

DOI :10.3876/j.issn.1000-1980.2011.04.003

基于偏最小二乘回归的洪水预报方法

丁胜祥¹,董增川²,张莉¹,李庆航³

(1.长江水利委员会水文局,湖北 武汉 430010;2.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;
3.长江勘测规划设计研究院,湖北 武汉 430010)

摘要:针对常规多元回归模型无法克服预报因子间复相关性的问题,提出了基于偏最小二乘回归的洪水预报方法.通过结合遗传算法与偏最小二乘回归来进行因子筛选,对筛选后的因子采用 Bootstrap 方法进行检验,再建立基于偏最小二乘回归的预报模型.实例分析结果表明,该方法建立的洪水预报模型结构简单,能依据给定目标进行因子筛选,有助于克服水文要素间复杂的多重相关性,具有较高的预报精度.

关键词:洪水预报;偏最小二乘回归;遗传算法;因子筛选

中图分类号: P338 **文献标志码:** A **文章编号:** 1000-1980(2011)04-0367-05

在水文研究中,水文要素值可以根据其他水文要素进行预测,如可以根据上下游的水位或流量建立相关关系来进行预测,常用的方法是建立统计模型,采用回归分析方法对事先拟定的因子进行筛选和系数求解^[1-2].这种方法简单易行且具有一定的精度,所以有较广泛的应用.然而,水文过程中各个水文要素,如流量、水位等是复相关的,因子间的复相关性会导致回归分析的正则方程组严重病态,从而使最小二乘法的参数估计不稳定,回归效果也因此大大降低^[2-6],使得传统的统计模型在某些实际应用时,效果不甚理想或失效,极大地限制了方法的应用. Wold 等^[7]提出了一种新型的多元统计分析方法——偏最小二乘回归(PLSR)方法.该方法能有效地抽取对系统解释性最强的综合信息,从而实现了对高维数据空间的降维处理,较好地克服变量间的复相关性.徐洪钟等^[3]应用 PLSR 建立了坝顶水平位移的统计模型;杨杰等^[4-5]针对最小二乘法难以克服多重共线性对回归模型精度影响的不足,将 PLSR 引入大坝安全监测变量及其影响因子分析中;邓念武等^[6]利用 PLSR 对大坝位移监测资料进行建模,并应用于土石坝的沉降资料分析;陈南祥等^[2]将 PLSR 与神经网络耦合,建立了泉流量预报模型;蒲前超等^[8]应用 PLSR 建立了河流病因诊断模型;李世欣等^[9]将 PLSR 与神经网络相耦合,建立了小流域降雨侵蚀产沙模型.

水文过程影响因素众多,单从物理成因上无法确定预报因子.一般可以判断待预报的洪水要素与其他某些要素相关,但是无法对相关程度给出判断,只能依靠试错或经验来选取预报因子.而实际预报时,不可能将所有可能与预报洪水要素相关的其他要素均包含进预测模型,不仅因为资料难以获得,过多的因子也使模型过于复杂.在模型精度和预测效果相当的情况下,自变量越少越好,这是多元回归分析的一条基本原则.因此,在利用回归模型进行洪水预报时,应对因子进行适当分析,删除与预报要素相关性较弱的因子,以达到简化模型与增强模型适用性的目的.遗传算法(GA)具有相当强的搜索能力^[10-14],当它与 PLSR 及人工神经网络(ANN)等建模方法相结合后,在一定条件下它们能够在有限的时间内搜寻到变量空间的最佳模型^[10],达到筛选预报因子的目的.

本文尝试用 PLSR 模型进行水文预报,利用 GA 能进行全局搜索的特点,将 GA 与 PLSR 模型结合起来进行因子筛选,根据筛选后的因子建立回归方程并进行预报.

1 基于 PLSR 的洪水预报模型

1.1 PLSR 算法

记因变量为 $Y = (y_1, y_2, \dots, y_q)$, 自变量为 $X = (x_1, x_2, \dots, x_p)$, 设因变量 Y 与自变量 X 均有 n 个观测

收稿日期:2010-08-10

基金项目:教育部科学技术研究重点项目(104197);水利部公益性行业科研专项(200901001, 200901002)

作者简介:丁胜祥(1982—),男,安徽潜山人,工程师,博士,主要从事水文水资源研究. E-mail: dingsx82@126.com

值,由此构成了数据表 $X_{n \times p}$ 和 $Y_{n \times q}$. 其中: X 为自变量矩阵; Y 为因变量矩阵; n 为样本个数; p 为自变量个数; q 为因变量个数.

