

DOI :10.3876/j.issn.1000-1980.2009.03.004

投影寻踪自回归模型在长江径流量预测中的应用

于国荣¹,叶 辉²,夏自强³,赵小勇⁴

(1.昆明理工大学电力工程学院,云南 昆明 650051;2.大连普兰店市水利勘测设计院,辽宁 普兰店 116200;
3.河海大学水文水资源与水利工程科学国家重点实验室,江苏 南京 210098;4.烟台市水利局,山东 烟台 264000)

摘要:为准确预测长江径流量,引入投影寻踪自回归模型,用自相关分析技术确定预测因子,构造了新的投影指标函数,并用基于实数编码的加速遗传算法优化投影指标函数,建立了投影寻踪自回归流量预测模型.将该模型应用于长江径流量的预测,结果表明用该模型预测长江流量时序是可行和有效的.

关键词:投影寻踪自回归模型 流量预测 非线性时间序列 遗传算法 长江
中图分类号:TV121+.4 TP338+.2 **文献标识码:**A **文章编号:**1000-1980(2009)03-0263-04

受气象、人为等众多不确定性因素的影响,长江流量时序常表现出弱相依性、突变性和随机性等复杂非线性特征,属于非线性动力系统.目前常用的线性时序模型的共同特点是采用“对数据结果或分布特征作某种假定—按照一定准则寻找最优模拟—对建立的模型进行证实”这类证实性数据分析方法^[1-2].该类方法由于过于形式化且受束缚较大,往往不能充分利用资料所提供的信息,建立的预测模型不确定性较大,用于预测长江流量这类非线性、非正态系统具有一定的局限性.近 20 年来,国际统计界大都采用“直接审视数据—计算机分析模拟—设计软件程序检验”这类探索性数据分析方法^[3-5].为了更客观地获取长江流量数据的结构特征,本文引入投影寻踪自回归模型(projection pursuit auto regression model,PPAR),并用基于实数编码的加速遗传算法(real coded accelerating genetic algorithm,RAGA)来优化投影指标函数,建立投影寻踪自回归流量预测模型,解决具有非线性特征的序列问题.模型线性投影克服了“维数祸根”问题,通过一系列岭函数和的逼近,使模型精度、稳健性、实用性都得到提高,较好地解决了非正态、非线性问题,为流量时间序列分析提供了一条新的途径.同时,将该模型应用于长江径流量的预测,结果表明用投影寻踪自回归流量预测模型预测长江流量时序是可行和有效的.

1 投影寻踪自回归流量预测模型建模

目前的投影寻踪方法的计算量相当大,在一定程度上限制了投影寻踪方法的深入研究和广泛应用.投影寻踪自回归是投影寻踪技术和时间序列自回归分析的结合,即通过线性投影,将不同滞后时段的高维时间序列数据投影至低维子空间,优选出客观反映数据结构特征的岭函数,再用这些不确定形式的岭函数的和来逼近回归函数,从而构造模型.本文应用投影寻踪自回归模型时,用基于实数编码的加速遗传算法来优化投影指标函数.投影寻踪自回归模型的函数为 Hermite 多项式^[6-10],投影寻踪自回归流量预测模型建模具体过程如下:

步骤 1 确定预测因子.时序 x_i 延迟 k 步的自相关系数 R_k 为

$$R_k = \frac{\sum_{i=k+1}^n (x_i - E_x)(x_{i-k} - E_x)}{\sum_{i=1}^n (x_i - E_x)^2} \quad (k = 1, 2, \cdots, m) \tag{1}$$

$$E_x = \sum_{i=1}^n x_i/n \tag{2}$$

式中 n 为实测时序的容量, $m < n/4$. R_k 的方差随 k 的增大而增大, R_k 的估计精度随 k 的增加而降低, 因此 m 应取较小的数值. 根据 R_k 的抽样分布理论, 在置信水平 $1 - \alpha$ 的情况下, 当自相关系数值

$$R_k \notin \left[-1 - \frac{u_{\alpha/2}(n - k - 1)^{0.5}}{n - k}, -1 + \frac{u_{\alpha/2}(n - k - 1)^{0.5}}{n - k} \right] \tag{3}$$

时则推断时序 x_i 延迟 k 步相依性显著, x_{i-k} 可作为 x_i 的预测因子; 否则时序 x_i 延迟 k 步相依性不显著. 式 (3) 中的分位值 $u_{\alpha/2}$ 可从正态分布表中查得.

