International Hydrographic Review, Monaco, LXXVI(2), September 1999

ROBUST METHOD FOR THE DETECTION OF ABNORMAL DATA IN HYDROGRAPHY

by HUANG MOTAO, ZHAI GUOJUN, WANG RUI
OUYANG YONGZHONG, GUAN ZHENG [1]

Abstract

Blunder detection is a topic of great interest to hydrographers because undetected blunders significantly distort the observed parameters, e.g., soundings. Based on an analysis of the characteristic of marine surveying, a robust method for the detection of abnormal data (include blunders) in hydrography is proposed in this paper, which is called the robust interpolation comparison test based on robust M-estimation by an iterative calculation procedure. Some questions involved in the implement of the suggested method are discussed in detail. Compared to the existing methods, the new method has more strong capacity of locating abnormal data. A simulation study and an actual numerical example for the process of multibeam soundings are given to test the performance of the proposed method. The results have illustrated the effectiveness of the method in the detection and identification of multiple blunders. The use of the new method will play an important role in improving the quality and reliability of marine measurements in our country.

INTRODUCTION

Compared to terrestrial survey, marine survey is strongly characterised by the dynamic effect. The observations in marine survey are affected not only by atmosphere, but also by the movement and physical property of ocean water. There exist, therefore, more noise sources in marine survey than in terrestrial survey. Taking the shipboard depth sounding for example, the pulse signals emitted from echo sounder could be reflected by floating and swimming living-things (e.g. fishes) and plants during their propagation. These false echoes and additional round trip echoes may result in a big discrepancy between the observed value and the true depth. It is the so-called blundering problem during data acquisition in marine survey. It is obvious that, due to the influences from different kinds of error sources, the blundering possibility in marine survey is much greater than that in terrestrial survey. In addition, it is difficult to make re-observations in marine survey at the

[1] Tianjin Institute of Hydrographic Surveying and Charting, 40, Youyi Rd, Tianjin, 300061, People's.Republic of China

exactly same position and find out the blunders due to lacking necessary checking-conditions. Thus the blundering problem remains in a very important position in the data processing.

In the statistical and geodetic literature, however, a blunder is not precisely defined. It is difficult to define since we are not sure if a blunder is a mistake or if the mathematical model is lacking. Someone calls those observations that are far away from the bulk of the data as blunders or gross errors or outliers. Someone believes that blunders are large in magnitude and much larger than the accidental errors. Obviously, blunder is a kind of abnormal data (called pseudo-abnormal data). In addition to blunders, there exists another kind of abnormal data in marine survey, i.e., the so-called true abnormal data, which are the true records of the observed parameters. These data are of great value to navigation with safety and design of marine engineering. It has been shown that there exist two kinds of abnormal data in marine survey, and it is essential for us to check and find out these observations in the data processing. However whether an abnormal observation is a blunder or not should be made a further investigation on the basis of sea state. As to this problem, this paper is not going to make a detail discussion. The main subject of this paper is to study how to find the existence of abnormal data and then locate them.

The problem of identifying is relatively simple when observations contain a single blunder. But if the observations contain more than one blunder, the problem of identifying becomes more difficult due to the masking or swamping effects. Masking occurs when some abnormal observations go undetected because of the presence of other, usually adjacent, abnormal observations. Swamping occurs when normal observations are incorrectly identified as abnormal ones because of the presence of other, usually remote, abnormal observations. In the past, those abnormal or suspected observations could only be extracted by manually editing blocks of data. With the widespread application of computer for the data processing in marine survey, some methods have been put forward to automatically remove the erroneous data through computer in recent two decades. These methods could be classified into two kinds. One kind is called statistical test based on some hypotheses, e.g., LUO (1984), LI (1988), CHEN (1991), ZHANG (1992) and EEG (1995). The other is called comparison test based on function interpolation and/or stochastic estimation (collocation), e.g., HERLIHY et al. (1992), WARE et al. (1992), SEVILLA (1993), GIL et al. (1993) and ZHU (1998). It should be admitted that if the methods mentioned above are reasonably used, part of abnormal observations could be correctly located and removed during the data processing. The pity is that all the methods mentioned above are based on the classical least-squares estimation. And it is well known that the least-squares estimation is not robust, even a single blunder can spoil the solution. If we use these contaminated estimates to construct statistic variable and then make statistical test, the results will certainly be unreliable. And unavoidably, the problem of masking and swamping effects mentioned above will arise, especially in the multiple-blunder case. To overcome the problem above, this paper makes an attempt to use the robust estimation to increase the reliability of the conventional testing methods for the detection of abnormal data in marine survey. The main idea is to combine the conventional interpolation comparison test with robust estimation, and then a so-called robust interpolation comparison test based on robust M-estimation by an iterative calculation procedure is proposed.

