5-2001

Automatic Detection of Outliers in Multibeam Echo Sounding Data

Tianhang Hou
University of New Hampshire, Durham, Tianhang.Hou@unh.edu

Lloyd C. Huff
University of New Hampshire, Durham

Larry A. Mayer
University of New Hampshire, larry.mayer@unh.edu

Automatic Detection of Outliers in Multibeam Echo Sounding Data

Tianhang Hou, Lloyd C Huff, and Larry Mayer

Center for Coastal and Ocean Mapping
University of New Hampshire
Durham NH 03824

Abstract

The data volumes produced by new generation multibeam systems are very large, especially for shallow water systems. Results from recent multibeam surveys indicate that the ratio of the field survey time, to the time used in interactive editing through graphical editing tools, is about 1:1. An important reason for the large amount of processing time is that users subjectively decide which soundings are outliers. There is an apparent need for an automated approach for detecting outliers that would reduce the extensive labor and obtain consistent results from the multibeam data cleaning process, independent of the individual that has processed the data.

The proposed automated algorithm for cleaning multibeam soundings was tested using the SAX-99 (Destin FL) multibeam survey data [2]. Eight days of survey data (6.9 Gigabyte) were cleaned in 2.5 hours on an SGI platform. A comparison of the automatically cleaned data with the subjective, interactively cleaned data indicates that the proposed method is, if not better, at least equivalent to interactive editing as used on the SAX-99 multibeam data. Furthermore, the ratio of acquisition to processing time is considerably improved since the time required for cleaning the data was decreased from 192 hours to 2.5 hours (an improvement by a factor of 77).

1.0 Introduction

When using multibeam sonar systems to map the true seabed topography, hydrographers expect their end product, bathymetric maps and charts, to be free from the effect of outliers or bad depths. In many instances, clean data is the result of the outliers or bad depths having been manually detected by operators using a computer graphic editing tool. In this data cleaning technique, a group of soundings are presented to users through a graphical interface and the user decides which sounding or soundings are the outliers [3]. With the continually increasing volume of data from advanced multibeam sonar systems, the manual editing process has become very time consuming, especially in shallow water surveys. Human intervention for removal of outliers or bad depths is based on subjective decisions make by an individual. Consequently the multibeam data manual cleaning process is not repeatable for the same or different users. Automatic outlier detection is an approach that can overcome many disadvantages of the manual approach to data cleaning, such as: time requirement, repeatability, and human error. However, it is necessary to find a set of objective criterions that adapt to the complicity of multibeam data in order to remove outliers and keep targets. Automatic outlier detections, are based on two primary assumptions. Firstly, that the natural sea-floor topography varies smoothly and continuously, as measured with the vertical and horizontal resolutions of new generation multibeam systems and secondly, that significant deviation from the assumption of smoothly varying bathymetry (target) is indicated when neighbor soundings from different pings hit the target point, two or more times. Two of the requirements that automatic outlier detections must satisfy are: (1) Process large volumes of data in a time efficient manner, (2) Employ objective criteria for detecting the outliers, and (3) Must not remove small targets or create a bias in detailed seafloor topography.

2.0 Proposed Outlier Detection Algorithms

The proposed algorithm is based on a buffer of 60 successive pings, represented by a rectangular matrix that is populated by soundings that are indexed in the vertical by the ping number and indexed in the horizontal by the beam number. Due to the high ping rate of shallow water surveying, the arrangement of close neighbor points used in the outlier detection algorithm is relatively insensitive to variations in the ship's roll, pitch, and yaw. The rectangular matrix is operated on using the algorithm for global and local variance estimation, the algorithm for two-sample variances[4], and the algorithm for bad ping detection. If one or more of the three individual algorithms detects a point as an outlier, then that point is considered to be an outlier.

If the soundings were geo-referenced, the Tin Method could be used to obtain the close neighbor points used in detecting outliers. However, there are drawbacks to using the Tin Method, such as: (1) long computing time, (2) fewer close neighbor points for performing outlier detection, and (3) boundary distortion.

2.1 Outlier detection by global and local variances

2.1.1 Global variance estimation

In multibeam surveying, the seabed can reasonably be modeled using linear slopes, which may fluctuate with short and long scale spatial variations. Removing long scale spatial trends from the seabed representation is critical when estimating global parameters for the background noise levels, which will serve as the dynamic threshold that is used in outlier detection.