步骤 2 设预测对象及其预测因子的样本为 $\{x_i | i = 1, 2, \dots, n\}$ 及 $\{x_{i-k} | k = 1, 2, \dots, p; i = k + 1, k + 2, \dots, n\}$, 其中 n, p 分别为样本容量和预测因子数目, x_i, x_{i-k} 已进行标准化处理(即原数据减去样本均值再除以样本标准差), 以消除时序的量纲效应. 现在的目的就是建立 x_{i-k} 与 x_i 之间的数学关系. 投影寻踪方法是把 p 维数据 x_{i-k} 综合成以 $a = (a_1, a_2, \dots, a_p)$ 为投影方向的一维投影值 z_i :

$$z_i = \sum_{k=1}^p a_k x_{i-k} \quad (i = p + 1, p + 2, \dots, n) \tag{4}$$

其中 a 为单位长度向量.

步骤 3 对散布点 (z, x) , 用基于正交 Hermite 多项式拟合, 此时投影寻踪自回归模型为

$$\hat{x} = \sum_{i=1}^m \sum_{j=1}^r c_{ij} h_{ij}(z) \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, r) \tag{5}$$

式中: r ——多项式阶数; c ——多项式系数, 可用最小二乘法获得; h ——正交 Hermite 多项式.

步骤 4 优化投影指标函数. 在优化投影方向 a 时, 同时考虑多项式系数 c 的优化问题, 可以通过求解投影指标函数最小化问题来估计最佳 a, c 值, 即

$$\min Q(a, c) = \frac{1}{n - p_{i=p+1}} \sum_{i=p+1}^n (x_i - \hat{x}_i)^2 \tag{6}$$

$$\text{s.t.} \quad \sum_{j=1}^p a_j^2 = 1 \tag{7}$$

这是一个以 a, c 为优化变量的复杂非线性优化问题, 用传统的优化方法处理较难. 本文应用模拟生物优胜劣汰与群体内部染色体信息交换机制的基于实数编码的加速遗传算法来解决其高维全局寻优问题.

步骤 5 计算第一次的拟合残差 $r_1 = x - \hat{x}$, 如果满足要求则输出模型参数, 否则, 进行步骤 6 计算.

步骤 6 用 r_1 代替 x , 回到步骤 1 开始下一个岭函数的优化, 直到满足一定要求, 停止增加岭函数个数, 输出最后结果.

2 投影寻踪自回归流量预测模型在长江径流量预测中的应用

以 1956 ~ 2005 年长江宜昌站流量监测数据进行分析(表 1). 利用表 1 中 1956 ~ 2004 年流量资料序列 $\{x_i | i = 1, 2, \dots, 49\}$ 来建立投影寻踪自回归流量预测模型, 预留 2005 年样本作预测检验, 再对 2006 年和 2007 年流量进行预测.

计算该序列前 12 阶自相关系数 R_k 和与之相应的式 (3) 右边上、下限 R_{2k}, R_{1k} 值, 结果见表 2, 其中置信水平取 70%. 表 2 显示, 只有 R_1, R_4, R_5 的相依性在置信水平 70% 的条件下是显著的, 故这里预测 x_i 的因子取 $x_{i-1}, x_{i-4}, x_{i-5}$.

把建模样本 $\{x_i | i = 1, 2, \dots, 49\}$ 及 $\{x_{i-k} | k = 1, 4, 5; i = 6, 7, \dots, 49\}$ 进行标准化处理后依次代入式 (4)、式 (5)、式 (6) 和式 (7), 采用 1 个岭函数进行拟合, 多项式的阶数为 8, 即得此例的投影指标函数, 然后用基于实数编码的加速遗传算法优化该函数, 得最小指标函数值为 9.6303×10^5 , 参数 a, c 值见表 3.

