文章编号:1671-8860(2010)02-0168-04　　**文献标志码**:A

利用 Bayes 估计进行多波束测深异常数据探测

黄贤源[1]　隋立芬[1]　翟国君[2]　柴洪洲[1,2]

(1　信息工程大学测绘学院,郑州市陇海中路 66 号,450052)
(2　天津海洋测绘研究所,天津市友谊路 40 号,300061)

摘　要:在海底地形变化连续、平缓的假设条件下,基于 Bayes 估计理论提出了多波束测深异常数据探测方法,并与选权迭代加权平均滤波法进行了分析和比较。结果证明,该方法可以解决测深异常值判断标准可靠性的问题,而且能合理、有效地探测出异常值。

关键词:多波束;Bayes 估计理论;异常数据
中图法分类号:P229.2

多波束测深系统是一个全覆盖式声纳测深系统,由于仪器噪声、复杂的海况或多波束声纳参数设置的不合理等因素,使得测深数据中含有少量的异常数据,这些异常数据直接影响到海底的真实反映,需要对其进行滤波处理。由于海洋测量的动态性,不可能通过重复测量来检测异常数据,并提高精度,为此,在海底地形变化连续、平缓的假设下,许多学者在异常检测方面取得了不少的研究成果[1-7],归纳起来主要有中值滤波法、趋势面滤波法以及选权迭代加权平均滤波法,这些方法判定异常值的标准都是通过残差序列与 2σ 或 3σ 进行比较得到的,而且可靠的残差序列是依赖于精确的海底地形,当残差序列不准确时,相应的标准也不可靠,这就可能引起异常值的漏判,这些缺陷使得上述方法在判定异常值的可靠性方面大大降低。基于上述情况,本文受到文献[8,9]思想的启发,对水深观测值直接引入识别变量,基于 Bayes 估计理论提出了一种新的多波束异常值探测方法。

1　海洋测深数据的预处理及局部区域水深观测方程的建立

多波束测深采用的是广角度定向发射以及多通道信息接收技术,因此,通过该系统获得的水深值具有高密度、不规则的特点[10,11]。利用获得的全部海底测深数据对海底地形进行真实的反映固然是一种很好的方法,但是由于海底测深数据在空间上的不规则性使得数据处理的过程变得繁琐甚至不可行,而且目前很多等高线的算法也是基于规则格网的,因此在对测深数据进行取舍的同时进行规则格网化。本文采用的是最小曲率的格网化方法,该方法保证了海底地形连续变化的特点,有利于测深异常值的检测。

多波束测深系统获得的海底地形数据是呈条幅式的,并且由于边缘波束的质量较差,因此只取靠近中央波束的数据进行处理。根据生产和科学实验可知,在取得的水深观测数据中,异常值的出现仅占 1%～10%。采用最小曲率法对数据进行规则格网化,仅仅使其满足海底地形是连续变化的,而不是平缓的。因此,为了更有效地检测出异常值,有必要对规则格网化后的测深数据按行列数变化作进一步细分,细分后的区域应满足海底测深值相对稳定,且仅包含少量的(或没有包含)异常值。

经过规则格网化及细分,得区域内测深点的观测方程为:

$$\underset{n\times 1}{\boldsymbol{L}} = \underset{n\times 1}{\boldsymbol{A}}\ \underset{1\times 1}{\boldsymbol{X}} + \underset{n\times 1}{\boldsymbol{\Delta}} \tag{1}$$

式中,$\boldsymbol{L}$ 为测深点的水深向量;$\boldsymbol{A}=(1,\cdots,1)^{\mathrm{T}}$ 为设计矩阵,且列满秩;$\boldsymbol{X}$ 为未知参数向量;$\boldsymbol{\Delta}=(\Delta_1,\cdots,\Delta_n)^{\mathrm{T}}$ 为观测误差向量;n 为细分区域的节点数。

在区域内,任何一个节点 j 均可作为检测点,区别于选权迭代加权平均滤波法[2,3],检测点的观测值 L_j 也参与平差,其他水深观测值 $L_i(i\neq j,i$