The rectangular matrix consists of a buffer of 60 successive pings. Each ping has 128 beams for EM3000, 60 beams for EM1000. The rectangular matrix is given as:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & a_{3m} \\ a_{n1} & a_{n2} & a_{n3} & a_{nm} \end{bmatrix}$$

De-trending to remove long spatial scale variations in the seabed representation is accomplished using second order differentiation, which is applied in both along track and across track directions:

along track $b_{11} = a_{21} - \frac{a_{11} + a_{31}}{2}$, cross track $C_{11} = a_{12} - \frac{a_{11} + a_{13}}{2}$.

The average of the global variances in both along and cross track differentiations is computed by

$$\sigma^2{}_{gloabel} = \frac{0.5}{p-1} \sum_{k=1}^{p} (b_k - \bar{b})^2 + \frac{0.5}{q-1} \sum_{k=1}^{q} (c_k - \bar{c})^2$$

Where: p = (ping number-2)* (beam number) ,
q = (beam number-2)*(ping number)

2.1.2 Local variance estimation

The close neighbor depth points are defined in a regular grid (2-D matrix) that is used to compute the local variance σ^2_{local}. The local variance may change significantly from point to point when the local data are noisy. In which case, it is unlikely that a satisfactory result will be obtained without somehow using a combination of the global variance and the local variance, since a criterion using only the local variance is insensitive to small outliers when the local data are noisy.

In this algorithm, the global and local variances are combined usisng the following logic:

If global variance > local variance, then the global variance is used for local outlier detection.

If global variance < local variance, then the variance used for local outlier detection is defined as: 0.5*(global variance + local variance)

The program allows the use of different ad hoc detection threshold scaling of a reference sigma, depending on whether the suspected outlier is shoaler or deeper than the local area. For the examples presented in this paper, a value of 2.2 times the reference sigma was used for shoal outliers and value of 2 times the reference sigma was used for deep outliers.

2.2 Outlier detection by testing two sample variances

In some cases, outliers that would have been detected using manual editing, are not detected by the global and local variances. Therefore a additional algorithm was designed and implemented in the program along with the rational that detection by one or more algorithms was indeed a detection. Assuming an outlier occurs at the center of the fixed working window, the sample H1 and sample H2 can be defined from their close neighbor depth points as:

H1= $\begin{bmatrix} a_{i-1j-1} & a_{i-1j} & a_{i-1j+1} \\ a_{ij-1} & a_{ij} & a_{ij+1} \\ a_{i+1j-1} & a_{i+1j} & a_{i+1j+1} \end{bmatrix}$, and H2= $\begin{bmatrix} a_{i-1j-1} & a_{i-1j} & a_{i-1j+1} \\ a_{ij-1} & & a_{ij+1} \\ a_{i+1j-1} & a_{i+1j} & a_{i+1j+1} \end{bmatrix}$.

The test variable G is constructed using the expression

$$\mathbf{G} = \frac{\frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{3}(a_{ij}-\bar{a}_{H1})^2}{\frac{1}{8}\sum_{i=1}^{3}\sum_{j=1}^{3}(a_{ij}-\bar{a}_{H2})^2} > \alpha \cdot \sigma\ global\,.$$

Where α is an ad hoc constant such that the term $\alpha \cdot \sigma$global=3.73.

2.3 Bad ping detection

A single bad ping can be found in manual editing of multibeam bathymetric data. It is hard to automatically detect these outliers in a small sample of close neighbor points by the methods described above. For detecting a single bad ping, the close neighbor matrix is re-arranged into three matrixes as:

$$\mathrm{LL} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_{21} & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}$$

$$\mathrm{L1} = \begin{bmatrix} a_1 & a_2 & a_3 \\ c_1 & c_2 & c_3 \end{bmatrix},\ \mathrm{L2} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix},\ \mathrm{L3} = \begin{bmatrix} b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}$$

Then, computing the standard deviations of sub-sets L1, L2, L3 based on the average $\overline{L1}$ of the set L1,

$$S_i^2 = \frac{1}{5}\sum_{k=1}^{6}(Li_k - \overline{L1})^2\,,$$

If both $\frac{S_2^2}{S_1^2}$, and $\frac{S_3^2}{S_1^2}$ are greater than a constant K, a significant difference is detected between the middle row and the other two rows. In the next step, the point evaluated must be tested from subset L1 as

$$(b_{21} - \overline{L1}) > s_{LL}.$$

If all three conditions are satisfied, then the point under evaluation is detected as a bad ping and therefore is deemed to be an outlier.

