文章编号: 1007-2284(2011)06-0004-03

抗差递推校正模型的风险分析

赵 超[1,2]

(1. 厦门理工学院水资源环境研究所,福建 厦门 361005;
2. 近海海洋环境科学国家重点实验室 厦门大学环科中心,福建 厦门 361005)

摘 要: 抗差递推校正模型在抵御流量异常值对实时校正精度影响方面效果明显,但同时也存在将真实值作为异常值剔除,校正误差更大的风险。利用蒙特卡洛方法,分析不同状况下,抗差递推校正模型风险与效果的关系。结果表明抗差递推校正模型的风险受异常值发生频率影响较大。

关键词: 抗差估计; 风险分析; 蒙特卡洛方法; 频率
中图分类号: TV213.4 文献标识码: A

An Analysis of the Risk of Robust Recursive Updating Model

ZHAO Chao[1,2]

(1. Water Resources and Environmental Institute, Xiamen University of Technology, Xiamen 361005, China; 2. State Key Laboratory of Marine Environmental Science, Environmental Science Research Center, Xiamen University, Xiamen 361005, China)

Abstract: Robust recursive updating model is insensitive to the outliers and is effective to stable flood updating accuracy. At the same time, it is risky to detect falsely good value as outliers. Based on Monte Carlo Method, the relations between risk and effect of model are obtained. The research results indicate that the risk of the model is affected by frequency of outliers.

Key words: robust estimation; risk analysis; Monte Carlo Method; frequency

抗差递推校正模型是一种引入抗差估计理论的自回归实时校正模型[1-3]。模型利用抗差系统具有的抗差能力,推导出带有遗忘因子的抗差递推最小二乘算法,以抵御异常值对模型参数估值的影响,提高实时校正精度。抗差递推校正模型在选取的多个流域效果良好[4],但方法同时存在一定的风险。必须对风险进行定量分析,方可客观、全面地评价抗差递推校正模型的效果。

本文利用蒙特卡洛方法,人工生成多种分布、多种频率的异常误差,定量分析抗差递推校正模型的风险,以正确评价方法的可靠性。

1 抗差递推校正模型的简介

传统实时校正模型,多利用实测流量采用 AR 模型以及递推最小二乘估计(RLS),进行实时修正。当实测流量不可避免地存在一些未知分布的异常误差时,采用递推最小二乘估计校正模型参数,估值有偏,致使校正结果不可信。抗差递推校正模型采用的抗差递推最小二乘算法(RRLS),将抗差估计与最小二乘算法结合,参数估值可以抵御异常值的影响,提高实时校正精度。

通过将抗差估计理论与最小二乘算法相结合,推导出抗差递推最小二乘算法[1,2,5],结果如下:

$$\theta(t+1)=\theta(t)+w(t+1)P_tX_{t+1}[\lambda+w(t+1)X_{t+1}^TP_tX_{t+1}]^{-1}[y(t+1)-X_{t+1}^T\theta(t)] \quad (1)$$

$$P_{t+1}=\frac{1}{\lambda}\{I-w(t+1)P_tX_{t+1}[\lambda+w(t+1)X_{t+1}^TP_tX_{t+1}]^{-1}X_{t+1}^T\}P_t \quad (2)$$

式中:$\theta(t+1)$ 为 $t+1$ 时刻 AR 模型对应的参数;$y(t+1)$ 为利用 AR 模型预测的 $t+1$ 时刻流量与实测流量间的偏差,当实测流量存在异常值时,就转化为 y 存在异常值;$X_{t+1}=[y(t),y(t-1),\cdots,y(t-n+1)]^T$;$w(t+1)$ 为 $t+1$ 时刻对应的权函数;λ 为遗忘因子。

选取的权函数为三段式函数:

收稿日期: 2010-12-28
基金项目: 国家自然科学基金项目(50909084),福建省自然科学基金项目(2009J05107);厦门理工学院杰出青年科研人才培育计划资助。
作者简介: 赵 超(1977-),女,副研究员,博士,主要从事水文水资源方面的研究工作。E-mail: zhaochao@xmut.edu.cn。

$$w(t)=\begin{cases}w(t) & |\varepsilon(t)|\leqslant 1.5\phi\\ w(t)\ 1.5\phi/|\varepsilon(t)| & 1.5\phi<|\varepsilon(t)|\leqslant 2.5\phi\\ 0 & |\varepsilon(t)|>2.5\phi\end{cases}\quad(3)$$