THE ROBUST METHOD

As mentioned above, the conventional least-squares estimation is not robust. It means that the least-squares procedure tends to smooth blunders into good (normal) observations. In other words, the blunders will be spread to other (good and bad) observations and thus the sizes of the actual blunders will be distorted. As a result, incorrect decisions may be derived from the statistical test based on the conventional methods, i.e., a good observation may be rejected or a bad (abnormal) observation may not be detected at all. The advantage of robust estimation is that the negative effect of the blunders on the estimator is greatly softened or even eliminated altogether when making the solution, although the statistical properties of a robust estimator are not as clearly defined and also the efficiency of the estimator is inferior to a least-squares estimator when no blunders are present.

As we know, robust estimation, on the whole, can be classified into three kinds (see HUANG, 1990; YANG, 1993). One is the maximum likelihood type estimation, shortly called M-estimation. Another is the linear combination of order statistics estimation, shortly called L-estimation. And the third one is the rank estimation, shortly called R-estimation. Among the three kinds of robust estimation above, robust M-estimation is the most often used one in geodesy. The method suggested in this paper for the detection of abnormal data is just based on the M-estimation by an iterative calculation procedure. The basic principle of this estimation method with equivalent weights is first summarised as follows:

Consider a general error equation

$$V = A\hat{X} - L \tag{1}$$

$$v_i = a_i\hat{X} - L_i \tag{2}$$

Where $\hat{X}$ is an $m \times 1$ estimate vector of unknown parameters; L is a $n \times 1$ observation vector with a $n \times n$ weight matrix P ; V is a $n \times 1$ residual vector of L; v_i and L_i is the *i*th element of V and L respectively; A is a $n \times m$ design matrix, and a_i is the *i*th row vector of A . The M-estimation with equivalent weights mentioned above means that a suitable function $\rho(v)$ is chosen to satisfy the following condition

$$\sum_i p_i\rho(v_i) = \min \tag{3}$$

Differentiating the expression above with respect to the unknown $\hat{X}$ yields

$$\sum_i p_i\psi(v_i)a_i = 0 \tag{4}$$

Where $\psi(v_i)$ is the derivative of $\rho(v_i)$.

Let

$$\overline{p}_i = p_i \psi(v_i)/v_i \tag{5}$$

Then by substituting (5) into (4), we have

$$A^T \overline{P} V = 0 \tag{6}$$

Or

$$A^T \overline{P} A \hat{X} - A^T \overline{P} L = 0 \tag{7}$$

And

$$\hat{X} = (A^T \overline{P} A)^{-1} A^T \overline{P} L \tag{8}$$

Where $\overline{P}$ is called equivalent weight matrix. The calculation of (8) can be made by iterations. Suppose we have obtained the *k*th estimates of unknown parameters $\hat{X}^{(k)}$ and the residuals $V^{(k)}$, then from equation (8), we get the *(k+1)*th robust estimates

$$\hat{X}^{(k+1)} = (A^T \overline{P}^{(k)} A)^{-1} A^T \overline{P}^{(k)} L \tag{9}$$

The robustness of the estimates above is mainly dependent on the determination of the equivalent weights. Some expressions for equivalent weight, which were derived and modified from conventional $\rho(v)$ and $\psi(v)$ functions, have been proposed by statisticians and geodesists (see HUANG, 1990; YANG, 1993). Here we use the following equivalent weight function based on the IGG scheme which was originally developed by ZHOU (1989) (YANG, 1994):

$$\overline{p}_i = \begin{cases} p_i & |v_i'| \le k_0 \\ p_i k_0 [(k_1 - |v_i'|)/(k_1 - k_0)]^2 / |v_i'| & k_0 < |v_i'| \le k_1 \\ 0 & |v_i'| > k_1 \end{cases} \tag{10}$$

Where $v_i' = v_i / \hat{\sigma}_i$, and $\hat{\sigma}_i^2 = \hat{\sigma}_0^2 / p_i$; $\hat{\sigma}_0^2$ is the estimate of unit weight variance. The constant k_0 is proposed to be 1.0~1.5 and k_1 to be 2.0~3.0.