2.4 Considerations for additional improvements

All the above methods are presently implemented into one automated program which functions without human interactive control during the data editing process. Can this be further improved? Probably, yes. Consider that the cross track beam-to-beam intervals and the random error in depth estimation vary between the nadir and off-nadir portions of the multibeam swath and that outlier detection using the close neighbor algorithm is more reliable for beams closer to nadir, as compared to beams that are farther from nadir. Therein lies good cause for further investigations into whether or not there is a benefit of having the thresholds for detecting off-nadir beam outliers systematically relaxed, compared to those for detecting near-nadir beam outliers. Such a modification to the automatic editing may provide more robust editing of bathymetry data associated with small-scale topographic seabed features that have been sampled with reduced spatial resolution. Another topic for investigation and possible inclusion in the algorithms for automatic detection of outliers in multibeam echo sounding data is the use of multibeam backscatter information in the detection and/or validation of potential outliers.

3.0 Examples

Examples are presented using Numerical and Visual Displays to illustrate the performance of the automatic cleaning algorithms under different circumstances. The examples also serve to demonstrate that the automatic cleaning of multibeam data requires several different algorithms in order to detect the same sorts of outliers that would be detected by manual editing.

3.1 Numerical Examples

An example where an outlier, that would have been detected by manual editing, and can be detected by the local and global algorithm, but cannot be detected by the two-sample variance testing algorithm[4].

$$\begin{bmatrix} 18.64 & 18.48 & 18.51 \\ 18.47 & 18.92 & 18.60 \\ 18.40 & 18.60 & 18.59 \end{bmatrix},$$

Here, the mean is18.5789 and the global sigma is 0.0924. The local sigma is 0.1498, the confident range is {18.3125, 18.8211}, and G = 3.2324. The outlier is not detected because G is < 3.73. However, the point is still defined as an outlier because it was detected using the local and global algorithm.

An example where the outlier, that would have been detected by manual editing, is detected by the two-sample variance testing algorithm and could not be detected by the local and global algorithm;

$$\begin{bmatrix} 18.48 & 18.56 & 18.54 \\ 18.53 & 18.69 & 18.54 \\ 18.52 & 18.50 & 18.56 \end{bmatrix},$$

The mean is 18.5467, the local sigma is 0.0598, the confident range is {18.3434, 18.7315}, and G = 4.5604. The outlier is detected because G is >3.73. Detection by one or more of the three individual algorithms constitutes a detection of the outlier.

An example where the outlier, that would have been detected by manual editing, can be detected using the global and local variance, but could not be detected using only the local variance. This example illustrates the type of misdetections that could occur without inclusion of the rule stated above for combining global and local variances when local variance >> global variance.

$$\begin{bmatrix} 18.48 & 18.45 & 18.33 \\ 18.48 & 18.21 & 18.48 \\ 18.57 & 18.57 & 18.62 \end{bmatrix},$$

The mean is 18.4656, the local sigma is 0.1276, the confident range is {18.1849, 18.7207}, and G= 2.0078. The outlier is not detected using the local variance algorithm because G< 3.73. If however the local and global parameter is used, the acceptance range changes to {18.2236, 18.6855} and the outlier is detected. A detection by one or more of the three individual algorithms constitutes a detection of the outlier.

An example of an outlier, that would have been detected by manual editing and could not be detected in the local and global algorithm or by the two-sample variance test algorithm, but is detected by the bad ping algorithm.

$$\begin{bmatrix} 18.48 & 18.45 & 18.43 \\ 18.65 & 18.68 & 18.69 \\ 18.40 & 18.39 & 18.38 \end{bmatrix},$$

$\frac{S_2^2}{S_1^2}$ = 26.0780, $\frac{S_3^2}{S_1^2}$ =25.9354, $(b_{21} - \overline{L1})$ =(18.68-18.4217)=0.2583, and S_{LL} =0.1299. The 18.68 sounding can be detected as an outlier in the bad ping detection algorithm, because the test value $(b_{21} - \overline{L1})$ is greater than the threshold parameter S_{LL}. The outlier is detected using the bad ping algorithm, whereas it was not detected in the local and global algorithm or by the two-sample variance test algorithm. A detection by one or more of the three individual algorithms constitutes a detection of the outlier.