式中:ε 为残差;ϕ 为加权残差均方差。

当残差绝对值较小时($\leqslant 1.5\phi$),权重保持不变,表示 y 可信;当残差绝对值较大时($>2.5\phi$),权重为 0,表示 y 异常干扰严重,将其淘汰;当残差绝对值位于两者之间时,采取降权处理,减小其对参数估计的影响。

抗差递推最小二乘算法之所以可以抵御异常干扰对参数的影响,关键在于式(3)三段式权函数的探测功能。它将所有的实测资料划分进 3 个区间:可信区间、怀疑区间、淘汰区间。位于淘汰区间的实测资料不参与参数估计,即可避免其对参数估值的影响。而三段函数的分界系数取值,将直接关系到 3 个区间的划分。若淘汰区间范围过大,那被划进淘汰区间的实测值的几率将增加,即将可信实测值错误探测为异常值的风险增大。若淘汰区间过小,异常值未被探测出的概率增加,方法的抗差性降低。

根据多个流域的实测资料计算统计,本文取三段分界系数为 1.5 和 2.5,将对此三段式权函数造成的风险做定量分析。

2　蒙特卡洛方法简介

目前,用于风险分析的方法很多,如:重现期法、直接积分法、蒙特卡罗法、均值一次二阶距法、加强一次二阶距法和二阶距法。本文采用目前被广泛应用于科研和生产实际工作中的蒙特卡罗方法。该方法的优点在于无需以显示给出风险因子与研究变量之间的关系,而是通过从实测资料的统计分析中,求得其经验关系。

蒙特卡洛方法(简称 MC 方法)又称统计试验方法,是人工产生和利用随机数方法的总称。它是一类通过对有关的随机变量或随机过程的随机抽样,来求解数学、物理和工程技术问题近似解的数值方法。具体来说,就是对对所要求解的问题,构造一种随机变量或随机过程,使其某一数值特征(如数学期望)为所求问题的解,然后对所构造的随机变量或过程进行抽样,并由得到的样本算出相应的参数值,作为所求问题的近似解。

用 MC 方法模拟产生满足一定分布的异常误差,首先要用某种特定的方法产生该种分布的随机数,这一过程称为随机抽样。随机变量的分布有多种(如:均匀分布、正态分布、P-Ⅲ型,F 分布等),不同分布对应的随机数序列也不同。但就随机数的产生而言,最基本也是最重要的随机变量是在区间[0,1]上服从均匀分布的随机变量。服从其他分布随机变量的随机数可由[0,1]上均匀分布的随机数变换产生。

产生单位均匀分布随机数的方法很多,有物理方法、随机数表法和数学方法。目前使用最广的是在计算机上通过各种数学运算产生随机数序列,特点是速度快,占用计算机内存小,对所模拟的问题可以进行复查。数学方法产生随机数序列到一定的时刻又回到初始值,即序列具有一定的周期。因此,严格地说,这样生成的序列并非真正相互独立、均匀分布的随机变量子样。但是,通过选取适当的参数,使产生出来的序列可以通过各种关于均匀分布与相互独立性的统计检验,因而可以被接受为独立均匀随机数序列使用,也称之为伪随机数。

用数学方法产生伪随机数的方法有多种。本文采用乘加同余[6]方法获得伪随机数。其递推公式为:

$$\begin{cases}r_{i+1}=r\,x_i+c(\mathrm{mod}M)\\ \xi_{i+1}=r_{i+1}/M\qquad i=1,2,\cdots\cdots\end{cases}\quad(4)$$

式中:a 为乘子,c 为常数,M 为模,均为非负整数。

3　抗差递推校正模型的风险分析

由于权函数的作用,抗差递推校正模型的风险和效果都达到最佳是不可能的,只能求的某种意义上的最佳平衡。是冒着损失一部分有用信息的风险求得一定程度的抗差效果。

抗差递推校正模型的风险主要有两大类,其一是模型的探测风险(P_f),即把流量可信值探测为异常值的风险;其二是模型的校正风险,即利用模型参数估值校正获得的校正流量相比未校正的流量误差更大的风险。