PLSR 通过成分提取的方法来达到有效建模的目的. 该方法分别在自变量矩阵 X 和因变量矩阵 Y 中提取成分 t 和 u . 提取后的成分 t 是 $x_1, x_2, \dots, x_p$ 的线性组合, 而 u 则是 $y_1, y_2, \dots, y_q$ 的线性组合. 为了回归分析的需要, PLSR 在提取成分时加上了以下 2 个目标^[1] (a) t 和 u 尽可能多地携带各自数据表中的变异信息; (b) t 和 u 的相关程度能够达到最大.

满足目标(a)则 t 和 u 可以尽可能好地代表数据表 X 和 Y , 达到目标(b)则自变量的成分 t 对因变量的成分 u 有很强的解释能力.

在对 X 和 Y 进行一次成分提取后, 分别实施 X 对 t 的回归和 Y 对 t 的回归. 如果回归方程已经达到满意的精度, 则停止计算, 否则, 将 X 关于 t 回归后的残差矩阵与 Y 关于 t 回归后的残差矩阵进行新一轮的成分提取. 如此反复, 直至达到一个较满意的精度为止. 以上为 PLSR 算法的思路, 目前已有许多 PLSR 计算方法, 主要为非线性迭代偏最小二乘(NIAPLS)算法和简单偏最小二乘(SIMPLS)算法 2 大类, 本文采用 NIAPLS 算法进行 PLSR 计算.

1.2 基于遗传算法的预报因子选择

假设一简单洪水预报问题如下(图 1): 要预报 A 区的水位, 可能因子有 B 区的降雨 P 和水位 Z , C 区的水位 Z , D 闸、E 闸和 F 闸的流量 Q , 还有 A 区自身的降雨 P 和水位 Z , 一共 8 种. 某一时刻 A 区的水位应与之前某几个时刻的这 8 种因子值都相关, 倘若将 8 种因子同时向前延长 5 个时段, 则共有 40 个潜在因子. 在这 40 个因子之中如何筛选, 最简单也是最繁琐的方法是枚举法, 即建立因变量和预报因子全部子集组合的回归模型, 然后从中选择最好的模型. 然而, 对于含有 K 个预报因子的问题, 有 $2^K - 1$ 个组合方式, 即会产生 $2^K - 1$ 个回归模型, K 很大时采用枚举法显然是不科学的.

因子筛选实际上是在众多的因子之中寻找最优的因子组合来使得预报模型同时具备较高的精度和足够简单的结构. 从这个角度来说, 因子筛选是一个优化问题, GA 具备良好的全局搜索能力和信息处理的隐含并行性, 能很好地处理该类问题, 尤其当预报量的影响因子较多较复杂时, GA 不仅能较快收敛到最优组合, 而且常常能提供 1 组因子组合而不是 1 个. 因此, 本文采用 GA 对 PLSR 模型的因子进行筛选, 将 GA 的全局优化搜索优点与 PLSR 模型有效解决因子间多重线性相关性的能力结合起来, 以达到因子筛选与简化模型的目的.

根据 PLSR 模型计算效果的好坏选择最好的因子组合是优化的目标. 将各个因子按照顺序排列在一起, 如图 2 第 1 行所示(共 G 个因子), 比较简单的选择组合的方法是采用二进制编码. 根据二进制编码的遗传算法进行进化计算. 这种编码方法对 G 个因子各用 1 位, 最终的编码长度为 G 位. 编码时, 当第 i 个因子对应的第 i 位编码为 1 时, 则该因子起作用; 反之, 当第 i 个因子对应的第 i 位编码为 0 时, 则该因子不起作用, 即被剔除. 但是, 这种编码方式在因子众多时存在编码长度过长的问题, 编码长度的过长意味着搜索空间的巨大, 而搜索空间的过大会导致搜索计算的过于复杂并可能导致算法失效.