3.2 Visual Display Examples

In order to better understand the results obtained using the automated approach to multibeam data cleaning, the proposed algorithm was imported to the VISE – Visual Swath Editor[1]. This was accomplished with the help of the developer of VISE, Jorgen Eeg of The Royal Danish Administration of Navigation and Hydrography. In the VISE swath edit tool, outlier candidates (or spikes) can be marked on a 3-D view image, then sorted by their outlier characteristics.

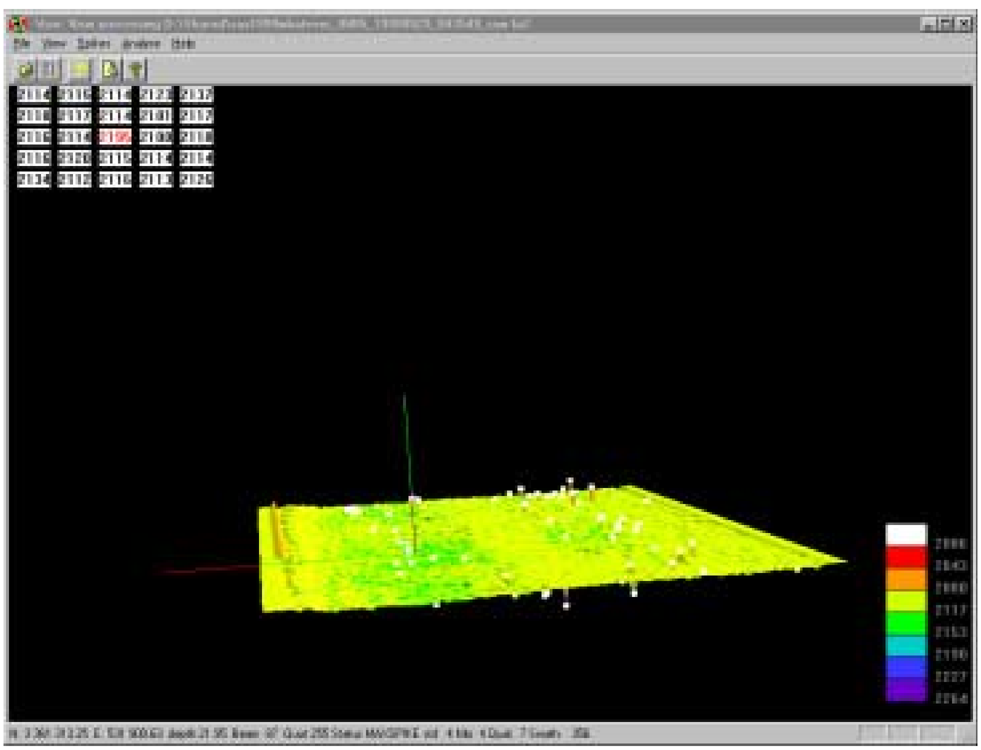

Figure 1 Three-dimensional surface view of Candidate Outliers

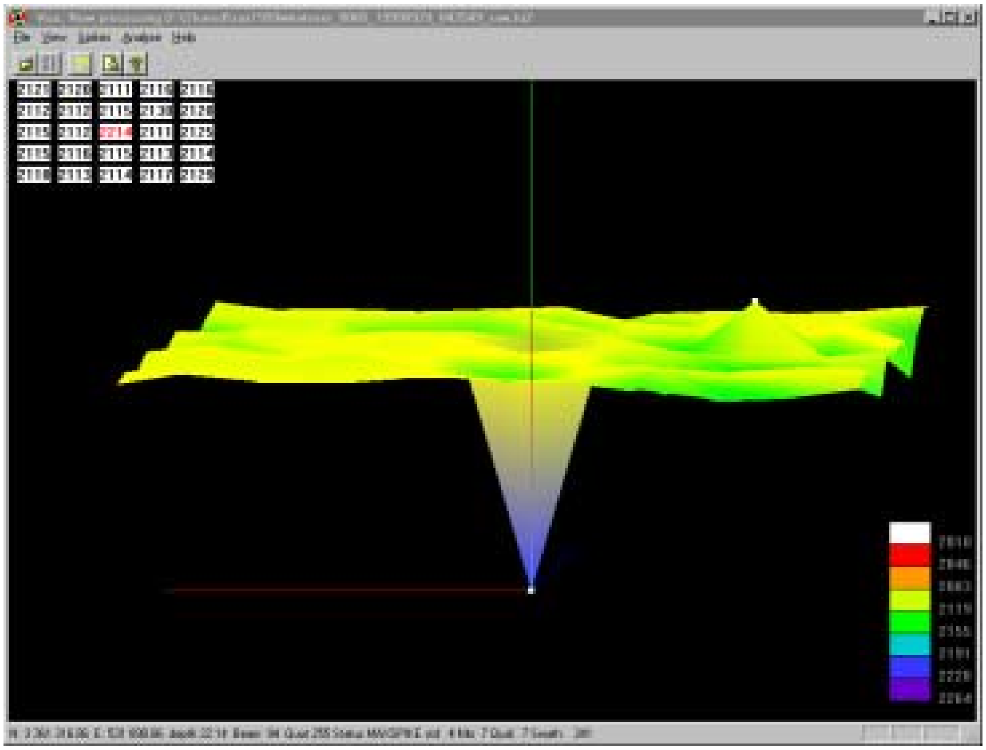

Figure 2 Three-dimensional surface view; zoomed to a single Candidate Outlier