The so-called robust interpolation comparison test suggested in this paper can be described as a three-step process. First each raw observation takes a turn being the comparison observation, or the observation currently under evaluation. Then the comparison observation is examined relative to the robust weighted average of the neighbour observations taken from the neighbourhood of the comparison point. And finally, a criterion is used to accept or reject the observation based on this comparison.

According to equation (9), the expression of the robust weighted average by an iterative calculation procedure can be directly written as

$$\hat{X}^{(k+1)} = \sum_i \overline{p}_i^{(k)} L_i / \sum_i \overline{p}_i^{(k)} \tag{11}$$

Where $L_i(i = 1,2,...,n)$ represents the neighbour observations taken from the neighbourhood of the comparison observation L_p. It can be seen from (11) that, formally, the general formula of the robust weighted average is identical to that of the conventional weighted average. The only difference is that the weight factor p_i of the conventional weighted average is now replaced by the equivalent weight factor $\overline{p}_i$ of the robust weighted average. Whereas it is this substitution that can make the robust weighted average resist the influence of blunders.

As usual, the initial weights $p_i(i = 1,2,...,n)$ used for the calculation of (10) can be assigned according to their horizontal distances to the interpolated point as follows:

$$p_i = 1/(d_i + \varepsilon)^2 \tag{12}$$

Where d_i indicates the horizontal distance between the observation point L_i and the interpolated point (i.e. comparison point) L_p. ε is an arbitrary constant to deal with p_i approaching infinitude as the denominator of the weight function approaches zero. In practical computation, we can set ε =0.01. It can be seen from (12) and (10) that the observations near the interpolated point, in normal case, have a great influence on the interpolation. Whereas when some observation is contaminated by blunder, the residual of the observation will increase, and the corresponding equivalent weight calculated from (10) will decrease. It means that the influence of the abnormal observation on the interpolation will be descending even if the observed point is so much close to the interpolated point. When $|v_i| > k_1\hat{\sigma}_i$, $\overline{p}_i = 0$, that is to say, the significantly abnormal observation has no influence on the interpolation.

Suppose $\hat{X}_p$ to be the convergence value of equation (11) and L_p the corresponding comparison observation (L_p does not take part in the calculation of $\hat{X}_p$). Then the predicted residual can be defined as

$$\Delta L_p = \hat{X}_p - L_p \tag{13}$$

Finally, the absolute value of ΔL_p can help us make a decision concerning the observation L_p based on a comparison between $|\Delta L_p|$ and a critical value or threshold $\Delta L_{\max}$. Here the critical value $\Delta L_{\max}$ is defined to be the maximum acceptable residual. The magnitude of $\Delta L_{\max}$ is dependent on the accuracy of observation and the perfection of interpolation model. In practical application, $\Delta L_{\max}$ is usually chosen as two or three times the standard deviation of observation.

The key to the above robust method for the detection of abnormal data lies in the determination of a starting value of the unknown parameter. For some cases, the starting solution even can determine whether or not a usable M-estimator is obtained. In our opinion, the starting value of robust solution should be chosen to be robust in order to get a reliable convergence. The median of observations, in this paper, is suggested to be taken as the starting value of $\hat{X}$ for the calculation of (11). That is

$$\hat{X}^{0} = \underset{i}{med}\{L_i\} \tag{14}$$

Where $\underset{i}{med}$ denotes median over i. It is known that the estimator defined by equation (14) has the highest possible breakdown point of 0.5 (Yang, 1993). So it can ensure the stability of the iterative procedure. Yang (1997) suggested that the variance factor used for the iterative procedure be calculated by

$$\hat{\sigma}_0^{(k)} = \underset{i}{med}\{\sqrt{p_i}\left|v_i^{(k)}\right|\}/0.6745 \tag{15}$$

And

$$\hat{\sigma}_i^{(k)} = \hat{\sigma}_0^{(k)}/\sqrt{p_i} \tag{16}$$

According to our experiences in application, it is found that when the equivalent weights are calculated through equation (10), the observations near the comparison point might be rejected to take part in the calculation of the interpolation due to their having a larger weight factor p_i, i.e., a smaller variance factor $\hat{\sigma}_i$, and a bigger ratio of $|v_i|/\hat{\sigma}_i$. As a result, it may cause the loss of interpolation efficiency. The reason for the above result is that we have chosen a special initial weight function (see equation (12)) which gives a great difference among the observations. Whereas such a initial weight function is essential to the improvement of interpolation accuracy in the normal case. In order to resolve the contradiction above, here the variance factor is suggested to be calculated directly by

$$\hat{\sigma}_i^{(k)} = \underset{i}{med}\{\left|v_i^{(k)}\right|\}/0.6745 \tag{17}$$

In this case, the initial weight factor p_i has no more direct influence on the determination of $\hat{\sigma}_i^{(k)}$. And as a result, all of the variance factors $\hat{\sigma}_i$ are kept stable in the whole iterative procedure.