收稿日期:2009-12-15。
项目来源:国家自然科学基金资助项目(40474007)。

$=1,\cdots,n$)相应的权为:

$$P_i = d_i^{-2} / \sum_{i=1}^{n} d_i^{-2} \tag{2}$$

其中,d_i为测深点L_i到检测点L_j的距离。为克服由于多波束数据过于密集($d_i=0$)而导致P_i无法正常求解的缺点,规定检测点的权$P_j=1$,其他水深观测值的权在满足式(2)的条件下,还满足相应的权比关系以及$P_i<P_j(i\neq j)$。

2　识别变量的引入及异常值判断标准的建立

基于 Bayes 估计的海洋测深异常数据探测方法的核心在于直接以水深观测值为研究对象,在(X,τ)($\tau=\sigma_0^{-2}$,σ_0^2为单位权方差)满足无信息先验假设[8,9],且有了水深观测值L的情况下,计算每个观测值L_i含有异常值的后验概率。为此,假定每个水深观测值L_i含异常值的先验概率相等,且都为α,从而Δ_i服从污染正态分布,即

$$\Delta_i \sim (1-\alpha)N(0,\sigma_0^2 P_i^{-1}) + \alpha N(0,k^2\sigma_0^2 P_i^{-1}) \tag{3}$$

其中,无异常观测值L_i对应的观测误差Δ_i服从正态分布$N(0,\sigma_0^2 P_i^{-1})$;含异常观测值L_i对应的观测误差Δ_i服从方差膨胀正态分布$N(0,k^2\sigma_0^2 P_i^{-1})$,$k>1$为一给定的常数。实验证明,当$k\in[3,10]$时,对异常值探测结果的影响不大[8,9]。

水深观测值是以正常值或异常值两种状态存在的,因此在上述假定条件下,对应于每个水深观测值L_i,引入一个识别变量,满足:

$$\delta_i = \begin{cases} 1, \text{第 } i \text{ 个测深值服从分布 } N(0,k^2\sigma_0^2 P_i^{-1}) \\ 0, \text{第 } i \text{ 个测深值服从分布 } N(0,\sigma_0^2 P_i^{-1}) \end{cases} \tag{4}$$

随着识别变量的引入,异常值探测的核心就转变为计算每个观测值L_i含有异常值的后验概率$q_i=p(\delta_i=1|L)$,$i=1,\cdots,n$。当$q_i>0.5$时,认为观测值L_i中含有异常值;否则,就认为L_i是正常值。正常值与异常值的归属是一个非此即彼的问题,因此 0.5 是一个自然而明显的判断标准。

3　后验概率的计算及测深异常值推估解的求取

3.1　MCMC 抽样设计基本原理

要计算识别变量的后验概率值q_i,需要确定识别变量δ_i的后验分布,但是该后验分布比较复杂,且无已有的概率密度函数可以借鉴,为此,考虑采用 MCMC(Markov chain monte carlo)抽样方法计算该后验概率的值。

考虑到:

$$q_i = p(\delta_i = 1 \mid L) = E(\delta_i \mid L) = \int \delta_i \pi(\delta_i \mid L)\mathrm{d}\delta_i = \int \delta_i \pi(\tau,\delta,X \mid L)\mathrm{d}\tau\mathrm{d}\delta\mathrm{d}X \tag{5}$$

记$\boldsymbol{Y}=(\tau,\delta_1,\cdots,\delta_n,X_1,\cdots,X_t)=(Y_1,Y_2,\cdots,Y_{n+t+1})$,则$q_i=\int\delta_i\pi(\boldsymbol{Y}|L)\mathrm{d}\boldsymbol{Y}$。采用 MCMC 方法计算后验概率$q_i$的基本思想是:通过建立一个平稳分布为后验分布$\pi(\boldsymbol{Y}|L)$的 Markov 链来得到后验分布的样本$\tilde{Y}^{(1)}$、…、$\tilde{Y}^{(R)}$,然后基于这些样本计算后验概率$q_i=p(\delta_i=1|L)$的值。

