International Hydrographic Review, Monaco, LXXVI(1), March 1999

AUTOMATIC DETECTION OF PUNCTUAL ERRORS IN MULTIBEAM DATA USING A ROBUST ESTIMATOR

par N. DEBESE [1], H. BISQUAY [1]

Abstract

The Oceanographic and Hydrographic service of the Navy (**SHOM**) has been using two MultiBeam Echo-Sounders (MBES) since 1988. These systems enable swath coverage of the sea floor along a survey line.

Compared with single beam Echo-sounder systems, the resolution of the data provided by these systems has been considerably increased. Nevertheless, errors still remain and they must be detected and eliminated to meet the international standards of bathymetric charts.

The high volume of data, particularly in the case of very shallow water Echo-Sounder systems, makes manual validation of the data inappropriate. In order to reduce the operating costs of the data cleaning step, **SHOM** has developed algorithms to automatically detect huge datasets generated by MultiBeams.

The algorithm described in this paper is based on a local modelization of the seabed. The fitting of a quadratic surface over the raw data is carried out using a robust estimator. We retained *Tukey* robust estimator as the most effective choice due to its adaptative capabilities. Possible outliers are soundings with high residual values between measured depths and depths estimated from the local model. Retained outliers are deduced from this first outliers set, by computing local cross validation.

This algorithm has been tested on different bathymetric data sets. Its efficiency has been demonstrated whatever the depth or type of seabed. Moreover, its application only requires two parameters to be set, thus making it the obvious choice. It has currently been adopted and installed on board all the **SHOM's** ships.

[1] EPSHOM, 13 rue du Chatellier, BP 426, 29275 Brest Cédex.

1. INTRODUCTION

SHOM has been using MultiBeam Echo-Sounders (MBES) for 10 years to carry out bathymetric surveys. These systems apply the beamforming sonar technique to obtain several measurements from one acoustic ping. An acoustic pulse is sent to the bottom and depth measurements are computed from the travel times of the echos reflected by the sea bottom. The measurements are not limited to a vertical one, as is the case for Single Beam Echo-Sounders, but on either side of the vessel's course.

The use of a MultiBeam Echo-Sounder to carry out bathymetric charts gives more accurate measurements and greater resolution. Nevertheless, experience shows that these data contain sparse erroneous soundings. Errors can result from surface reflection, low signal to noise ratio in bad weather conditions or turbulent flows with bubbles that cause interference with the transducers. Even if the error rates still remain generally low, the cleaning step is essential, if accurate bathymetric charts ensuring navigational safety are to be drawn up. For example, a study carried out by **SHOM** (DEBESE, 1997) shows that the error rate is less than 0.5%, in the particular case of the SIMRAD EM12-dual Echo-Sounder .

Removing erroneous soundings from the dataset is a crucial post-processing step. Two approaches are generally encountered for this cleaning step:

- The first one is entirely manual. A trained operator has to visualize, one by one, all of the soundings of a survey. The identification of the erroneous soundings, which are based on local validations of the bathymetry, is the responsibility of a trained operator.
- The second one is, on the contrary, entirely automatic. The identification of the potentially erroneous soundings is obtained through the application of algorithms. It is a question of validating a set of *a priori* defined rules.

SHOM has chosen an intermediate solution by combining features from both approaches. The validation of the data is the responsibility of the trained operator who decides to validate or invalidate the doubtful soundings pointed out by an algorithm or a set of algorithms.

This hybrid approach was chosen to ensure the homogeneity of the various processings which were inevitably performed by different trained operators. Such an approach also gives a good ratio between the processing time and the quality of validation.

All the algorithms concerning the automatic detection of erroneous soundings that result from several studies directed by **SHOM** are based on the assumption of a local continuity of the bathymetry. However, these algorithms can be divided into two classes. The first one consists of algorithms deduced from an *a posteriori* defined classification of a set of approximately five million manually cleaned soundings. This study (DEBESE, 1997) has given rise to a protocol of three

algorithms specifically dedicated to data acquired by the SIMRAD EM12-dual, a deep water Echo-Sounder.

The algorithm described in this paper belongs to the second class. The detection of punctual errors relies on local modelization of the seabed. The retained model is a quadratic surface. As introduced in (DEBESE, 1998), the building of the model is directly applied to raw data using a robust estimator. The weighting estimators, also called W-estimators, are robust and straightforward to implement. In this class of robust estimators, we have retained the *Tukey* because of its adaptative capability. Like most robust methods, the estimator uses a residual measurement of the information to identify potentially erroneous soundings: a high residual value indicates a sounding which is largely deviated compared to the presupposed model. Soundings detected as outliers are deduced from this first set of doubtful soundings through local cross validation.