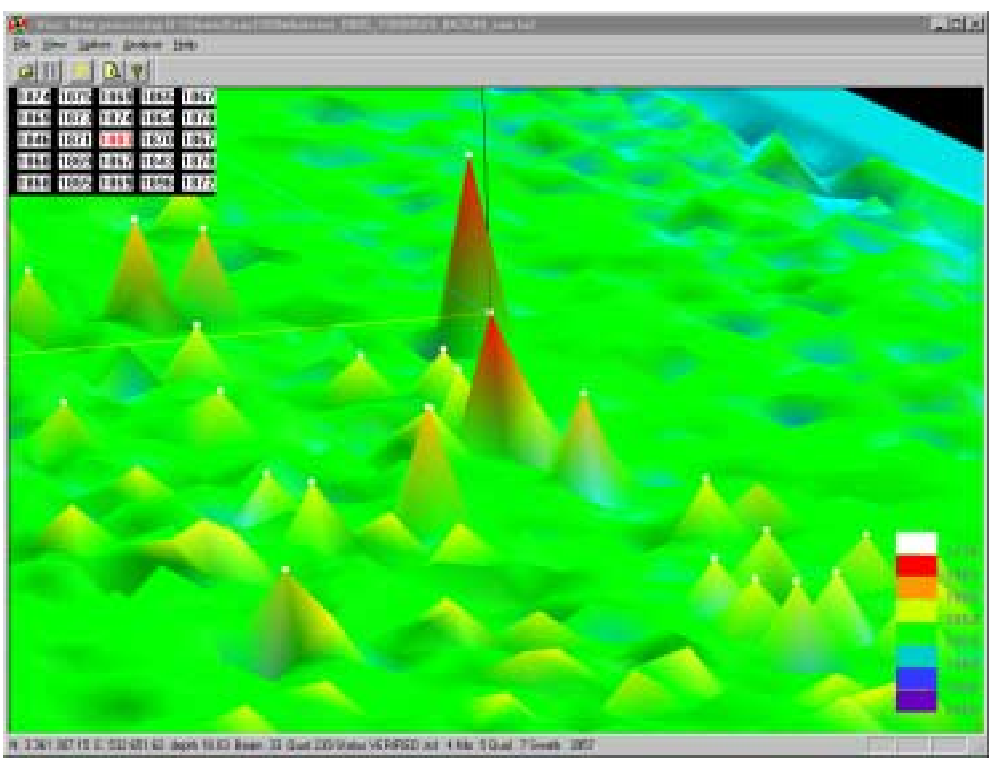

Figure 3 Three-dimensional surface view; zoomed to several Candidate Outliers

4.0 Comparison with human intervention method using SAX-99 data

The proposed automated technique for cleaning multibeam soundings was tested using the SAX-99 (Destin FL) multibeam survey data, acquired by Roger Flood [2]. Using the automatically cleaned data, manually cleaned data, and original (raw) data, three topographic surfaces were created with the pixel size of 0.5 meter. Figures 4, 5, and 6 show comparisons between the three topographic surfaces. The horizontal axes in those three figures are pixel index numbers (equivalent to 0.5 m in physical space).