其实, 在水文预报的因子筛选过程中, 同一类因子对预报量的影响往往是连续的, 如连续的 3 个时段雨量影响预报水位, 而预报时间前 1 个时段与前面第 3 时段雨量影响预报水位, 但是预报时刻前面第 2 时段雨量不影响预报水位的情况则一般不会出现, 在物理成因上也无法解释. 因此, 针对水文预报因子筛选的实际问题, 可以对编码进行改进, 对每类因子采用 2 个整数来表示其起作用的因子时间范围, 如对于第 1 类因子 A 区面雨量, 编码 [2 5] 表示前 2 3 4 5 这 4 个时段的降雨量对预报水位均产生显著影响. 这样, 对于每类因子, 只用 2 个整数即可表达哪些因子产生作用. 本文采用如下所示的编码方式与适应度函数设计遗传算法并进行因子筛选:

a. 编码. 编码采用整数编码, 每类因子对应 2 个整数, 当有 g 类因子时, 则有 g 对整数, 即一个个体的编

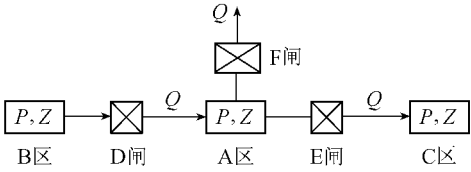


图 1 预报点 A 区情况

Fig. 1 Sketch of area A

1	2	3	4	5	6	...	...	G
1	0	1	0	1	1	...	...	1

图 2 因子排列与二进制编码

Fig. 2 Factor arrangement and binary coding

码形式为 $\{[a_1, b_1][a_2, b_2] \dots [a_g, b_g]\}$,见图 3。

b. 适应度函数.采用预测残差平方和与拟合的确定性系数综合作为适应度函数,具体如下:

$$F = \frac{1}{1000S_{SSE}} + R^2 \tag{1}$$

式中: S_{SSE} ——预报的残差平方和; R^2 ——确定性系数或测定系数。

在编码和适应度函数确定后,结合 PLSR 模型可以对具体的预报对象进行 GA 计算并实现因子筛选,具体步骤为 (a)根据因子类数 g 和最大可能影响时段数 g_b 初始化群体 (b)对每个个体的编码进行解码得到一组因子组合,进行 PLSR 计算,得到 S_{SSE} 和 R^2 ,计算适应度值 F ,对每个个体按照适应度值的大小进行排序 (c)进行选择操作,得到选择后的群体 (d)对选择操作产生的子代群体进行交叉操作,并对交叉后的个体进行修正,得到交叉后子代 (e)对交叉后的子代进行变异计算,并修正产生的不合理个体,得到变异子代; (f)检查是否收敛,是则输出收敛个体编码,否则转向步骤 (b)。

1.3 Bootstrap 检验

遗传算法在因子较多时能够较快找出最佳因子组合,但对各个因子是否具备显著性却不能判断。为检验 PLSR 模型中的各个因子是否具有显著性,常需要采用 Bootstrap 方法对 GA 方法筛选后的因子所组成的 PLSR 模型进行非参数检验,以确定每个自变量是否确实对因变量有显著的解释作用,从而删除对系统无显著解释作用的变量,给出一个更加合理、有效的自变量集合。因此,结合基于遗传算法的因子筛选方法与 Bootstrap 抽样检验所得到的预报因子可以作为最终的预报因子。具体步骤为 (a)根据预报变量,分析确定潜在的预报因子,记因子类数为 g (b)分析预报因子对预报量的最大可能影响步长,即该因子最长到前几个时段的价值会对预报量产生影响,记最大可能影响时间步长为 g_b (c)对 g 类因子,每个因子前推 g_b 步长,采用 GA 进行因子筛选 (d)对遗传算法筛选后的因子采用 Bootstrap 方法进行因子检验,通过检验的因子即为最终的预报因子。

1.4 基于 PLSR 的洪水预报模型

首先分析潜在的相关预报因子,然后采用遗传算法并根据 PLSR 模型的计算效果对预报因子进行筛选,再对筛选后的因子进行 Bootstrap 抽样检验,将通过检验的因子作为 PLSR 模型的输入进行计算,最终得到的回归方程即为预报方程。这样的模型既能够根据预报结果筛选因子,又能够充分利用筛选后的因子信息对因变量进行预报,消除各个因子间的复相关性。具体步骤为 (a)根据预报变量,分析确定潜在的预报因子; (b)利用基于遗传算法的因子筛选方法对预报因子进行筛选,并进行 Bootstrap 检验 (c)对筛选后的预报因子关于预报要素进行 PLSR 模型计算 (d)根据得到的回归方程进行预报。