The algorithm dealing with the automatic detection of punctual errors in bathymetric data is described in the paragraph hereafter. Its evaluation was carried out on five sets of real data acquired by several MultiBeam (shallow and deep water) EchoSounders. Retained datasets were chosen because of the diversity of their reliefs. These datasets are presented in paragraph 3, criteria and results of the evaluation are provided in paragraph 4.

2. DESCRIPTION OF THE ALGORITHM

2.1 General principle

The algorithm is based on the assumption that at least one representation scale of the seabed exists, and its topography can be modelized using a quadratic (1) surface. As this goal could not be achieved over all the geographic area, it is necessary to subdivide it into sub-areas. In this case, we propose a division into squared cells of identical size (L). If a local quadratic model is statistically verified, the residual between measured depths and depths estimated from the model can be used to check the consistency of each sounding with its neighbourhood.

$$z = a_5 x^2 + a_4 y^2 + a_3 xy + a_2 x + a_1 y + a_0 = A \cdot X \qquad (1)$$

When the problem addresses the detection of erroneous soundings in bathymetric data, measured residuals are due to two types of noise (GAUDIN, 1996), that is to say:

- the ***MBES** noise*, which we suppose to be gaussian,
- the *noise of punctual errors*, coming from erratic phenomena where the distribution function is unknown, but non-gaussian.

As data contain sparse erroneous soundings, as is the case here, a standard estimation procedure, such as the least squares one (2), cannot be used because all the soundings are taken into account in the same way, to estimate the parameters of the model.

$$\hat{A} = \arg\min_{A} \sum_{i=1}^{N} \left(z_i - A \cdot X_i\right)^2 \qquad (2)$$

As described in Fig.2-1 (a), valid soundings are pointed out as non-valid as the surface is adjusted according to erroneous soundings.

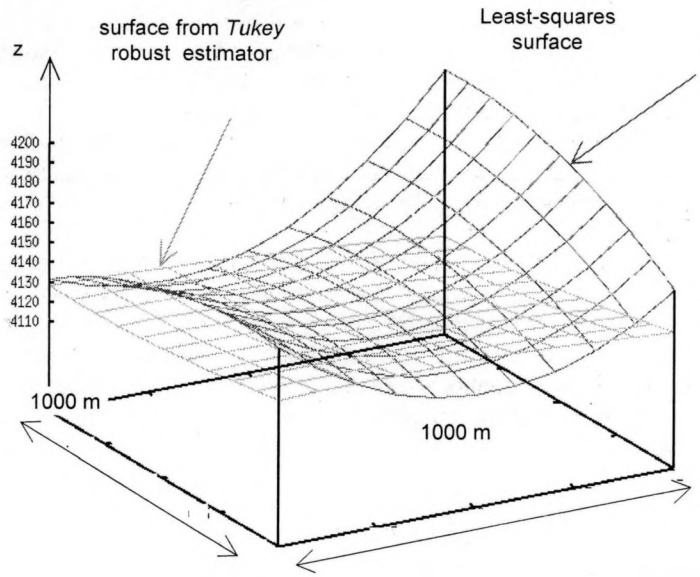

FIG. 2-1 (a).- Comparison of a least-squares surface with one obtained by a robust estimation. The *Tukey* robust estimator was used.

To be able to determine the parameters of the model, the use of a robust estimation procedure is required. Unlike a least squares estimation, a robust procedure does not take into account all of the soundings (*Cf.* Fig. 2-1 (b)).

As described in figure Fig. 2-1 (b), erroneous soundings located far away from the surface defined by validated ones, will consequently possess high residual values with respect to the robust fit (ROUSSEEUW, 1987).There must be less than 50% of erroneous soundings in the initial dataset, to superimpose a robust estimator surface over the valid soundings.

