模糊线性回归模型在河流水体总氮浓度预测中的应用*

周九州 刘 强** 荣湘民 彭建伟 谢桂先
(湖南农业大学资源环境学院,长沙 410128)

摘 要 以湘江熬洲断面为例,将该断面水体中总氮浓度及其有关影响因子用三角模糊数来表征。同时,结合已有的模糊线性回归模型成果,构造了带有三角模糊参数的水体中总氮浓度模糊线性回归预测模型。并应用所建模型预测该断面水体中2002—2005年总氮浓度,所得的预测值与已有的实测值之间的相对误差均小于20%,完全满足实际应用对误差的要求,预测合格率为100%,说明这种预测模型在预测河流水体总氮浓度变化中有一定的实用性,为今后开展河流水体中污染物浓度预测提供了新途径。

关键词 河流;总氮浓度;模糊线性回归;预测
中图分类号 X522 **文献标识码** A **文章编号** 1000-4890(2009)12-2628-05

Application of fuzzy linear regression model in predicting river water total nitrogen concentration. ZHOU Jiu-zhou, LIU Qiang, RONG Xiang-min, PENG Jian-wei, XIE Gui-xian (*College of Resources and Environment, Hunan Agricultural University, Changsha* 410128, *China*). *Chinese Journal of Ecology*, 2009, **28**(12): 2628-2632.

Abstract: Taking the Aozhou section of Xiangjiang River as a case, the total nitrogen concentration in water body and related affecting factors were characterized with triangular fuzzy number, and the triangular fuzzy parameters were introduced into the existing fuzzy linear regression model to predict the total nitrogen concentration in water body. The relative errors between the predicted and measured values of total nitrogen concentration in the section in 2002-2005 were less than 20%, with a qualified prediction rate being 100%, which suggested that the modified fuzzy linear regression model had definite practicability in predicting the total nitrogen concentration in river water.

Key words: river; total nitrogen concentration; fuzzy linear regression; prediction.

模糊数学作为一门新的数学分支,以"模糊集合论"为基础,提供了一种处理不确定性和不精确性问题的新方法,目前正广泛用于工业(刘易平和马邕文,2008)、农业(方东权和吴天吉,2008;刘松涛等,2008)、环境(侯素霞等,2008;王博等,2008;朱红玉等,2008)、医学(Lee *et al.*,2001)、机械、电子、管理、交通(向红艳和肖盛燮,2006)等领域。基于采用模糊数学改进后的线性回归模型在电力(耿光飞和郭喜庆,2002;游仕洪等,2006)、用水量(陈南祥和徐海洋,2007;孟丽丽等,2008)、发病率(程利军和高丽,2000)、水文(Bardossy *et al.*,1990)等预测中取得了比较好的预测效果。

本文以湘江熬洲断面为例,将该断面水体中1990—2001年共计12年的总氮浓度实测数据以及影响该断面总氮浓度的控制区域内农药使用量、氮肥使用量、生活污水和工业废水中氨氮总量、大家畜数量、家禽数量、水产品数量、总人口数以及湘江流入研究区域的断面总氮浓度等8个相关因子数据用三角模糊数来表征,结合已有模糊线性回归模型成果(于九如和杨泽华,1995;向阳,1998;吴冲等,2000;曾文艺等,2006),构造成带有三角模糊参数的水体中总氮浓度模糊线性回归预测模型。该方法通过将模型参数模糊化,达到弱化模型对历史数据准确度的依赖,能在历史数据不精确、不全面的情况下,获得较好的预测效果。

* 湖南省教育厅重点项目(05A024和07A028)和国家科技支撑计划资助项目(2007BAD87B11和2008BADA7B07)。
** 通讯作者 E-mail:lq8053@hunau.net
收稿日期:2009-04-21 接受日期:2009-08-15

1 模糊线性回归模型

1.1 模糊线性回归模型

和经典的线性回归分析类似，设变量 y 与它的相关因素 $x_1, x_2, \cdots, x_n$ 有线性关系：

$$y = A_1x_1 + A_2x_2 + \cdots + A_nx_n \tag{1}$$

回归分析的问题是利用已知的 n 组观测数据 $(y_j, x_{1j}, x_{2j}, \cdots, x_{nj}, j = 1, 2, \cdots, m)$，去估计回归系数 A_j。但在模糊线性回归分析中，认为模型具有模糊性，即回归系数 A_j 是模糊数，于是模型的拟合值 $\bar{y}_i$ 与观测值 y_i 之间的偏差是由这种模糊性引起的。