The numerical results of comparing of the Manually Cleaned and Raw Data topographic surfaces are: Mean difference =0.0816 m, Max difference= 0.5 m, Min difference=-0.39 m, and the standard deviation of differences=0.0652 m.

The numerical results of comparing of the Auto-Cleaned and Raw Data topographic surfaces are: Mean difference =0.0826 m, Max difference= 0.5 m, Min difference=-0.24 m, and the standard deviation of differences=0.065 m.

The numerical results of comparing of the Manually Cleaned and Auto-Cleaned topographic surfaces are: Mean difference =0.011 m, Max difference= 0.68 m, Min difference=-0.27 m, and the standard deviation of differences=0.0097 m.

.

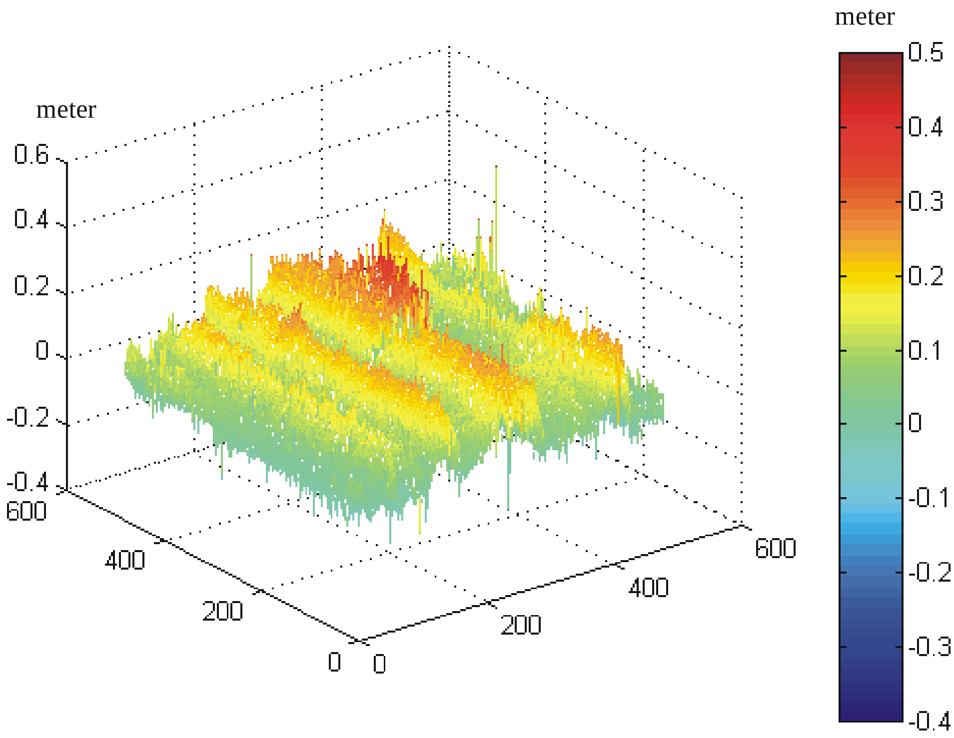

Figure 4 Differences between topographic surfaces of the Manually Cleaned and Raw Data

