Article ID:1002-1175(2010)04-0556-07

Brief Report

Parameter estimation for Muskingum routing model based on robust algorithm*

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*)

(Received 30 June 2009; Revised 3 February 2010)

Zhao C. Parameter estimation for Muskingum routing model based on robust algorithm [J]. Journal of the Graduate School of the Chinese Academy of Sciences, 2010,27(4):556-562.

Abstract There are a variety of techniques for estimating the parameters of the Muskingum routing model. However the robustness of these methods has to be questioned because of the tendency of outliers in data to strongly influence the outcome. A robust estimation has been presented. The robustness of this estimator has been compared with the least squares method by means of synthetic data sets, in which both Gaussian random errors and outliers have been introduced. The study demonstrates that the robust estimator has the potential to reduce estimation bias in the presence of outliers, and it has an advantage over the least squares method.

Key words Muskingum model, outlier, parameter estimation, robustness

CLC P338

1 Introduction

The Muskingum model, first developed by McCarthy for flood control studies in the Muskingum River basin in Ohio, representing a linear reservoir concept, is an example of the simplest form of the flood routing models. The Muskingum model uses continuity and storage equations which are stated, respectively, as:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = I - Q, \tag{1}$$

and

$$W = k[xI + (1 - x)Q], \tag{2}$$

where W is the channel storage, I and Q are channel inflow and outflow, respectively, k and x are the Muskingum parameters.

Equations (1) and (2), when expressed in finite difference form and solved for the outflow at time step

* Supported by the National Natural Science Foundation (50909084), the Natural Science Foundation of Fujian Province (2009J05107), and Xiamen University of Technology (YKJ08015R)

† E-mail: zhaochao@xmut.edu.cn

t, yield

$$Q_t = c_0 I_t + c_1 I_{t-1} + c_2 Q_{t-1}, \tag{3}$$

where c_0, c_1, and c_2 are expressed in terms of k, x, and $\Delta t = t_t - t_{t-1}$ (the routing time step) as

$$c_0 = (\Delta t - 2kx)/(2k - 2kx + \Delta t), \tag{4}$$

$$c_1 = (\Delta t + 2kx)/(2k - 2kx + \Delta t), \tag{5}$$

$$c_2 = (2k - 2kx - \Delta t)/(2k - 2kx + \Delta t), \tag{6}$$

that is

$$c_0 + c_1 + c_2 = 1. \tag{7}$$

If both the inflow and outflow hydrographs are available, the Muskingum routing model parameters, (k, x) or (c_0, c_1, c_2), can be determined by several estimation techniques. These techniques include trial-and-error graphical procedure and linear and nonlinear regression schemes based on the least-squares method [1-4], approximate methods [5], and genetic algorithms [6-7].

The trial-and-error procedure was the first method developed by McCarthy (1938), whereas it is somewhat subjective and time consuming. For the most part, it has become obsolete. To avoid subjective interpretations of data in estimating parameters, numerical techniques have been developed. However, in the course of these studies, disturbances of inflows and outflows are generally assumed stochastic and Gaussian. When the disturbances are not Gaussian, but corrupted by outliers, above-mentioned methods are significantly affected and induce a bias of the estimated parameters. Even an outlier can result in a strong degradation. This brings us to the issue of robustness. An estimator that is insensitive to departures from the assumptions is considered robust. While the problem of robustness is well recognized in statistics [8-9], few approaches have been proposed for applications in hydrology.

This paper develops a robust procedure for estimating parameters of Muskingum routing model by mitigating the influence of the outliers. A brief description of robust estimation is given in section 2, comparison of the results obtained from the robust procedure with those obtained from a least-squares scheme with synthetic data sets is reviewed in section 3, and the conclusions are summarized in section 4.

2 Robust estimation

Equation (3) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{XC}, \tag{8}$$

where $\boldsymbol{X} = \begin{pmatrix} I_2 & I_1 & Q_1 \\ I_3 & I_2 & Q_2 \\ \cdots & \cdots & \cdots \\ I_m & I_{m-1} & Q_{m-1} \end{pmatrix}$, $\boldsymbol{Y} = (Q_2, Q_3, \cdots, Q_m)^{\mathrm{T}}$, $\boldsymbol{C} = (c_0, c_1, c_2)^{\mathrm{T}}$ and $c_0 + c_1 + c_2 = 1$, m is the data size.