把参数 a, c 值代入投影寻踪自回归流量预测模型中, 进行拟合检验、预测检验, 结果见表 1. 由表 1 可知: 用投影寻踪自回归流量预测模型对 44 个历史数据拟合检验中, 相对误差绝对值的平均值为 5.81%, 2005 年流量预测的相对误差为 -7.26%. 黄胜^[11]运用神经网络模型和小波网络模型对宜昌站 1991 ~ 2000 年年

表 1 流量实测值和投影寻踪自回归流量预测模型的拟合结果

Table 1 Measured runoff, fitting and predicted results of projection pursuit auto regression model									
年份	实测值/ ($\text{m}^3 \cdot \text{s}^{-1}$)	投影值	模型拟合值/ ($\text{m}^3 \cdot \text{s}^{-1}$)	相对误差/%	年份	实测值/ ($\text{m}^3 \cdot \text{s}^{-1}$)	投影值	模型拟合值/ ($\text{m}^3 \cdot \text{s}^{-1}$)	相对误差/%
1956	13 100				1981	14 000	- 0.087	13 601	- 2.85
1957	13 600				1982	14 200	- 0.377	13 708	- 3.46
1958	13 100				1983	15 100	- 0.647	14 185	- 6.06
1959	11 600				1984	14 300	0.309	13 718	- 4.07
1960	12 700				1985	14 500	0.661	13 377	- 7.75
1961	14 000	- 0.565	13 879	- 0.86	1986	12 100	0.549	13 738	13.54
1962	14 700	- 0.093	13 606	- 7.45	1987	13 700	0.113	13 531	- 1.23
1963	14 300	- 0.502	13 764	- 3.75	1988	13 300	0.752	12 916	- 2.89
1964	16 500	- 0.901	16 302	- 1.20	1989	15 100	0.366	13 784	- 8.72
1965	15 600	0.626	13 502	- 13.33	1990	14 100	0.325	13 739	- 2.56
1966	13 600	1.051	12 167	- 10.53	1991	13 700	- 0.476	13 793	0.28
1967	14 300	0.562	13 711	- 4.12	1992	13 000	- 0.092	13 605	4.65
1968	16 300	1.275	16 105	- 1.20	1993	14 600	0.050	13 527	- 7.35
1969	11 600	2.526	11 590	- 0.08	1994	11 000	0.985	11 983	8.93
1970	13 300	0.063	13 526	1.70	1995	13 400	- 0.693	14 448	7.82
1971	12 300	0.031	13 532	10.01	1996	13 300	- 0.277	13 703	3.03
1972	11 300	0.568	13 696	21.20	1997	11 500	- 0.117	13 624	18.47
1973	13 600	- 0.343	13 708	0.79	1998	16 500	- 1.124	16 217	- 1.72
1974	15 900	- 0.952	16 730	5.22	1999	15 200	- 0.219	13 686	- 9.96
1975	13 700	0.171	13 563	- 1.00	2000	14 900	0.276	13 676	- 8.21
1976	12 900	- 1.229	12 615	- 2.21	2001	13 100	- 0.388	13 709	4.65
1977	13 400	- 1.209	13 587	1.39	2002	12 500	- 0.217	13 685	9.48
1978	12 400	0.532	13 768	11.03	2003	12 900	1.178	13 659	5.89
1979	12 600	0.470	13 826	9.73	2004	13 100	0.707	13 156	0.42
1980	14 600	- 0.566	13 882	- 4.92	2005	14 600	0.127	13 540	- 7.26

表 2 流量序列自相关系数及其上、下限值(置信水平 70%)

Table 2 Upper and lower limits of auto-correlation coefficient of runoff series (confidence level 70%)							
<i>k</i>	<i>R</i> _{1<i>k</i>}	<i>R</i> _{<i>k</i>}	<i>R</i> _{2<i>k</i>}	<i>k</i>	<i>R</i> _{1<i>k</i>}	<i>R</i> _{<i>k</i>}	<i>R</i> _{2<i>k</i>}
1	- 0.169 4	0.166 6	0.127 7	7	- 0.182 4	- 0.037 8	0.134 7
2	- 0.171 4	0.018 4	0.128 8	8	- 0.184 8	- 0.099 5	0.136 0
3	- 0.173 4	- 0.008 2	0.129 9	9	- 0.187 4	0.067 2	0.137 4
4	- 0.175 5	- 0.297 2	0.131 1	10	- 0.190 0	0.095 3	0.138 7
5	- 0.177 7	- 0.314 0	0.132 3	11	- 0.192 8	- 0.094 5	0.140 2
6	- 0.180 0	0.000 4	0.133 5	12	- 0.195 7	- 0.156 0	0.141 6