NUMERICAL EXAMPLES

In order to show the efficiency of the robust method proposed above, two numerical examples are given to test its performance. The first example is a simulation study where a set of data are simulated, firstly, which consist of twenty-five observations with an equivalent interval and follow a normal distribution $N(\mu = 10, \sigma^2 = 1)$. The order numbers and magnitudes of the simulated observations are given in Table 1:

Tab.1 The distribution of the simulated data(order number/magnitude)

1 8.18	2 8.42	3 9.00	4 9.49	5 10.16
6 8.49	7 8.64	8 9.14	9 9.79	10 10.54
11 9.19	12 9.38	13 9.93	14 10.46	15 10.96
16 10.20	17 10.55	18 10.76	19 11.11	20 11.49
21 10.47	22 10.81	23 11.11	20 11.69	25 11.94

Then each of the initial observations takes a turn being the interpolated point. And two interpolations for each point are performed by using the robust weighted average and the conventional weighted average, respectively. Based on the predicted residuals (see equation (13)), two standard deviations corresponding to the two methods are computed as: σ (robust)=0.50 and σ (conventional)=0.53. It is shown that, in the normal case (without blunders), the prediction accuracies of the two methods are nearly the same. Finally, to get bad observations, four blunders are added to four initial observations, respectively. Two cases are treated:

Case one—4σ size of blunder is added on observation 8, 12, 14 and 18 at the same time.

Case two—4σ size on observation 8 and 10σ size on observation 12, 14 and 18, respectively.

The predicted residuals corresponding to the two cases above by using the interpolation methods of robust and conventional weighted average are calculated and given in Table 2 and Table 3, respectively.

Tab.2 The predicted residual errors corresponding to case one(robust/conventional)

-0.95 -1.44	-0.82 -1.40	-0.73 -1.36	-0.42 -0.92	0.12 -0.30
-0.77 -1.39	-0.84 -1.89	3.54 2.91	-0.64 -1.36	0.23 -0.26
-0.68 -1.39	3.59 3.03	-0.69 -1.85	3.96 3.57	0.27 -0.46
0.03 -0.44	0.10 -0.78	4.14 3.76	0.02 -0.74	0.43 0.05
0.04 -0.30	0.26 -0.25	0.25 -0.47	0.43 0.12	0.75 0.56

Tab.3 The predicted residual errors corresponding to case two(robust/conventional)

-0.94 -1.90	-0.80 -1.86	-0.68 -1.83	-0.42 -1.38	0.13 -0.76
-0.73 -2.08	-0.75 -2.91	3.56 2.05	-0.58 -2.37	0.26 -0.96
-0.52 -2.59	9.59 8.51	-0.51 -3.82	9.98 9.05	0.31 -1.66
0.13 -1.26	0.27 -2.37	10.19 9.07	0.21 -2.34	0.51 -0.77
0.10 -0.92	0.30 -1.09	0.32 -1.75	0.60 -0.73	0.90 -0.06

As shown in Tab.2 and Tab.3, in both cases, the comparison tests based on the robust interpolation method proposed in this paper can always find the blunders correctly, even for small sizes (4σ). Conversely, the tests based on the conventional interpolation are not so satisfactory, especially in case two, where the predicted residual of good observation (number 13) is even larger than that of bad observation (number 8). It is shown that when there exist multiple blunders and the sizes of them are not uniform, the problem of masking and swamping effect mentioned in the previous section may arise by using the conventional method, whereas it does not by using the proposed robust method here.

The second example is a set of actual soundings produced by the Chinese-developed H/HCS-017 swath bathymetry system. The data set consists of 38 swathes and a total of 198,928 soundings. By using the robust interpolation comparison test to check the data set, a total of 4,742 soundings are found to be abnormal, which account for a 2.38% of the total soundings. The statistical results of test for the data subsets of each swath are given in Table 4.