2 实例分析

在太湖流域防洪调度中太湖水位(指望亭太、西山、夹浦、小梅口和大浦口 5 站平均水位)最为关键,太湖水位高低决定望虞河、太浦河工程运行方式以及周围地区工程的运行方式,本文采用以上介绍的基于 PLSR 的洪水预报方法来预报每日太湖 8 时水位。根据分析,影响太湖水位的因素有 (a)各分区前期降雨量 (b)琳桥、甘露、枫桥和平望等站水位 (c)前期入湖水量 (d)重要防洪枢纽(望亭立交、太浦闸和常熟枢纽)的前期泄量。

记太湖水位时间序列为 Z_t , t 表示时间;记琳桥、甘露、枫桥与平望的水位为 $z_{i,t}$ ($i = 1, 2, 3, 4$), $z_{i,t}$ 表示 i 控制点第 t 时刻水位(单位为 m), i 的取值从 1 到 4 分别表示琳桥、甘露、枫桥与平望;各分区前期面平均降雨量时间序列记为 $P_{i,t}$ ($i = 1, 2, \dots, 6$), 其中 i 为分区标号, $P_{i,t}$ 表示 i 分区第 $t-1$ 时刻至第 t 时刻雨量(单位为 mm), i 的取值从 1 到 6 分别表示湖西山丘区、武澄锡虞区、阳澄淀泖区、太湖湖区、杭嘉湖区和浙西区;记湖西山丘区与浙西山区入太湖流量为 $q_{j,t}$ ($j = 1, 2$, $j = 1$ 表示湖西山丘区入流; $j = 2$ 表示浙西山区入流), $q_{j,t}$ 表示 j 分区第 $t-1$ 时刻至第 t 时刻入太湖流量(单位为 m^3/s);记几个防洪控制枢纽的前期泄量为 $Q_{k,t}$ (单位为 m^3/s ; $k = 1, 2, 3$, $k = 1$ 表示望亭立交泄量; $k = 2$ 表示太浦闸泄量; $k = 3$ 表示常熟枢纽泄量), $Q_{k,t}$ 表示 k 分区第 $t-1$ 时刻至第 t 时刻的泄量(单位为 m^3/s)。经过对实测资料的详细分析得知,各变量中对太湖水位的影响时间最长的为降雨,影响时间为 7 d,因此潜在的预报因子有 105 个。本文以日为时段长,预测下一日 8 时的太湖水位。由于各种水文因素之间的相互作用非常复杂,为了简化问题的复杂性、消除太湖水位

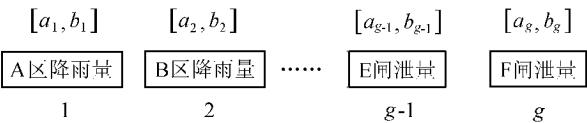


图 3 因子组合的整数编码

Fig. 3 Integer coding of factor combination

的非平稳性,对太湖水位进行一阶差分.设太湖水位时间序列为 Z_t ,表示第 t 日8时水位(单位为m),对太湖水位作一阶差分记为 ΔZ_t ,即 $\Delta Z_t = Z_t - Z_{t-1}$,然后再对 ΔZ_t 进行预报.以1999年4月1日—9月30日的资料对基于PLSR的洪水预报模型进行因子筛选与参数率定,从105个因子中筛选出14个因子,经过Bootstrap抽样检验剔除2个因子,剩下12个因子,得到最终的预报公式为

$$\begin{aligned} Z_t = & -0.00177Z_{t-5} - 0.00158Z_{t-4} - 0.00118Z_{t-3} + 1.00109Z_{t-1} + \\ & 0.0000142q_{1,t-5} + 0.0000153q_{2,t-1} - 0.0000299Q_{2,t-1} - \\ & 0.00181Z_{1,t-3} - 0.00225Z_{1,t-2} - 0.000962Z_{4,t-5} + \\ & 0.00141P_{6,t-1} + 0.3203\Delta Z_{t-1} + 0.0177 \end{aligned} \tag{2}$$

预报模型的率定结果见图4.在率定阶段,水位误差小于或等于3cm的天数为总天数的92.2%,确定性系数为0.99.用2000年4月1日—9月30日的水文观测资料进行模型检验,模型在检验期水位误差小于或等于3cm的天数占总天数的95%,确定性系数为0.98,模型在检验阶段的模拟效果见图5.