In the least-squares algorithm, the parameters are determined by minimizing the sum of squared residuals

$$\mathrm{Min}\, J = \sum_{i=2}^{m} (Q_i - \hat{Q}_i)^2 - \lambda\left(\sum_{j=0}^{2} c_j - 1\right) = \sum_{i=2}^{m} \varepsilon_i^2 - \lambda\left(\sum_{j=0}^{2} c_j - 1\right), \tag{9}$$

where Q_i is the observed outflows, $\hat{Q}_i$ is the corresponding Muskingum model result, which is a function of the parameter vector $\boldsymbol{C}$. λ is Lagrange multiple. This algorithm is optimal when the residuals, ε_i are Gaussian, otherwise performance of the least squares method degrades.

To the channels running into reservoir, channel outflows are equal to reservoir inflows. In China, reservoir inflows are not measured immediately, but calculated by reservoir water-balance according to the observed reservoir stage and observed reservoir release. In windy days, the measures of reservoir stage often are affected by wind that causes the outlying reservoir inflows, even negative values. In this case, ε_i displays a non-Gaussian distribution. The non-Gaussian nature of ε_i can be modeled as a mixture distribution, such that a large portion of the residuals obey a normal distribution with small variance, while a small portion has an unknown distribution with much bigger variance. The probability density function (pdf) of ε_i is

$$p(\varepsilon) = (1-\delta)N(\varepsilon|0,\sigma^2) + \delta h(\varepsilon), \quad 0 \leqslant \delta < 1, \tag{10}$$

where $N(\cdot|0,\sigma^2)$ denotes the zero mean normal pdf with variance σ^2, and $h(\cdot)$ is a pdf of outliers with variance $\sigma_h^2 \gg \sigma^2$.

It is well known that outliers, generated by the pdf $h(\cdot)$ in equation (10), have an unsuitably large influence on the least squares estimation. Least squares method is evidently influenced by outliers by equivalently taking account of all residuals including outliers. Therefore, robust procedures have been developed to modify the least-squares estimates in order to down-weight the influence of outliers [9].

For reducing the effect of outliers, the robust estimation adopts another function instead of equation (9),

$$\text{Min } J_r = \sum_{i=2}^{m} \rho(\varepsilon_i) - \lambda\left(\sum_{j=0}^{2} c_j - 1\right), \tag{11}$$

where $\rho(\cdot)$ is a robust loss function, which suppresses the outliers. Moreover, regarding the contaminated Gaussian character of the discharge under study, this function has to provide high efficiency for the nominal normal model in equation (10), as well as for $h(\cdot)$ of outlier models. Thus, $\rho(\cdot)$ should look like a quadratic function for small values of the argument. Furthermore, it is desirable that its derivative $\psi(\cdot) = \rho'(\cdot)$, the so-called influence function, should be bounded and continuous. Boundedness ensures that no single observation can have an arbitrarily large influence, while continuity ensures that rounding or quantization errors will not have a major effect.

The desired value $\hat{C}$ can be obtained by solving the equation

$$\partial J_r / \partial c_j = \sum_{i=2}^{m} w(\varepsilon_i) \cdot \varepsilon_i \cdot \frac{\partial \varepsilon_i}{\partial c_j} - \lambda = 0, j = 0,1,2, \tag{12}$$

where $w(\varepsilon_i) = (1/\varepsilon_i) \cdot \psi(\varepsilon_i)$ is a weight function.

Matrix form of equation (12) is

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} & \boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B} & 0 \end{bmatrix} \begin{bmatrix} \hat{C} \\ \lambda \end{bmatrix} + \begin{bmatrix} -\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Y} \\ -1 \end{bmatrix} = 0, \tag{13}$$

where $\boldsymbol{B} = \{1,1,1\}$, $\boldsymbol{W} = \begin{pmatrix} w(2) & 0 & \cdots & 0 \\ 0 & w(i) & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & w(m) \end{pmatrix}$,

$$\begin{bmatrix} \hat{C} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} & \boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Y} \\ 1 \end{bmatrix}. \tag{14}$$

Equation (14) defines the robust estimation.