采用乘加同余法,人工生成如下异常误差,叠加到闽江七里街流域 1997－1998 年 7 场实测洪水流量资料,以形成含异常值的"实测流量资料"。

异常误差生成模式:

$$\varepsilon_i=\begin{cases}(r-0.5)\ \overline{Q}\ p & i=\mathrm{int}(i/L)\ L\\ 0 & i\neq\mathrm{int}(i/L)\ L\end{cases}\quad(5)$$

式中:r 为随机数;$\overline{Q}$ 为实测流量的均值;p 为常数,可控制异常误差的大小;L 为异常值产生的频率间隔,可控制异常误差产生的频率。

3.1　探测风险

对以下 4 种情况采用蒙特卡洛方法模拟 0.5 万次,计算抗差递推校正模型的探测风险和探测效果:

(1) p、L 均为定值,$p=1$、$L=5$;

(2) p 在[1,3]间均匀分布,$L=5$;

(3) $p=1$,在[5,20]间均匀分布;

(4) p 在[1,3]间均匀分布,在[5,20]间均匀分布。

探测风险与探测效果采用式(6)、(7)计算,计算结果见表 1。

$$P_f=\frac{\text{探测失误的个数}}{\text{资料总数}}\quad(6)$$

$$P_s=\frac{\text{探测准确的个数}}{\text{异常值个数}}\quad(7)$$

表 1　探测风险和探测效率

Tab. 1 Detection risk and detection efficiency

洪　号	P_f				P_s			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
970605	0.009	0.009	0.003	0.002	0.747	0.748	0.758	0.790
970618	0.004	0.004	0.001	0.001	0.826	0.832	0.830	0.858
970701	0.008	0.008	0.004	0.003	0.825	0.829	0.830	0.864
970808	0.002	0.002	0.001	0.001	0.829	0.824	0.835	0.879
980215	0.003	0.003	0.002	0.002	0.820	0.829	0.822	0.865
980301	0.003	0.003	0.002	0.001	0.800	0.806	0.805	0.846
980509	0.004	0.004	0.002	0.001	0.737	0.745	0.760	0.800
平均	0.005	0.005	0.005	0.004	0.798	0.802	0.806	0.843
均方差	0.003	0.003	0.003	0.003	0.039	0.039	0.033	0.034

由表 1 发现,抗差递推校正模型对于不同场次、不同量级的洪水,探测风险和探测效率的均方差都较小,且比较稳定。说明探测风险和探测效率比较稳定。同时模型的探测风险比较小。相比较而言,970 605 和 970 701 两场洪水的探测风险较其他洪水偏大,分析发现,此两场洪水中存在突变时段流量,模型将其探测为异常值,增大了探测风险。

对比情况(1)和情况(2)以及情况(3)和(4),当异常误差发生频率相同,模型对同一场洪水的探测风险基本保持不变,说明探测风险(P_f)与异常误差极值的关系不大;而探测效果(P_s)随误差极值的增大而提高。

对比情况(1)和(3)以及情况(2)和(4),当异常误差极值相同,异常误差发生间隔时间增长,模型的探测风险(P_f)变小,探测效果(P_s)变大。

3.2 校正风险

抗差递推校正模型的校正风险,是根据参数估值,进行实时校正,校正后的流量误差(e_i)大于不校正的流量误差(e)概率,此时的风险和效果之和为 1。校正风险公式如下:

$$P_j = P(e_i > e) \tag{8}$$

上述 4 种情况下的校正风险见表 2。

表 2 校正风险

Tab. 2 Updating risk

洪 号	P_j			
	(1)	(2)	(3)	(4)
970 605	0.121	0.126	0.080	0.080
970 618	0.153	0.150	0.098	0.097
970 701	0.054	0.051	0.039	0.040
970 808	0.126	0.121	0.076	0.076
980 215	0.091	0.094	0.055	0.056
980 301	0.080	0.080	0.062	0.062
980 509	0.125	0.135	0.076	0.079
平均	0.107	0.108	0.069	0.070

由表 2 可以看出,4 种情况抗差递推校正模型的校正风险均小于 11%,说明模型能抵御绝大部分异常数据的干扰,获得更精确的结果。

对比情况(1)和(2)以及情况(3)和(4),对于每一场洪水,当异常值发生频率不变,校正风险与异常误差极值的变化不明显。当误差极值不变,异常值发生时间间隔长,校正风险减小。