3.2　MCMC 抽样设计步骤

1) 确定识别向量的初始值$\boldsymbol{\delta}^{(0)}=(\delta_1^{(0)},\cdots,\delta_n^{(0)})$

从区域水深观测值中选取n_0个观测值构成初始子集S_0,n_0的确定原则是以很高的概率保证S_0是仅含正常观测值的子集,同时也满足初始子集的容量大于或等于必要观测的条件。本文采用文献[8,9]提出的方法来确定n_0,当n_0确定后,相应的识别变量的初始值$\boldsymbol{\delta}^{(0)}$根据下式进行确定:

$$\delta_i^{(0)} = \begin{cases} 1, L_i \notin S_0 \\ 0, L_i \in S_0 \end{cases}, i = 1,\cdots,n \tag{6}$$

2) 给出初始值向量$\boldsymbol{Y}^{(0)}=(\tau^{(0)},\boldsymbol{\delta}^{(0)},\boldsymbol{X}^{(0)})$

其中识别向量的初始值$\boldsymbol{\delta}^{(0)}$的选择具有基础性的作用,当$\boldsymbol{\delta}^{(0)}$确定后,可以根据公式:

$$\tilde{\boldsymbol{X}} = (\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{L} \tag{7}$$

$$\tilde{\boldsymbol{P}} = \mathrm{diag}\left(\frac{P_1}{1+\delta_1(k^2-1)},\cdots,\frac{P_n}{1+\delta_n(k^2-1)}\right) \tag{8}$$

求得未知参数的初始解$\boldsymbol{X}^{(0)}$。因为$\{\tau|L,\boldsymbol{\delta}^{(0)},\boldsymbol{X}^{(0)}\}\sim\Gamma[n/2,\sum\tilde{\Delta}_i^2/2]$,且$\tilde{\Delta}_i=\frac{(L_i-\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{X})\sqrt{P_i}}{1+\delta_i(k-1)}$,因此可以随机产生出$\tau^{(0)}$,从而形成初始值向量$\boldsymbol{Y}^{(0)}=(\tau^{(0)},\boldsymbol{\delta}^{(0)},\boldsymbol{X}^{(0)})$。

3) 迭代产生样本值向量$\boldsymbol{Y}^{(s)}=(\tau^{(s)},\delta^{(s)},X^{(s)})$

假定第$s\geqslant1$次抽样开始时的样本值向量为$\boldsymbol{Y}^{(s-1)}=(\tau^{(s-1)},\delta^{(s-1)},X^{(s-1)})$,则第$s$次抽样产生的样本值向量为$\boldsymbol{Y}^{(s)}=(\tau^{(s)},\delta^{(s)},X^{(s)})$。其中,$\tau^{(s)}$从条件分布$p(\tau|L,\delta^{(s-1)},X^{(s-1)})$中抽取,而$\{\tau|L,\boldsymbol{\delta},\boldsymbol{X}\}\sim\Gamma[n/2,\sum\tilde{\Delta}_i^2/2]$;$X^{(s)}$从条件分布$p(\boldsymbol{X}|L,\delta^{(s-1)},\tau^{(s)})$中抽取,而$\{\boldsymbol{X}\mid L,\boldsymbol{\delta},\tau\}\sim N(\tilde{\boldsymbol{X}},(\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{A})^{-1}/\tau)$;$\delta^{(s)}$从条件分布$p(\delta_i|L,\delta_1^{(s)},\cdots,\delta_{i-1}^{(s)},\delta_{i+1}^{(s)},\cdots,\delta_n^{(s-1)},X^{(s)},\tau^{(s)})$中抽取,而$\{\delta_i|$

$L,\delta_{-i},\boldsymbol{X},\tau\}\sim b(1,\tilde{q}_i)$，$i=1,\cdots,n$，其中，$\tilde{q}_i=P\left\{\delta_i=1\middle|L,\delta_{-i},\boldsymbol{X},\tau\right\}=\alpha f(\Delta_i\sqrt{p_i\tau}/k)/[\alpha f(\Delta_i\sqrt{p_i\tau}/k)+k(1-\alpha)f(\Delta_i\sqrt{p_i\tau})]$，进行抽样，直到 Markov 链达到稳定，此时，分布收敛到 $\pi(\boldsymbol{Y}|L)$，即得一个 MCMC 样本 $\widetilde{\boldsymbol{Y}}$。