To robust estimation, the choice of the ρ and w functions is most important. Many functions have been proposed in the literature [8,10]. In this paper, IGG Ⅰ estimator has been selected for this study [11-12].

$$\rho(\varepsilon) = \begin{cases} \varepsilon^2/2 & |\varepsilon| \leqslant a\sigma \\ a \cdot |\varepsilon| & a\sigma < |\varepsilon| \leqslant b\sigma, \\ b \cdot k & |\varepsilon| > b\sigma \end{cases} \tag{15}$$

$$w(\varepsilon) = \begin{cases} 1 & |\varepsilon| \leqslant a\sigma \\ a/|\varepsilon| & a\sigma < |\varepsilon| \leqslant b\sigma. \\ 0 & |\varepsilon| > b\sigma \end{cases} \tag{16}$$

In contrast to robust estimation, $w(\varepsilon)$ is 1 for all ε in least squares method. This means that weights in least squares method are equal for all residuals while weights in robust estimator decrease to zero as residuals become large.

Robust estimator cuts off data with an absolute value of ε greater than $b \cdot \sigma$ and suppresses data with an absolute value of ε between $a \cdot \sigma$ and $b \cdot \sigma$. With reduction of $a \cdot \sigma$ and $b \cdot \sigma$, estimation could be closer to the true parameter sets, however the good value will increase the risk of being treated as outliers. The parameters a and b in equations (15) and (16) are important for the estimation because the parameter value affects the estimation results. The reasonabe values are 1.5 and 2.5 for the constants a and b, respectively[13]. Such a choice will lead to much higher efficiency than the least squares method when the underlying noise pdf has a heavy-tailed non-Gaussian shape in equation (10), while it will still maintain efficiency if the pdf is Gaussian.

Moreover, w is dependent on σ, while σ depends on C determined by w. So w has to be computed by iteration. The iterative solution at the kth step is

$$\begin{bmatrix} \hat{\boldsymbol{C}}^{(k+1)} \\ \hat{\boldsymbol{\lambda}}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(k)}\boldsymbol{X} & \boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(k)}\boldsymbol{Y} \\ 1 \end{bmatrix}. \tag{17}$$

The first iteration starts with $w(\varepsilon_i) = 1$, which is equivalent to the least squares method.

It is well known that iterative algorithm needs adaptive initial values. Note that if ρ function is convex, we have found good results using the result of least squares estimate as the initial estimate. However if ρ function is not convex, result of least squares estimate sets the robust estimator at a wrong starting point leading to a local solution that we are not interested in. In this situation, initial values should be around the optimal parameters, so that global optimum could be found.

Since equation (15) is not convex, initial values around the optimal parameters must be selected. Assuming that the number of the outliers is less than $m/6$, the observed data is divided into $m/6$ groups. There are 6 values in each group, then there is at least one group without outlier. Then six values of every group are used into equation (9), parameters of $m/6$ groups are produced. Parameters of every group and all observed data are used to compute equation (11). Parameters that lead to the least J_r are elected as initial values of parameters. Generally, the initial values are got by the group of data without outliers. It is obviously that the breakdown point of this method is 16.7%[14]. Of course, the number of the data in every group can be adjusted according to practical demands.

3 Performance comparison

In this section the performance of the robust estimator in comparison with the least squares method is shown using synthetically generated data sets. The reason for using synthetic data is that the error structure is known and can be varied to test different hypotheses.

We consider an ideal experiment in which synthetic inflows are free of errors, and the true parameters are $c_0 = 0.28$, $c_1 = 0.52$, and $c_2 = 0.2$. The outflows are calculated by Muskingum routing model with synthetic inflows and true parameters.

First, we examine the performance of the two estimators in the presence of errors that are normally distributed. Gaussian noise is added to outflows to simulate random measurement errors. We generated 2000 realizations for each outflow point in Monte Carlo method.