通常取 A_j 为三角模糊数 $A(\alpha, c)$，其隶属函数：

$$\mu_A(z) = \begin{cases} 1 - \dfrac{|z - \alpha|}{c}, & \alpha - c \leqslant z \leqslant \alpha + c \\ 0, & \text{其它} \end{cases} \tag{2}$$

三角模糊数 $A(\alpha, c)$ 有通常的线性性质，即：

$$\begin{cases} kA(\alpha, c) = A(k\alpha, kc), & k > 0 \\ A_1(\alpha_1, c_1) + A_2(\alpha_2, c_2) = A(\alpha_1 + \alpha_2, c_1 + c_2) \end{cases} \tag{3}$$

得出(1)式中 y 的隶属函数为

$$\mu_y(y_i) = \begin{cases} 1 - \dfrac{|y_i - \sum\limits_j \alpha_i x_{ji}|}{\sum\limits_j c_j |x_{ji}|}, & \sum\limits_j \alpha_i x_{ji} - \sum\limits_j c_j |x_{ji}| \leqslant y \leqslant \sum\limits_j \alpha_i x_{ji} + \sum c_j |x_{ji}| \\ 0, & \text{其它} \end{cases} \tag{4}$$

为使拟合函数(1)对已知的 n 组观测数据：$(y_j, x_{1j}, x_{2j}, \cdots, x_{nj}, j = 1, 2, \cdots, m)$ 拟合最好，在FLR分析中，必须同时满足下述2个准则：

1) 必须使各回归系数的模糊幅度之和最小(即精度最大)，即：

$$\min J = c_1 + c_2 + \cdots + c_n \tag{5}$$

2) 按一定的置信水平 H，必须能“覆盖”所有的观测数据 y_i，即满足：

$$\mu_i(y_i) \geqslant H, \quad 0 \leqslant H \leqslant 1 \tag{6}$$

这在模糊集理论中称为“H 水平截集”，或通俗地称为“门槛”(图1)，保证了没有隶属度 $< H$ 的 y_i。

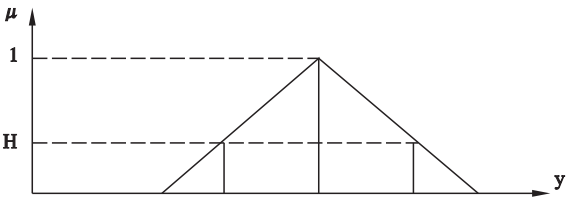

图1 置信水平 H
Fig. 1 Confidence level (H)

根据式(4)和式(6)可得：

$$1 - \frac{y_i - \sum\limits_j \alpha_j x_{ji}}{\sum\limits_j c_j |x_{ji}|} \geqslant H$$

$$\sum_j \alpha_j x_{ji} - (1 - H) \sum_j c_j |x_{ji}| \leqslant y_j \leqslant \sum_j \alpha_j x_{ji} + (1 - H) \sum_j c_j |x_{ji}| \tag{7}$$

结合式(5)和式(7)，求解FLR(1)的问题转化为求解线性规划：

$$\begin{cases} \min J = c_1 + c_2 + \cdots + e_n \\ s,t \begin{cases} \sum\limits_j \alpha_j x_{ji} - (1 - H) \sum\limits_j c_j |x_{ji}| \leqslant y_i \\ \sum\limits_j \alpha_j x_{ji} + (1 - H) \sum\limits_j c_j |x_{ji}| \geqslant y_i \end{cases} \end{cases} \tag{8}$$

式(8)的解中 $A_j(\alpha_j, c_j)$ 可能有下述4种情形：①$\alpha_j \neq 0, c_j > 0$；②$\alpha_j \neq 0, c_j = 0$；③$\alpha_j = 0, c_j > 0$；④$\alpha_j = 0, c_j = 0$。和经典回归一样，①和②表示 y 与 x_j 相关，但在②A_j 为精确数。③和④表示 y 与 x_j 不相关，可以在模型(1)中将 x_j 项剔除，但对于③，如果将 x_j 剔除，可能对模型有微小的影响。

1.2 模型的评价

式(1)求解后，可按下述标准评价观测值拟合的优劣，即：① 各观测值 y_i 对模型的隶属度 $\mu_i(y_i)$，一般认为各 $\mu_i(y_j)$ 的值 > 0.5，就是比较好的拟合；② 各拟合的中心值和观测值 y_i 的相对偏差：$\Omega_1 = \dfrac{|y_i - \sum\limits_j \alpha_j x_{ji}|}{y_i}$；③ 模糊幅度对观测值 y_i 的比，$\Omega_2 = \dfrac{\sum c_j |x_{ji}|}{y_i}$；$\Omega_1$ 和 Ω_2 如在30%以内，一般认为拟合是可以接受的。