4) 收敛性判断及识别变量后验概率的计算

给定一非常小的常数 $\xi>0$，当 $|\hat{q}_i^{(s)}-\hat{q}_i^{(s-1)}|<\xi(i=1,\cdots,n)$ 成立时，即认为 MCMC 抽样收敛了，从而得到的样本可以看作是来自于平稳分布为 $\pi(\boldsymbol{Y}|L)$ 的 Markov 链的一个 MCMC 样本。重复 MCMC 抽样 R 次，每次抽样均使 Markov 链达到稳定，这样就获得了一个容量为 R 的 MCMC 样本：

$$\widetilde{Y}^{(1)}=(\tau^{(1)},\delta_1^{(1)},\cdots,\delta_n^{(1)},X_1^{(1)},\cdots,X_t^{(1)})$$

$$\cdots$$

$$\widetilde{Y}^{(R)}=(\tau^{(R)},\delta_1^{(R)},\cdots,\delta_n^{(R)},X_1^{(R)},\cdots,X_t^{(R)})$$

在上述样本的基础上，由下述公式计算后验概率值：

$$\hat{q}_i=\frac{1}{R}\sum_{j=1}^{R}\frac{\alpha f(\Delta_i\sqrt{P_i\tau}/k)}{\alpha f(\Delta_i\sqrt{P_i\tau}/k)+k(1-\alpha)f(\Delta_i\sqrt{P_i\tau})} \tag{9}$$

式中，$f(\cdot)$ 为标准正态分布的概率密度函数。根据 $\hat{q}_i$ 的值标定出测深异常值，因为区域内海底的地形变化是连续、平缓的，对于不是异常值的水深，原则上保留原始观测值，而对于异常值，采用区域内正常水深值的加权平均值作为推估值。

4 实验数据与分析

4.1 实测算例

本文的实验数据来自于 CARIS 公司 HIPS&SIPS6.1 软件自带的一条测线的原始多波束测深数据，由于通常情况下边缘波束的质量比较差，所以只取靠近中央波束的数据。采用最小曲率法得到的规则格网构成的矩阵维数为 100×76，同时格网化后的数据仍保留有异常点及较小的随机噪声，测区最大水深为 19.657 0 m，最小水深为 15.696 5 m，如图 1 所示。由于实测海底地形比较平坦，选权迭代加权平均滤波法中的 k_0 取 0.5～1，k_1 取 1～2。对规则格网化后的数据进一步细分成 380 个 4 m×5 m 的局部窗口。为验证文中提出的方法，将其与选权迭代加权平均滤波法进行比较。由文献[2,3]可知，选权迭代加权平均滤波法随着经验值 k_0 与 k_1 取值的不同，所得的残差序列也不相同，分别设为 $v_1,\cdots,v_n$ 和 $v_1',\cdots,v_n'$，考虑如下几种方案：① 基于 $v_1,\cdots,v_n(k_0=0.8,k_1=1.5)$，采用选权迭代加权平均滤波法（$\hat{\sigma}_0^2=(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{v})/(n-t)$）；② 基于 $v_1',\cdots,v_n'(k_0=0.6,k_1=1.2)$，采用选权迭代加权平均滤波法（$\hat{\sigma}_0^2=(\boldsymbol{v}'^{\mathrm{T}}\boldsymbol{P}\boldsymbol{v}')/(n-t)$）；③ 基于 Bayes 估计的异常数据探测法。

