 Monte Carlo 法应用于多元非线性回归模型的模拟检验,可有效解决小样本空间多元回归模型有效性检验的问题。

参考文献:

[1] 罗定贵,王学军,郭青.基于 MATLAB 实现的 ANN 方法在地下水水质评价中的应用[J].北京大学学报:自然科学版,2004,40(2):296-302.

[2] 刘娜仁,王天国,马砚儒. Monte Carlo 方法在《概率统计》教学中的应用[J].内蒙古民族大学学报:自然科学版,2004(8):452-453.

[3] 陈丽华,常沁春,陈兴国,等. BP 网络应用于黄河水质的预测研究[J].兰州大学学报:自然科学版,2003,39(2):53-57.

[4] 罗先香,杨建强.径向基函数网络在水环境质量评价分区中的应用[J].环境工程:自然科学版,2000,18(6):50-55.

[5] 彭勃.2002 年黄河纳污量调查分析报告[R].郑州:黄河水资源保护局,2002.

(收稿日期 2007-11-28 编辑 舒建)