同时，还可通过配对的 t 检验来考察模拟结果与实际测定值的差异性是否显著，配对 t 检验的 $S\bar{d}$ 的值计算公式为：

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}}$$

式中，

$$s_{\bar{d}} = \sqrt{\frac{\sum (d - \bar{d})^2}{(n - 1)n}} = \sqrt{\frac{\sum d_i^2 - \dfrac{(\sum d_i)^2}{n}}{(n - 1)n}}。$$

2 应用实例

2.1 研究区概况

研究区域为湘江衡阳段的宜水、耒水、蒸水、洣

水等一级支流汇水区域，控制面积约 36000 km^2。区域内气候属中亚热带季风湿润性气候，水资源丰富，多年年均降雨为 1315.11 mm，多年年均径流量为 92.33 亿 m^3，但降雨时空分布不均，多集中在 4—9 月，以暴雨形式为主，水资源中过境客水占 82.4%。区域内水系发育，河网绸密，以湘江干流为中轴，形成树形辐聚式水系格局。

2.2 影响因子的选取

Lucey 和 Goolsby（1993）对 Iowa 州的 Raccoon River 以及 Mueller 等（1997）对美国中西部河流的研究表明，河水中氮含量与上游地区的耕作面积、农作物种植面积、氮肥施用量、人口密度及气候条件等有一定的相关性。结合我国具体情况及国内有关文献（冯绍元和郑耀泉，1996；吕耀，1998；李怀恩等，2004），本研究选取对水体总氮浓度影响比较强烈的农业面源污染影响因子（主要为农药使用量、氮肥使用量、大家畜数量、家禽数量、水产品数量等）、点源污染影响因子（主要为生活污水和工业废水中氨氮总量）、控制区域内总人口数以及湘江流入研究区域的断面总氮浓度作为研究湘江熬洲断面总氮浓度变化的相关因子，农药使用量、氮肥使用量、大家畜数量、家禽数量、水产品数量以及总人口数从相关县市统计局获取，生活污水和工业废水中氨氮总量、湘江流入研究区域的断面总氮浓度从相关县市环境监测站获取。

2.3 模糊线性回归预测模型的建立

以湘江熬洲断面 1990—2001 年共计 12 年的总氮浓度及其有关影响因子建立模糊线性回归模型 $Y = A_1x_1 + A_2x_2 + \cdots + A_8x_8$，其中回归系数 A_j 为三角模糊数 $A_j(\alpha_j, c_j)$，$c_j \geqslant 0$，$(j = 1,2,\cdots,8)$，Y 为熬洲断面总氮浓度，x_1 为研究区内农药使用量，x_2 为研究区内氮肥使用量，x_3 为研究区内生活污水和工业废水中氨氮总量，x_4 为研究区内大家畜数量，x_5 为研究区内家禽数量，x_6 为研究区内水产品总产量，x_7 为研究区内总人口数，x_8 为湘江流入研究区域的断面总氮浓度。

为了求出 A_j，解下列线性规划：

$$\begin{cases} \min J = c_1 + c_2 + \cdots + c_8 \\ s,t\begin{cases} \sum_j \alpha_j x_{ji} - (1-H)\sum_j c_j \mid x_{ji} \mid \leqslant y_i, i = 1,2,\cdots,12 \\ \sum_j \alpha_j x_{ji} + (1-H)\sum_j c_j \mid x_{ji} \mid \geqslant y_i, i = 1,2,\cdots,12 \end{cases} \end{cases} \quad (9)$$

将 1990—2001 年共计 12 年的相关数据代入式（9），取 $H = 0.5$，利用 lindo 软件，解得三角模糊数 $A_j(\alpha_j, c_j)$ 见表 1。

与经典回归一样，当 $\alpha_j = 0$ 时，表示 x_j 与 Y 不相关，因此将那些 x_j 由模型中剔除（包括那些 $\alpha_j = 0$，$c_j > 0$），最后得到：

$$Y = A_1x_1 + A_2x_2 + A_3x_3 + A_4x_4 + A_7x_7 + A_8x_8 \quad (10)$$

式中，$A_1 = A_1(-0.000079, 0)$，$A_2 = A_2(-0.000014, 0.000002)$，$A_3 = A_3(0.000194, 0)$，$A_4 = A_4(0.000884, 0)$，$A_7 = A_7(0.003054, 0)$，$A_8 = A_8(0.371357, 0)$。