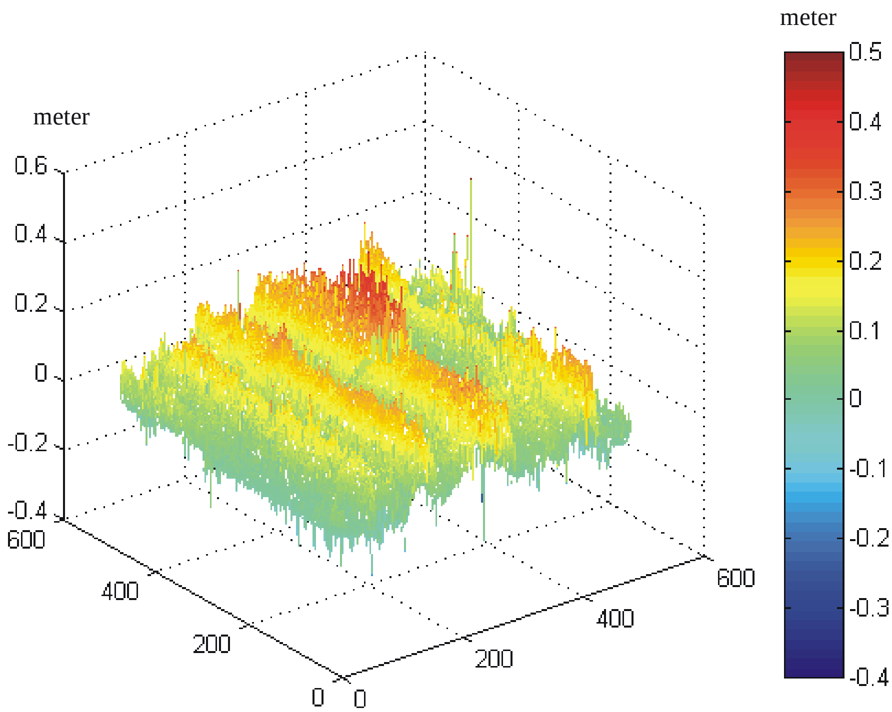

Figure 5 Differences between topographic surfaces of the Auto-Cleaned and Raw Data

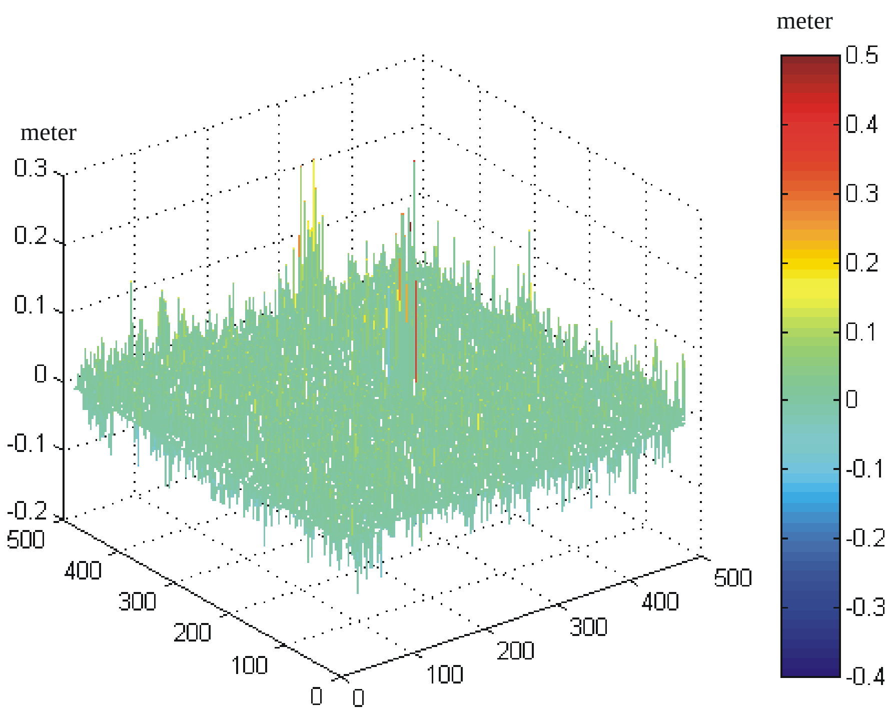

Figure 6 Differences between topographic surfaces of the Manually Cleaned and Auto-Cleaned data

Figure 6 presents a clear indication that the results of manual cleaning and the results of this scheme for automatically cleaning multibeam data s are very similar. The automated cleaning has not created a bias in the seafloor topography, relative to the manually cleaned data. However, the two methods do differ in some particular instances. In the opinion of the authors, those few instances of significant (greater than 2 sigma) differences in the two data cleaning techniques are attributable to the subjective decisions made during their manual cleaning of the data.

Figure 7 shows two applications of the 2-D Graphic Viewer, developed by John Hughes-Clark. The original (raw) data is presented on the left and the data after automatic cleaning is presented on the right. The two views, left and right, of *Beam Profiles* and *Swath Profiles* provide a dramatic comparison between with the original (raw) data and automatically cleaned data, using a tool that is commonly applied to the editing and viewing of multibeam data.