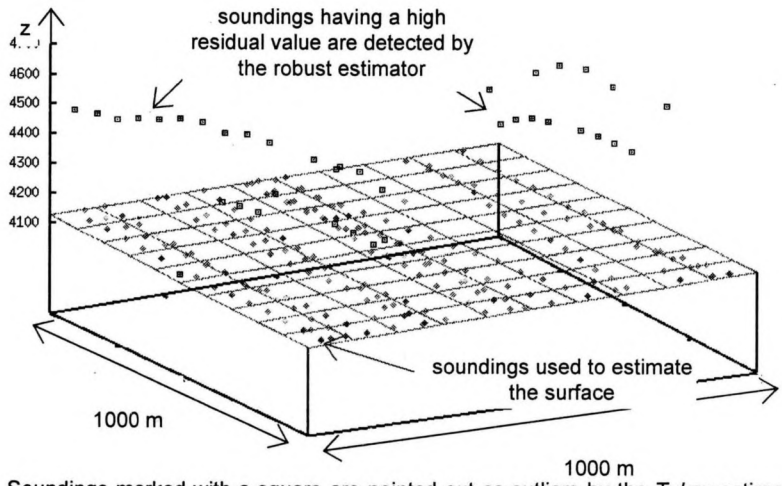

FIG. 2-1 (b).- Soundings marked with a square are pointed out as outliers by the *Tukey* estimator, those with a diamond are used to estimate the parameters of the model. The presented area is a squared cell of 1000 meters wide, that contains approximately 290 soundings (with almost 15% erroneous soundings).

2.2 *Tukey* robust estimator
Main features

The robust estimators which are straightforward to implement are the W-*estimators* (ROUSSEEUW 1987), also called the *IRLS-Estimators* (*as Iterative Reweighted Least Squares*). Their iterative construction scheme is based on the generalized least squares technique. The first step allocates the same weight to each point. The residual values provide the information to compute the sounding weights that will be used in the following step.

$$\hat{A}^{(j)} = \arg \min_{A} \sum_{i} w_i^{(j-1)} r_i^{(j-1)2}$$

with (3)

$$r_i^{(j-1)} = \left| z_i - \hat{A}^{(j-1)} X_i \right|^2$$

A given *W-estimator* is associated with a mathematical function called the influence function of the estimator (HAMPEL, 1986). This function allocates a weight to each point which depends on its residual value point. As the aim of this paper is not to design a robust estimator (HAMPEL, 1986), we will simply point out that *W-estimators* are classified into three classes according to the behaviour of their influence function. Common W-estimators have a descending influence function, which is strictly positive. In other words, from one step to the next, there are no rejected points. The *Tukey* robust estimator is a particular case as regards *W-estimators* (SOMOGYL, 1996). Its influence function, which is one of the two main features of the estimator, can reject soundings from one step to the next. Soundings having a weight equal to zero at the last step are detected as possible erroneous soundings. Secondly, we chose the *Tukey* robust estimator because of its adaptative capabilities (HUANG, 1995). The rejected threshold of the soundings changes from one step to the next one (4). It depends linearly on the median value

$r_{\text{median}}^{(j-1)}$ computed over all the absolute residuals and on α the inverse-sensitivity factor of the estimator.

$$\begin{cases} w_i^{(j)} = \left(1 - \left(\dfrac{r_i^{(j-1)}}{\alpha \cdot r_{\text{median}}^{(j-1)}}\right)^2\right)^2 & \text{if} \quad r_i^{(j-1)} < r_{\text{median}}^{(j-1)} \\ w_i^{(j)} = 0 & \text{otherwise} \end{cases} \tag{4}$$

In our case, it is precisely this feature that will be used to distinguish punctual-error noise from **MBES** noise.

2.3 Parameters of the algorithm

The proposed algorithm simply requires the setting of two parameters, which are:

- The *inverse-sensitivity* of the estimator : α
- The *size* of the cells (*i.e.* area) : *L*

The *inverse-sensitivity factor* of the estimator, α, is a parameter used to compute the rejection threshold. As mentioned in paragraph 2.2, soundings having an absolute residual value α times greater than the median value of the absolute residuals are not taken into account to estimate the surface at the next step : these are considered as the current outliers set.

The size of the cells *L* (*cf.* 2.21) has to be chosen so as to statistically fit a quadratic surface over each cell.

The *Tukey* robust estimator behaviour was evaluated using an artificial data set obtained by adding a white noise to constant depth values. This test clearly demonstrates that in the presence of a white noise, which is supposed to represent the MBES noise, even if the size of the cells is correctly set - as is the case in a featureless seabed - soundings are still extracted from the dataset as potential outliers, mainly because of the adaptative qualities of the estimator.

Consequently, even in the case of an optimal setting of the cell size *L*, the soundings of the zero-centered mode of the histogram computed by the residual values of the detected soundings are not rejected as punctual errors. This is an inevitable part inherent to the adaptative behaviour of the detection process. It is therefore essential to introduce, independently of the process, a global parameter defining the minimal residual value of the punctual erroneous soundings. This threshold can be defined, in a more accurate way, from the intrinsic characteristics (*i.e.* resolution) of the **MBES**.