2.4 模糊线性回归预测模型的评价

模型建立之后，将 1990—2001 年的相关数据代入式（10），求出拟合值，并与实测数据进行比较和评价（表 2）。

由表 2 可知，除了 $\mu_i(y_j)$ 有几个指标没有达到 0.5 之外，Ω_1 和 Ω_2 的值均 < 0.3，表明所拟合的模型较好。T 检验表明，模拟结果与实际测定值的 $t = 0.068$，$t < t_{0.05(11)} = 2.201$，说明模拟结果与实际测定值之间差异性不显著，模拟效果好。

2.5 模糊线性回归预测模型的预测效果

根据所建立的模糊线性回归模型以及影响熬洲断面水体中总氮浓度的 2002—2005 年相关因素值，预测 2002—2005 年该断面水体中总氮浓度为 1.233、1.332、1.706 和 1.790 mg · L^{-1}。将熬洲断面水体中 2002—2005 年总氮浓度的预测值和已有的实测值（1.099、1.270、1.570 和 1.600 mg · L^{-1}）在 MATLAB 程序中编程，可得 2002—2005 年熬洲断面水体中总氮浓度的预测值和实测值曲线，以及预测值和实测值之间的误差曲线（图 2）。由图 2 可知，2002—2005 年熬洲断面水体中总氮浓度的预测

表 1 三角模糊数 $A_j(\alpha_j, c_j)$ 值

Tab. 1 Value of triangular fuzzy number $A_j(\alpha_j, c_j)$

j	1	2	3	4	5	6	7	8
a_j	−0.000079	−0.000014	0.000194	0.000884	0.000000	0.000000	0.003054	0.371357
c_j	0.000000	0.000002	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

表 2 水体总氮浓度拟合值与评价(mg · L^{-1})

Tab. 2 Fitted values of total N concentration in water body and evaluation

i	年份	实测值	拟合值			评价		
			中心值 $\sum ax$	模糊度 $\sum cx$	区间	$\mu_i(y_i)$	Ω_1	Ω_2
1	1990	1.083	1.205	0.146	[1.059,1.351]	0.163	0.113	0.135
2	1991	1.243	1.169	0.180	[0.989,1.349]	0.593	0.059	0.145
3	1992	1.053	1.211	0.188	[1.023,1.398]	0.158	0.150	0.178
4	1993	1.085	1.009	0.186	[0.823,1.196]	0.594	0.070	0.172
5	1994	0.772	0.871	0.189	[0.682,1.056]	0.478	0.128	0.245
6	1995	0.850	0.924	0.208	[0.717,1.132]	0.643	0.087	0.244
7	1996	1.428	1.546	0.201	[1.345,1.747]	0.412	0.083	0.141
8	1997	1.724	1.642	0.198	[1.444,1.840]	0.585	0.048	0.115
9	1998	1.407	1.324	0.203	[1.120,1.426]	0.588	0.059	0.144
10	1999	1.481	1.433	0.203	[1.230,1.636]	0.766	0.032	0.137
11	2000	1.170	1.342	0.206	[1.136,1.549]	0.164	0.148	0.176
12	2001	1.372	1.523	0.210	[1.313,1.733]	0.280	0.110	0.153

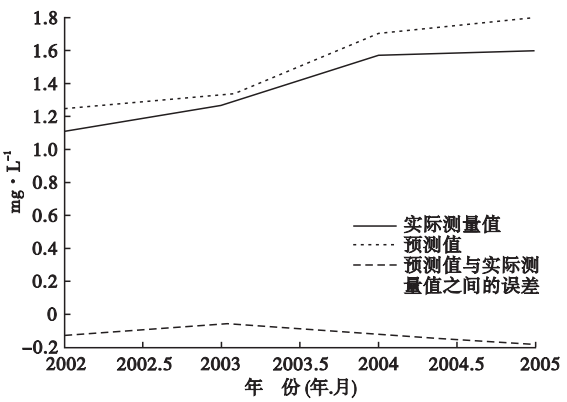

图 2 水体总氮浓度实测值与预测值之间的误差曲线

Fig. 2 Error curve between predictive and measured values of total N concentration in water body

值和实测值之间的误差小于 ±0.2 mg · L^{-1},进一步计算二者的相对误差,结果分别为 12.2%、4.9%、8.7%和 11.9%,均小于 20%,完全满足实际应用对误差的要求,预测合格率为 100%。