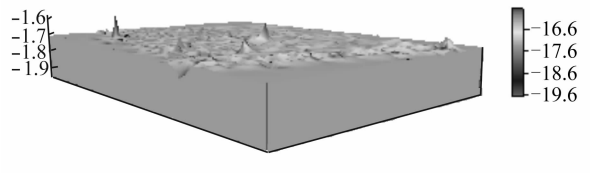


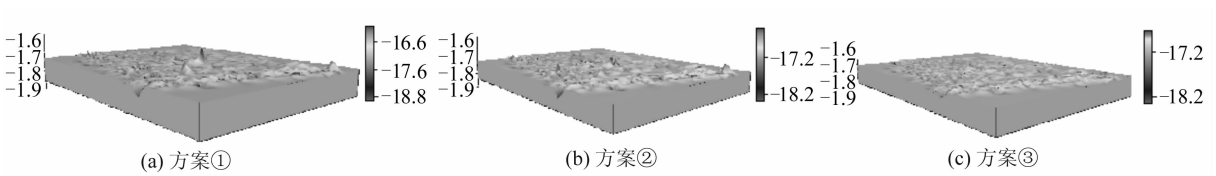
图 1 原始的水深数据

Fig. 1 Original Depth Data

各方案的计算结果如图 2 所示。表 1 的数据是采用方案③计算的第 192 个 4 m×5 m 内所有点的后验概率值。

4.2 模拟算例

为充分体现基于 Bayes 估计的异常数据探测法的优越性，模拟了一组数据，并对 10 个点加入异常，异常值从 −3～3 m，其中在第53个5 m×



(a) 方案① (b) 方案② (c) 方案③

图 2 各方案剔除异常值后的水深数据（实测数据）

Fig. 2 Depth Data by Different Schemes(Real Data)

表 1 采用方案③计算第 192 个 4 m×5 m 内所有点的后验概率值

Tab. 1 Probability of All Nodes of the 192th by the Scheme ③

观测值序号	q_i	观测值序号	q_i	观测值序号	q_i	观测值序号	q_i
$L_{192}(1,1)$	0.017 5	$L_{192}(2,1)$	0.206 5	$L_{192}(3,1)$	0.610 3	$L_{192}(4,1)$	0.013 2
$L_{192}(1,2)$	0.016 4	$L_{192}(2,2)$	0.636 4	$L_{192}(3,2)$	0.999 5	$L_{192}(4,2)$	0.020 0
$L_{192}(1,3)$	0.016 0	$L_{192}(2,3)$	0.013 1	$L_{192}(3,3)$	0.027 6	$L_{192}(4,3)$	0.189 1
$L_{192}(1,4)$	0.016 2	$L_{192}(2,4)$	0.019 3	$L_{192}(3,4)$	0.022 1	$L_{192}(4,4)$	0.068 0
$L_{192}(1,5)$	0.022 9	$L_{192}(2,5)$	0.032 9	$L_{192}(3,5)$	0.027 5	$L_{192}(4,5)$	0.035 3

5 m 内的 $L_{53}(2,2)$ 数据加入−3 m 的异常。将数据格网化后构成的矩阵维数是 100×75,测区最大水深为 17.771 6 m,最小水深为 4.431 1 m,如图 3 所示。对规则格网化后的数据进一步细分为 300 个 5 m×5 m 的局部窗口,仍采用上述的三种方案进行数据处理,其中方案①中取 $k_0=1.2, k_2=2.0$;方案②中取 $k_0=1.0, k_1=1.8$。各方案的计算结果如图 4 所示。表 2 的数据是采用方案③计算的第 53 个 5 m×5 m 内所有点的后验概率值。

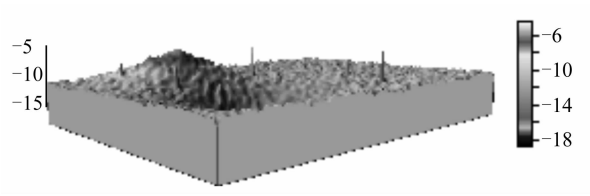
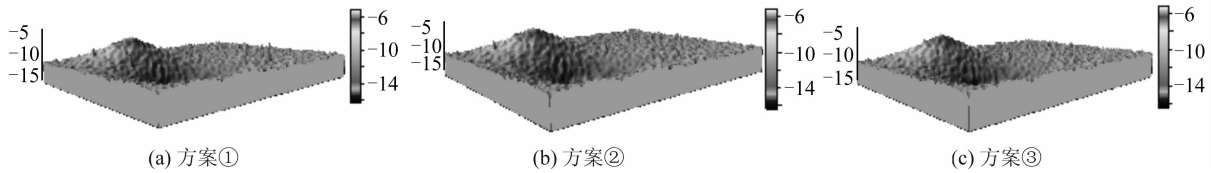


图 3　原始的水深数据

Fig. 3　Original Depth



(a) 方案①　(b) 方案②　(c) 方案③

图 4　各方案剔除异常值后的水深数据(模拟数据)

Fig. 4　Depth Data by Different Schemes(Model Data)