表 3 参数 *a*, *c* 值

Table 3 Values of parameters <i>a</i> and <i>c</i>									
<i>a</i> ₁	<i>a</i> ₂	<i>a</i> ₃	<i>c</i> ₁	<i>c</i> ₂	<i>c</i> ₃	<i>c</i> ₄	<i>c</i> ₅	<i>c</i> ₆	<i>c</i> ₈
0.551 1	0.497 6	- 0.669 8	95 889	- 158 825	291 773	- 355 333	318 726	- 316 001	- 203 966

径流量进行预测,其相对误差及本文投影寻踪自回归流量预测模型相对误差见表 4。由表 4 可以看到,投影寻踪自回归流量预测模型相对误差(除 1993 年、1995 年和 2000 年外)明显小于其他模型。因此,投影寻踪自回归流量预测模型仅利用流量时序延迟 1 步、4 步和 5 步的相依信息得到这样的结果令人满意。把参数代入该模型,得到 2006 年和 2007 年流量预测值分别为 13 703 m³/s 和 14 372 m³/s。

3 结 语

实例的计算结果说明,预测流量时序这类非线性系统,投影寻踪自回归流量预测模型是可行而有

表 4 宜昌站 1991~2000 年年径流量预测相对误差对比

Table 4 Comparison of relative error of annual runoff at Yichang station from 1991 to 2000				%
年份	神经网络模型	小波网络模型	投影寻踪自回归流量预测模型	
1991	1.5	4.0	0.28	
1992	11.0	7.8	4.65	
1993	- 4.9	- 6.1	- 7.35	
1994	32.0	13.5	8.93	
1995	- 0.6	- 1.3	7.82	
1996	10.8	9.7	3.03	
1997	22.2	19.6	18.47	
1998	- 19.2	- 9.1	- 1.72	
1999	13.2	- 0.4	- 9.96	
2000	- 1.1	- 5.0	- 8.21	

效的,该模型计算简便、适用性强,克服了传统投影寻踪方法计算复杂、编程实现困难的缺点,有利于投影寻踪方法的推广应用,为处理非线性时序预测问题提供了一条值得探索的新途径。

参考文献:

- [1] 金菊良,魏一鸣.投影寻踪门限自回归模型在海洋冰情预测中的应用[J].海洋预报,2002,19(4): 60-66.
- [2] 付强,梁川.节水灌溉系统建模与优化技术[M].成都:四川科技大学出版社,2006: 127-133.
- [3] FRIEDMAN J H, TURKEY J W. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Trans On Computer, 1974, 23(9): 881-890.
- [4] 李祚泳.投影寻踪技术及其应用进展[J].大自然探索,1998,17(63): 47-50.
- [5] 金菊良,杨晓华,丁晶.基于实数编码的加速遗传算法[J].四川大学学报:工程科学版,2000,32(4): 20-24.
- [6] 丁晶,邓育仁.随机水文学[M].成都:成都科技大学出版社,1988: 33-26.
- [7] 任若思,王惠文.多元统计数据分析:理论、方法、实例[M].北京:国防工业出版社,1998.
- [8] 杨位钦,顾岚.时间序列分析与动态数据建模[M].北京:北京工业学院出版社,1986: 391-411.
- [9] 李祚泳,邓新民.大气污染物浓度的投影寻踪自回归预测[J].上海环境科学,1997,16(7): 17-19.
- [10] 付强,赵小勇.投影寻踪模型原理及其应用[M].北京:科学出版社,2006.
- [11] 黄胜.长江上游干流区径流变化规律及径流预测研究[D].四川:四川大学,2006.

Application of projection pursuit auto regression model in predicting runoff of Yangtze River

YU Guo-rong¹, YE Hui², XIA Zi-qiang³, ZHAO Xiao-yong⁴

(1. Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming 650051, China;

2. Pulandian Water Resources Investigation and Design Institute of Dalian City, Pulandian 116200, China;

3. State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China;

4. Yantai Water Conservancy Bureau, Yantai 264000, China)

Abstract: In order to accurately predict the runoff of the Yangtze River, a new model—the projection pursuit auto regression mode is introduced. The prediction factors were determined using of the auto-correlation analysis technique. A new function of projection indices was established and optimized using a real coded accelerating genetic algorithm. A projection pursuit auto regression model was established and applied in the predication of the runoff of the Yangtze River. The calculated results show that the time series of the predicted runoff of the Yangtze River by means of the present model is feasible and effective.

Key words: projection pursuit auto regression model; runoff predication; nonlinear time series; genetic algorithm; the Yangtze River