In Table 1, the results obtained with the two estimators are summarized. The parameter means are near to the true parameter values, confirming that both the least squares and robust estimators are unbiased. The standard deviation is a measure of estimation uncertainty. The uncertainty of robust estimator is close to that of least squares estimator. It is obvious that the performance of two estimators is quite similar if the normal distribution of errors is satisfied.

Table 1 Sample statistics of Muskingum model for the two estimators

	Mean c_0	Mean c_1	Mean c_2	s. d. c_0	s. d. c_1	s. d. c_2
Least squares	0. 2794	0. 5215	0. 1991	0. 0047	0. 0100	0. 0057
Robust estimator	0. 2790	0. 5220	0. 1990	0. 0079	0. 0141	0. 0068

Next, we examine the performance of the two estimators in the case where errors are not normally distributed. In this study, errors of outflows are Gaussian errors augmented with outliers. Outliers are generated by

$$e_i = \begin{cases} (r - 0.5) \cdot \bar{I} \cdot p, & i = \mathrm{int}(i/L)L, \\ 0, & i \neq \mathrm{int}(i/L)L, \end{cases} \tag{18}$$

where r is a random number, $\bar{I}$ is the average of inflows, p is a constant that controls the maximum of e, L is the frequency of outliers. Adjusting p and L, outliers with different magnitudes and sizes can be generated. Two thousand realizations are generated for each outflow point in Monte Carlo method.

The statistical results of the two estimators with different p and L values are displayed in Table 2 and Fig. 1. The results show that least squares method gives significantly biased estimation due to the high sensitivity to the outliers. The deviation is obvious with an increase in p and a decrease in L, demonstrating that the performance of this method is evidently affected by numbers and magnitudes of outliers.

Table 2 Statistics of Muskingum model parameters for the two estimators

		Least squares			Robust estimator		
		c_0	c_1	c_2	c_0	c_1	c_2
true value		0. 28	0. 52	0. 2	0. 28	0. 52	0. 2
$L=15$	$P=0.5$	0. 2254	0. 6662	0. 1084	0. 2786	0. 5226	0. 1988
Contaminated	$P=1$	0. 1899	0. 763	0. 0471	0. 2783	0. 5225	0. 1992
(1/15) = 6. 67%	$P=2$	0. 176	0. 8035	0. 0205	0. 2786	0. 525	0. 1963
$L=8$	$P=0.5$	0. 2141	0. 7077	0. 0782	0. 2782	0. 5249	0. 1969
Contaminated	$P=1$	0. 191	0. 7841	0. 0250	0. 2792	0. 5212	0. 1996
(1/8) = 12. 5%	$P=2$	0. 1951	0. 8022	0. 0027	0. 2801	0. 521	0. 1989

Instead of minimizing the variance of all residuals, robust algorithm preferentially matches the less affected outflows, and assigns zero weights to the outliers. So robust method is not sensitive to outliers and gives more accurate parameter estimation. With an increase in p and a decrease in L, the results plotted in Fig. 1 for the robust estimator are further away from the least squares solution and identically close to the true parameter set. However, it is important to realize that the robust method performs favorably only for a certain proportion of outliers. If the majority of the data is corrupted, the robust estimator is not expected to perform better than least squares method and may in fact bias the solution toward the wrong parameter set by discarding the good data. In this paper, choice of the initial values decides that maximum proportion is 16.7%.

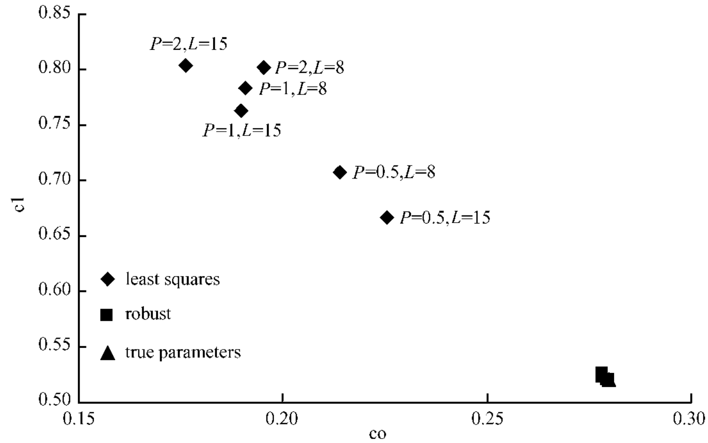

Fig. 1 Mean estimated parameter sets from 2000 hypothetical data sets with outliers using the two estimators