Tab.4 The statistical results of detecting abnormal observations from multibeam sounding

Swath number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Total soundings	4688	4384	4144	6272	5200	4720	4448	4560	4768	4720	4272	3904	3984	5472	4576	4496	4128	4544	4224
Abnormal soundings	191	195	56	243	135	114	134	85	127	137	17	47	64	18	37	75	113	127	68
Percentage	4.06	4.45	1.35	3.73	2.60	2.42	3.01	1.86	2.66	2.90	0 40	1.20	1 61	0.33	0.81	1.67	2.74	2.79	1.61
Swath number	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Total soundings	4144	4256	4176	6048	7664	7872	6224	5024	3952	5808	7376	6272	4208	7408	6368	5728	7744	6992	4160
Abnormal soundings	78	102	35	273	256	252	230	181	87	68	265	181	0	199	109	67	159	153	73
Percentage	1.88	2.40	0.84	4.51	3.34	3.20	3.70	3.60	2.20	1.17	3.59	2.89	0	2.69	1.71	1.17	2.05	2.19	1.75

As shown in Tab.4, the percentages of abnormal observations found in this example are a little larger than that listed in HERLIHY (1992). Having made a further analysis on the detected abnormal data, it is found that more than 90% of abnormal observations are located in the border areas of each swath. It gives an indication of where improvement should be made in this new sounding system in order to increase the accuracy and reliability of soundings.

CONCLUSIONS

High volume data acquisition techniques for mapping the seabed, e.g., multibeam echosounding system and airborne laser depth sounding system, have recently become available and adopted for use in China. These systems have a number of features in common. A high data rate is one of them. As an important part of the quality control of data, it is essential for us to develop a valid method in time for detecting automatically the abnormal data from the high volume observations. This paper has introduced the theory of robust estimation to the data processing of hydrography for the first time. The purpose is to promote people to pay more attentions on the quality control and reliability of data in hydrography in China. Our proposed procedure for the detection of abnormal data is a composite of the robust estimation and the conventional interpolation comparison. The simulation study and the practical process of swath bathymetry data have proven the new method to be an effective procedure of detecting abnormal data. It is especially impressive for the multiple blunder case.

References

[1] CHEN, S. J. and MA, J. R., The application and analysis of marine observation data, Beijing: Marine publishing house, 1991 (in Chinese)

[2] EEG, J., On the identification of spikes in soundings, International Hydrographic Review, 1995, LXXII(1):33~41

[3] GIL, A. J., et al., A method for gross-error detection in gravity data, International Geoid Service, Bulletin No.2, 1992: 25~31

[4] HERLIHY, D. R., et al., Filtering erroneous soundings from multibeam survey data, International Hydrographic Review, 1992, LXIX(2): 67~76

[5] HUANG, Y. C., Robust estimation and data snooping, Beijing: Publishing house of surveying and mapping, 1990 (in Chinese)

[6] LI, D. R., Error processing and reliability theory, Beijing: Publishing house of surveying and mapping, 1988 (in Chinese)

[7] LUO, N. X., Data processing and survey error, Beijing: Measurement publishing house, 1984 (in Chinese)

[8] SEVILLA, M. J., Analysis and validation of the D.M.A gravimetric data of the Mediterranean sea, GEOMED Report-3, 1993:78~101

[9] WARE, C., et al., A system for cleaning high volume bathymetry, International Hydrographic Review, 1992, LXIX(2): 77~94

[10] YANG, Y. X., The theory and application of robust estimation, Beijing: Bayi publishing house, 1993 (in Chinese)

[11] YANG, Y. X., Robust estimation for dependent observation, Manuscr geod, 1994(19): 10~17

[12] YANG, Y. X., Estimators of covariance matrix at robust estimation based on influence functions, ZfV, 1997(4): 166~174

[13] ZHANG, F. R. and J. T. ZHANG, The distribution and statistical test of survey errors, Beijing: Measurement publishing house, 1992 (in Chinese)

[14] ZHOU, J. W., Classical theory of errors and robust estimation, ACTA GEODAETICA et CARTOGRAPHIC SINICA, 1989(2): 115~120

[15] ZHU, Q. and D. R. LI, Error analysis and processing in multibeam soundings, Journal of Wuhan Technical University of Surveying and Mapping, 1998(1):1~4 (in Chinese).