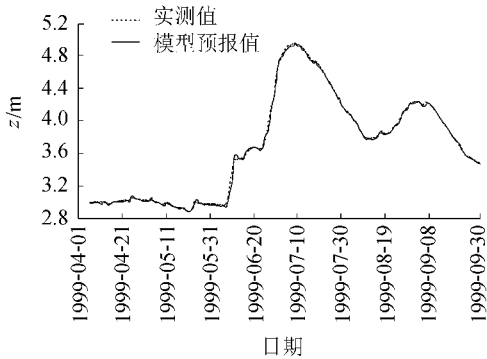


图4 太湖8时水位的模型率定结果

Fig. 4 Measured versus predicted water levels of Taihu lake at 8:00 in model calibration stage

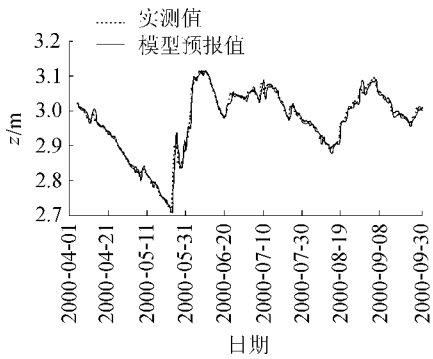


图5 太湖8时水位的模型检验结果

Fig. 5 Measured versus predicted water levels of Taihu lake at 8:00 in model verification stage

从图5可以看出,基于遗传算法的因子筛选方法效果良好,筛选后的因子都是对预报变量解释作用较强的因子,PLSR方法能够克服水文要素之间的复相关性,建立的预报模型简单有效,进行预报时方便可靠且具有很高的预报精度.

3 结 论

本文提出的基于PLSR的洪水预报方法将预报因子筛选与PLSR相结合,对所需预报的洪水要素进行预报,得到了理想的结果.这种基于遗传算法的预报因子筛选与PLSR方法建立的PLSR模型,具有2个方面的优点:(a)由于模型将基于遗传算法的因子筛选与Bootstrap抽样检验结合起来,因此最终模型的预报因子对预报变量均有较好的解释与预报作用;(b)由于采用了PLSR模型,因此模型能有效地抽取对系统解释性最强的综合信息,从而实现对高维数据空间的降维处理,较好地克服水文要素之间的多重相关性.

本文对预报因子筛选的部分还仅仅是根据预报效果来确定,虽然进行了Bootstrap检验,但是对因子筛选的物理意义没有进行分析研究,如何选择能同时考虑预报效果、因子规模和物理意义的更为合理的适应度函数来进行因子筛选还有待进一步研究.

参考文献:

[1] 王惠文,吴载斌,孟洁.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,2006:97-127.
[2] 陈南祥,黄强,曹连海.基于偏最小二乘回归与神经网络耦合的岩溶泉预报模型[J].水利学报,2004,35(9):68-72.(CHEN Nan-xiang,HANG Qiang,CAO Lian-hai. Model for prediction of karst spring flow based on the coupling of neural network model with partial least square method[J]. Journal of Hydraulic Engineering, 2004, 35(9): 68-72.(in Chinese))
[3] 徐洪钟,吴中如.偏最小二乘回归在大坝安全监控中的应用[J].大坝观测与土工测试,2001,25(6):22-23.(XU Hong-zhong,WU Zhong-ru. Partial least square regression and its application to dam safety monitoring[J]. Dam Observation and Geotechnical Tests, 2001, 25(6): 22-23.(in Chinese))