4 Conclusions

We have shown that the robust estimation produces more efficient and less biased estimates than the conventional least squares method. Muskingum model parameter estimation based on the minimization of a robust weight function will therefore give more accurate results. Testing on synthetic data sets indicates that the influence of outliers on the parameter estimation is reduced significantly by using the robust estimation schemes. In summary, the robust estimation algorithm can provide satisfactory results, no matter whether there are outliers or not in inflow and outflow.

Another advantage of robust estimation method is that it provides useful diagnostic information for detecting outliers and influential observations. Of course, one would have to pay some price in efficiency loss by using these robust estimators if the normal model is indeed correct. However, in the situations of contamination and non-normal errors, robust methods provide more accurate results.

References

[1] Aldama A A. Least-squares parameter estimation for Muskingum routing [J]. Journal of hydraulic engineering, 1990, 116(4): 580-586.

[2] Tung Y K. River flood routing by nonlinear Muskingum method [J]. Journal of hydraulic engineering, 1985, 111(12): 1447-1460.

[3] Yoon J, Padmanabhan G. Parameter estimation of linear and nonlinear Muskingum models [J]. Journal of water resources planning and managemen, ASCE, 1993, 119(5): 600-610.

[4] Zhai G J. Research on parameter estimation for Muskingum mode [J]. Hydrology, 1997, 3: 40-43.

[5] Jasem M H, Ismail I E. Approximate methods for the estimation of Muskingum flood routing parameters [J]. Water resources management, 2006, 20: 979-990.

[6] Chen J J, Yang X H. Optimal parameter estimation for Muskingum model based on gray-encoded accelerating genetic algorithm [J]. Communications in nonlinear science and numerical simulation, 2007, 12: 849-858.

[7] Mohan S. Parameter estimation of nonlinear Muskingum models using genetic algorithm [J]. Journal of hydraulic engineering, 1997, 123(2): 137-142.

[8] Andrews D F, Bickel P J, Hampel F R, *et al*. Robust estimates of location: Survey and advances [M]. Princeton: Princeton University press, 1972.

[9] Huber P J. Robust statistics [M]. New York: John Wiley, 1981.

[10] Zhou J W, Huang Y C, Yang Y X, *et al*. Robustified least squares approaches [M]. Wuhai: Huangzhong University of Science and Technology Press, 1995.

[11] Bao W M, Qu S M, Huang X Q, *et al*. Analysis of robust weighting functions for hydrological systems [J]. Journal of Tsinghua University, 2003, 43(8): 1127-1129.

[12] Zhao C, Hong H S, Bao W M, *et al*. Robust recursive estimation of auto-regressive updating model parameters for real-time flood forecasting [J]. Journal of Hydrology, 2008, 349(3):376-382.

[13] Zhou J W. Classical theory of errors and robust estimation [J]. Aota Geodetica et Cartographica Sinica,1989,2: 115-120.

[14] Rousseeuw P J. Unconventional features of positive-breakdown estimators [J]. Statis and Pro Letters, 1994,19: 417-431.

基于抗差算法的马斯京根参数估计

赵 超[1,2]

(1 厦门理工学院水资源环境研究所，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，厦门 361005)

摘　要　用于估计马斯京根模型参数的方法很多，但这些方法在数据存在异常值时缺乏抵御异常值影响的抗差性能. 推导出一种有限制条件的参数抗差估计算法，通过含有随机误差和异常误差的人工数据和真实数据比较抗差算法与传统最小二乘算法的抗差性. 研究表明抗差估计算法能减小异常值对参数估值的影响.

关键词　马斯京根模型，异常值，参数估计，抗差性