对比情况(1)和(3)以及情况(2)和(4),当异常误差极值相同,异常误差发生间隔时间增长,模型的校正风险(P_f)变小。

模型校正风险主要来源于模型的探测风险,所以校正风险与探测风险的结论基本一致。

4 结 语

针对抗差递推校正模型的风险定量分析发现,模型的风险较稳定且较小,同时模型的风险与异常值发生的频率关系较大,与异常误差的极值关系较小。抗差递推校正模型处理具有异常值的流量数据时,抗差效果较好,风险稳定且较小。本文只针对权函数分界系数为 1.5 和 2.5 效果和风险的关系作了讨论,对于其他取值的分析以及在更多流域的应用将有待进一步研究。 □

参考文献:

[1] Zhao Chao, Hong Hua-sheng, Bao Wei-min, et al. Robust recursive estimation of auto-regressive updating model parameters for real-time flood forecasting [J]. Journal of hydrology, 2008, 349: 376-382.

[2] 赵 超, 洪华生, 张珞平. 实时校正模型的抗差递推算法[J]. 中国科学院研究生院学报, 2008, (5): 665-670.

[3] 包为民, 王 浩, 赵 超. AR 模型参数的抗差估计研究[J]. 河海大学学报, 2006, (3): 258-261.

[4] 赵 超. 流域实时洪水抗差预报系统研究[D]. 南京: 河海大学, 2006.

[5] 周江文, 黄幼才, 杨元喜, 等. 抗差最小二乘法[M]. 武汉: 华中理工大学出版社, 1997: 115-116.

[6] 徐钟济. 蒙特卡罗方法[M]. 上海: 上海科学技术出版社, 1985.

(上接第 3 页) 区径流能力的减弱。

4 结 语

南流江、钦江 1970-2008 年月均实测径流与降雨量年内分配呈单峰型,径流、降雨量年内分配不均系数较小,但呈上升趋势。由于受水资源开发利用等人类活动的影响,实测径流多年变化和不均匀程度大于降雨的,径流量的年际变化均匀度较差。在降雨旱涝等级变化波动较小的情况下,实测径流丰枯等级变化波动频繁,丰枯历时较短。

南流江降雨呈上升趋势,而实测径流呈下降趋势;钦江降雨、径流都呈下降趋势,但径流下降幅度大于降雨的。钦江实测径流系数下降幅度大于南流江,且呈显著下降趋势。南流江、钦江降雨径流关系分为 3 个阶段,降雨径流关系呈现减小的规律。北部湾经济区开放开发后经济迅速发展,流域下垫面变化,水资源开发利用程度提高,耗水量增加,致使流域径流能力降低。 □

参考文献:

[1] 高卫平, 秦 毅, 黄 强, 等. 唐乃亥流域近期降雨径流特性变化分析[J]. 西安理工大学学报, 2005, 21(4): 429-432.

[2] 石教智, 陈晓宏, 吴 甜, 等. 东江流域降雨径流变化趋势及其原因分析[J]. 水电能源科学, 2005, 23(5): 8-10.

[3] 武夏宁, 江 燕. 潮河流域气候变化对径流量的影响分析[J]. 中国农村水利水电, 2010, (2): 5-7.

[4] 高 伟, 王西琴, 曾 勇. 太湖流域西苕溪 1972-2008 年径流量变化趋势与原因分析[J]. 中国农村水利水电, 2010, (6): 33-37.

[5] 栾兆擎, 胡金明, 邓 伟, 等. 人类活动对挠力河流域径流情势的影响[J]. 资源科学, 2007, 29(2): 46-51.

[6] 李道峰, 田 英, 刘昌明. GIS 支持下的黄河河源区降水径流要素变化分析[J]. 水土保持研究, 2004, 11(1): 144-147.

[7] 胡兴林. 甘肃省主要河流径流时空分布规律及演变趋势分析[J]. 地球科学进展, 2000, 15(5): 516-520.

[8] 曹 琨, 葛朝霞, 薛 梅, 等. 上海城区雨岛效应及其变化趋势分析[J]. 水电能源科学, 2009, 27(5): 31-33.

[9] 刘春蓁, 刘志雨, 谢正辉. 近 50 年海河流域径流的变化趋势研究[J]. 应用气象学报, 2004, 15(4): 387-393.