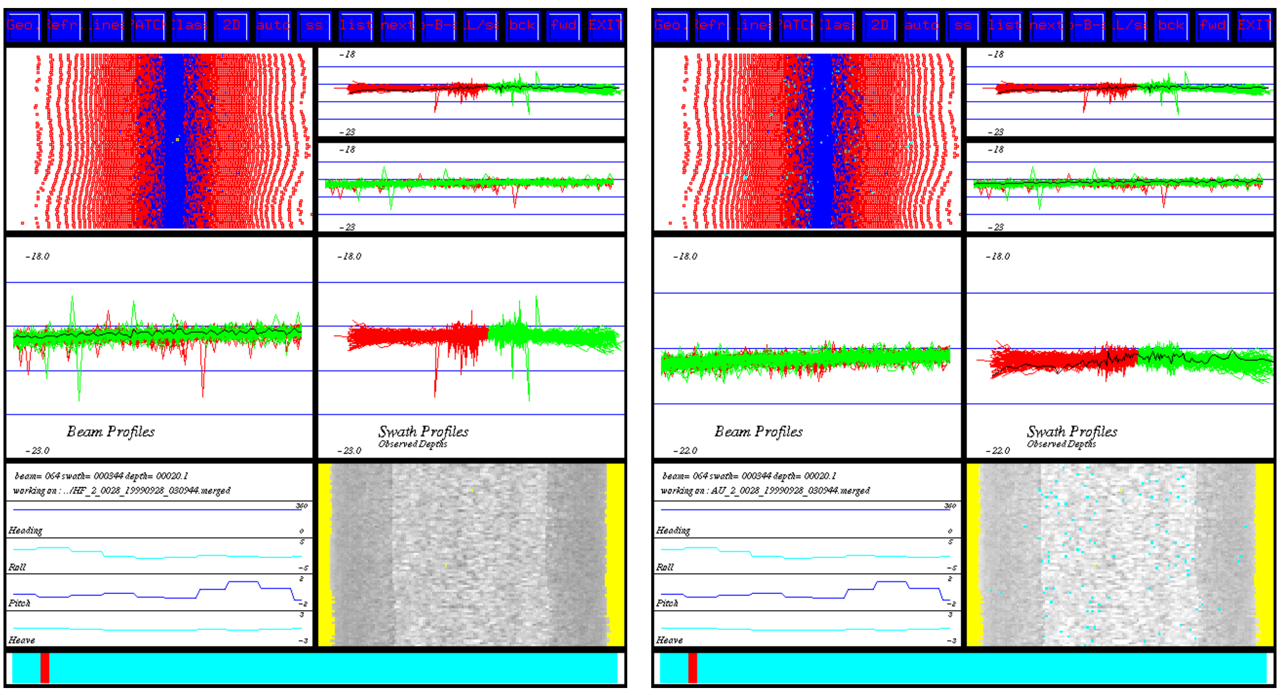

Figure 7 Comparison of un-clean and automatically cleaned data using 2D Graphic Viewer

5.0 Conclusions

This paper has described the rational, development and performance of a set of three individual algorithms for detecting specific types of outliers in multibeam data. A detection by one or more of the three individual algorithms constitutes a detection of the outlier, consequently the process can reliably detect outliers of the types most often detected during manually editing. Although the intent has been to emulate the performance of good manual editing by a well-trained, highly motivated operator, the algorithms are based on objective criteria. These algorithms are designed operate quickly on large volumes of data, while avoiding the variability of manual editing.

A comparison of the automatically cleaned data with the subjective, interactively cleaned data indicates that the method described in this paper is, if not better, at least equivalent to the manual editing when used on the SAX-99 multibeam (EM3000) data. Furthermore, the ratio of acquisition to processing time is considerably improved since the time required for cleaning the SAX-99 multibeam data was decreased from 192 hours for manual cleaning to 2.5 hours for automatic cleaning(an improvement by a factor of 77).

6.0 References

[1] Eeg, Jorgen, VISE – Visual Swath Editor, personal communications, September-December, 2000.
.
[2] Flood, Roger, EM3000 Multibeam survey data from SAX-99 (Destin FL), personal communication.

[3] Ware, C., L. Slipp, K. Wang, B. Nickerson, D. Wells, Y.C. Lee, D. Dodd and G. Costello, 1992, “A System for Cleaning High Volume Bathymetry”. The International Hydrographic Review. Vol. LXIX. No. 2, pp.77-94.

[4] John Neter, W. Wasserman and G.A. Whitmore, 1978, Applied Statistics. Allyn and Bacon, Inc., Toronto.