3 结 论

以湘江熬洲断面 1990—2001 年共计 12 年的实测总氮浓度及其有关影响因子建立模糊线性回归模型,并对所建模型进行相关评价以及对模拟结果与实际测定值进行配对 t 检验,所得结果表明该模型能较好地拟合所研究的实际观察数据。

利用所建模型对熬洲断面河流水体中 2002—2005 年总氮浓度进行预测,所得的预测值与实测值的相对误差分别为 12.2%、4.9%、8.7%和 11.9%,均小于 20%,完全满足实际应用对误差的要求,预测合格率为 100%,说明模糊线性回归模型在河流水体中总氮浓度预测方面是可行的。

参考文献

陈南祥,徐海洋. 2007. 模糊线性回归法在用水量预测中的应用. 人民黄河,29(5):33-34.

程利军,高 丽. 2000. 模糊线性回归分析在发病率预测中的应用. 滨州教育学院学报,6(3):74-75.

方东权,吴天吉. 2008. 基于模糊数学决策理论构建网络农业信息资源综合评价模型. 安徽农业科学,36(29):12979-12982.

冯绍元,郑耀泉. 1996. 农田氮素的转化与损失及其对水环境的影响. 农业环境保护,15(6):277-279.

耿光飞,郭喜庆. 2002. 模糊线性回归法在负荷预测中的应用. 电网技术,26(4):19-21.

侯素霞,刘新铭,钟 秦. 2008. 模糊数学在丹河水环境综合评价中的应用. 生态环境,17(4):1411-1414.

李怀恩,李 越,蔡 明,等. 2004. 河流水质与流域人类活动之间的关系. 水资源与水工程学报,15(1):24-28.

吕 耀. 1998. 农业生态系统中氮素造成的非点源污染. 农业环境保护,17(1):35-39.

刘松涛,赵喜茹,曹雯梅. 2008. 应用模糊数学法综合评判旱地小麦新品种. 安阳工学院学报,(4):94-96.

刘易平,马邕文. 2008. 模糊数学法在电子行业清洁生产水

平评价中的应用. 青岛科技大学学报(自然科学版), 29(5):467-470.
孟丽丽, 迟道才, 崔 屾, 等. 2008. α-加权模糊线性回归模型在参考作物需水量预测中的应用. 沈阳农业大学学报(自然科学版), 39(5):603-606.
王 博, 杨志强, 李慧颖, 等. 2008. 基于模糊数学和GIS的松花江流域水环境质量评价研究. 环境科学研究, 21(6):124-129.
吴 冲, 潘启树, 李汉铃. 2000. 模糊线性回归预测. 西安交通大学学报, 34(9):100-102.
向红艳, 肖盛燮. 2006. 模糊数学方法在交通流预测评价中的应用. 重庆交通学院学报, 25(4):106-108,112.
向 阳. 1998. 模糊回归分析及应用. 工业技术经济, 17(1):91-92,95.
于九如, 杨泽华. 1995. 模糊线性回归及其应用实例. 系统工程理论与实践, (4):32-37.
游仕洪, 程浩忠, 谢 宏. 2006. 应用模糊线性回归模型预测中长期电力负荷. 电力自动化设备, 26(3):51-53.
曾文艺, 李洪兴, 施 煜. 2006. 模糊线性回归模型(Ⅰ). 北京师范大学学报(自然科学版), 42(2):120-125.
朱红玉, 杜少少, 谷媛媛, 等. 2008. 模糊数学在地表水水质评价中的应用. 水科学与工程技术, (5):77-79.
Bardossy A, Bogardi I, Duckstein L. 1990. Fuzzy regression in hydrology. *Water Resources Research*, **26**:1497-1508.
Lee YW, Chung SY, Bogardi I, *et al*. 2001. Dose-response assessment by a fuzzy linear-regression method. *Water Science and Technology*, **43**:133-140.
Lucey KJ, Goolsby DA. 1993. Effects of climatic variations over 11 years on nitrate-nitrogen concentrations in the Raccoon River, Iowa. *Journal of Environmental Quality*, **22**:38-46.
Mueller DK, Ruddy BC, Battaglin WA. 1997. Logistic model of nitrate in streams of the upper-midwestern United States. *Journal of Environmental Quality*, **26**:1223-1230.

作者简介 周九州,男,1973年生,博士研究生,主要从事水环境N、P面源污染研究。E-mail:zjz22008@126.com

责任编辑 魏中青