表 2　采用方案③计算第 53 个 5 m×5 m 内所有点的后验概率值

Tab. 2　Probability of All Nodes of the 53th by the Scheme ③

观测值序号	q_i	观测值序号	q_i	观测值序号	q_i	观测值序号	q_i	观测值序号	q_i
$L_{53}(1,1)$	0.039 5	$L_{53}(2,1)$	0.020 9	$L_{53}(3,1)$	0.016 7	$L_{53}(4,1)$	0.013 7	$L_{53}(5,1)$	0.013 0
$L_{53}(1,2)$	0.024 1	$L_{53}(2,2)$	0.999 8	$L_{53}(3,2)$	0.015 5	$L_{53}(4,2)$	0.013 3	$L_{53}(5,2)$	0.013 3
$L_{53}(1,3)$	0.018 5	$L_{53}(2,3)$	0.015 0	$L_{53}(3,3)$	0.013 9	$L_{53}(4,3)$	0.013 0	$L_{53}(5,3)$	0.013 6
$L_{53}(1,4)$	0.014 0	$L_{53}(2,4)$	0.013 4	$L_{53}(3,4)$	0.013 0	$L_{53}(4,4)$	0.013 5	$L_{53}(5,4)$	0.015 7
$L_{53}(1,5)$	0.013 3	$L_{53}(2,5)$	0.013 0	$L_{53}(3,5)$	0.013 4	$L_{53}(4,5)$	0.014 1	$L_{53}(5,5)$	0.018 3

4.3　结果分析

由图 1、图 3 可知,利用观测数据对海底地形进行真实反映的过程中,异常值的影响是很大的,必须对其进行滤波处理。对比图 2(a)、2(b)及图 4(a)、4(b)可知,选权迭代加权平均滤波法虽然在一定程度上剔除了异常值,但是随着残差序列取值的不同,异常值剔除的程度也有所不同。

由表 1 及表 2 可知,利用基于 Bayes 估计的异常数据探测法能快速地计算出水深观测异常值的后验概率,并根据识别变量判断异常值的标准。由于实测算例中 $L_{192}(2,2)$、$L_{192}(3,1)$、$L_{192}(3,2)$ 及模拟算例中 $L_{53}(2,2)$ 的 q_i 值均大于 0.5,因此被确定为异常值,并准确标定。由图 2(c)及图 4(c)可以看出,由于异常值被清晰地标定及拟合,因此更能反映出海底的真实地形。

5　结　语

基于 Bayes 估计的异常数据探测法的基本思想是通过计算水深观测含有异常值的后验概率,并根据水深观测识别变量判别法对后验概率值进行判断,其判断标准清晰明了,并且由于 MCMC 抽样方法的引入,使得后验概率值的计算过程非常简单。实测和模拟算例表明,该算法能合理有效地探测出异常值。

参　考　文　献

[1] 阳凡林,刘经南,赵建虎. 多波束测深数据的异常检测和滤波[J]. 武汉大学学报・信息科学版,2004,29(1):80-83

[2] 何义斌,吴书帮,谢洪燕,等. 多波束异常测深数据检测方法实践[J]. 测绘科学,2004,29(1):50-52

[3] 赵建虎. 多波束深度及图像数据处理方法研究[D]. 武汉:武汉大学,2002

[4] 赵建虎,刘经南,阳凡林. 多波束测深数据系统误差的削弱方法研究[J]. 武汉大学学报・信息科学版,2004,29(5):394-397

[5] Shaw S, Arnold J. Automated Error Detection in Multibeam Bathymetry Data[C]. OCEAN'93, Victoria, BC, Canada,1993

[6] Lirakis C B, Bongiovanni K P. Automated Multibeam Data Cleaning and Target Detection[C]. IEEE Conference and Exhibition, RI, USA, 2000

[7] Bisquay H, Freulon X, Fouquet C, et al. Multibeam Data Cleaning for Hydrography Using Geostatistics[C]. OCEAN'98 Conference, Nice, France,1998

[8] 李新娜,归庆明,许阿裴. 基于识别变量的粗差探测 Bayes 方法[J]. 测绘学报,2008,37(3):355-360

(下转第 214 页)