- [4] 杨杰,胡德秀,吴中如.大坝安全监控模型因子相关性及其不确定性研究[J].水利学报,2004,35(12):99-105.(YANG Jie, HU De-xiu, WU Zhong-ru. Multiple co-linearity and uncertainty of factors in dam safety monitoring model[J]. Journal of Hydraulic Engineering, 2004, 35(12):99-105.(in Chinese))
- [5] 杨杰,吴中如.观测数据拟合分析中的多重共线性问题[J].四川大学学报:工程科学版,2005,37(5):19-24.(YANG Jie, WU Zhong-ru. Research on the multicollinearity existing in observation data simulation and analysis[J]. Journal of Sichuan University: Engineering Science Edition, 2005, 37(5):19-24.(in Chinese))
- [6] 邓念武,徐晖.单因变量的偏最小二乘回归模型及其应用[J].武汉大学学报:工学版,2001,34(2):14-16.(DENG Nian-wu, XU Hui. Model of simple partial least squares regression and its application[J]. Engineering Journal of Wuhan University, 2001, 34(2):14-16.(in Chinese))
- [7] WOLD S, MARTENS H, WOLD H. The multivariate calibration problem in chemistry solved by the PLS method[C]//RUHE A, KÅGSTRÖM B. Proc Conf Matrix Pencils, Lecture Notes in Mathematics. Heidelberg: Springer-Verlag, 1983: 286-293.
- [8] 蒲前超,夏继红,韩玉玲,等.基于偏最小二乘回归法的河流病因诊断[J].河海大学学报:自然科学版,2009,37(1):27-31.(PU Qian-chao, XIA Ji-hong, HAN Yu-ling, et al. Reason diagnosis of river health based on partial least squares regression[J]. Journal of Hohai University: Natural Sciences, 2009, 37(1):27-31.(in Chinese))
- [9] 李世欣,温建,邵孝侯,等.偏最小二乘法与人工神经网络耦合的小流域产沙模型[J].河海大学学报:自然科学版,2010,38(2):149-153.(LI Shi-xin, WEN Jian, SHAO Xiao-hou, et al. Sediment yield model for small watersheds based on coupling of partial least square regression and artificial neural network[J]. Journal of Hohai University: Natural Sciences, 2010, 38(2):149-153.(in Chinese))
- [10] HOPFINGER A J, DAVID R. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships[J]. Journal of Chemical Information and Computer Sciences, 1994, 34(4):854-866.
- [11] 刘卫林,董增川,王德智.混合智能算法及其在供水水库群优化调度中的应用[J].水利学报,2007,38(12):1437-1443.(LIU Wei-lin, DONG Zeng-chuan, WANG DE-zhi. Hybrid intelligent algorithm and its application in dispatch optimization for water supply reservoir group[J]. Journal of Hydraulic Engineering, 2007, 38(12):1437-1443.(in Chinese))
- [12] 陆桂华,郦建强,杨晓华.水文模型参数优选遗传算法的应用[J].水利学报,2004,35(2):50-56.(LU Gui-hua, LI Jian-qiang, YANG Xiao-hua. Application of genetic algorithms to parameter optimization of hydrology model[J]. Journal of Hydraulic Engineering, 2004, 35(2):50-56.(in Chinese))
- [13] 陈立华,梅亚东,董雅洁,等.改进遗传算法及其在水库群优化调度中的应用[J].水利学报,2008,39(5):550-556.(CHEN Li-hua, MEI Ya-dong, DONG Ya-jie, et al. Improved genetic algorithm and its application in optimal dispatch of cascade reservoirs[J]. Journal of Hydraulic Engineering, 2008, 39(5):550-556.(in Chinese))
- [14] 刘卫林,董增川,陈南祥,等.基于多变量相空间重构的地下水动态神经网络预测模型[J].河海大学学报:自然科学版,2008,36(1):15-18.(LIU Wei-lin, DONG Zeng-chuan, CHEN Nan-xiang, et al. Neural network model for groundwater prediction based on multivariate phase space reconstruction[J]. Journal of Hohai University: Natural Sciences, 2008, 36(1):15-18.(in Chinese))

A method of flood forecasting based on partial least squares regression

DING Sheng-xiang¹, DONG Zeng-chuan², ZHANG Li¹, LI Qing-hang³

(1. Bureau of Hydrology, Changjiang Water Resources Commission, Wuhan 430010, China;

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

3. Changjiang Institute of Survey, Planning, Design and Research,
Changjiang Water Resources Commission, Wuhan 430010, China)

Abstract: In order to solve the problem of multiple correlation among forecast factors which cannot be avoided in the conventional multiple regression model, a flood forecasting method based on the partial least squares regression was developed. This method combines partial least squares regression with a genetic algorithm to select variables. The selected variables were tested by the Bootstrap method. Case study results show that this method can effectively select variables by given objects and avoid multiple correlation among hydrological elements. The method has the advantages of simple structure and high accuracy.

Key words: flood forecasting; partial least squares regression; genetic algorithm; factor selection