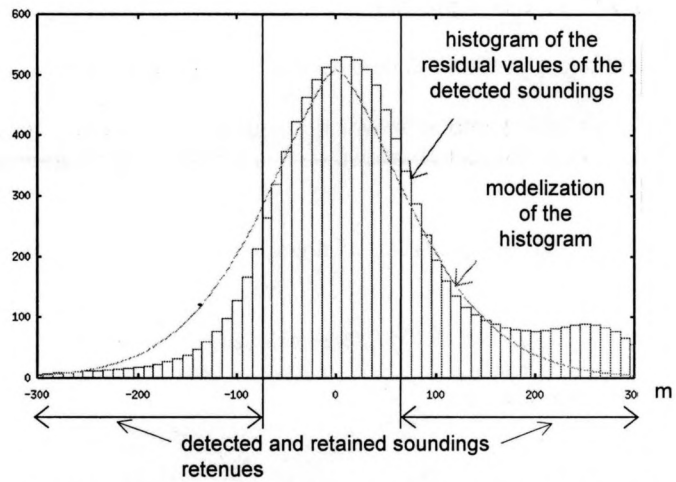

FIG. 2-3 (a).- The automatic detection of the zero-centered mode of the histogram of the residual value of the detected soundings is used to extract the erroneous soundings from the potential ones. In the example above, the lobe width is fixed at 2/3 of its height.

As illustrated in figure Fig. 2-3 (a), it is also possible to accurately fit this threshold, by visualizing the histogram of the residual values of the detected soundings. Nevertheless, this can only be achieved if the sampling interval of the histogram is correctly chosen (*cf.* FIG - 2-3 (b)). This step must be great enough compared to the intrinsic uncertainty of the sounder in order to obtain a zero-centered histogram mode.

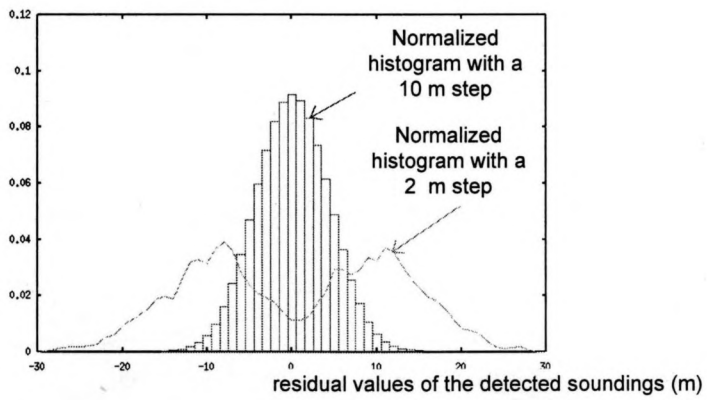

FIG. 2-3 (b).- Normalized histograms of the residual values of the detected soundings, with 10 and 2 meter steps, are obtained in the case of an artificial dataset. This dataset was built by adding a white noise N(0, σ^2) to constant depth values.

If the trained operator requests it, this threshold could also be added, in an automated way, as an option, from a modelization of the central histogram (such a modelization is shown by the dotted line in figure Fig. 2-3 (a).

2.4 Running modes of the algorithm

The described algorithm has two running modes, which are:

- the *fast running* mode, in which a sounding is observed once only,
- the *cover* mode, which allows each sounding to be tested several times.

The fast running mode consists of a sequential and separated scan of the *LxL* adjacent cells.

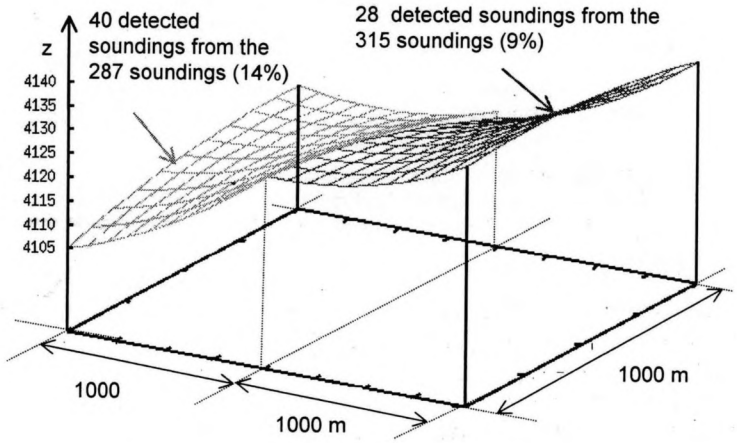

FIG. 2-4 (a).- Fast running mode algorithm.

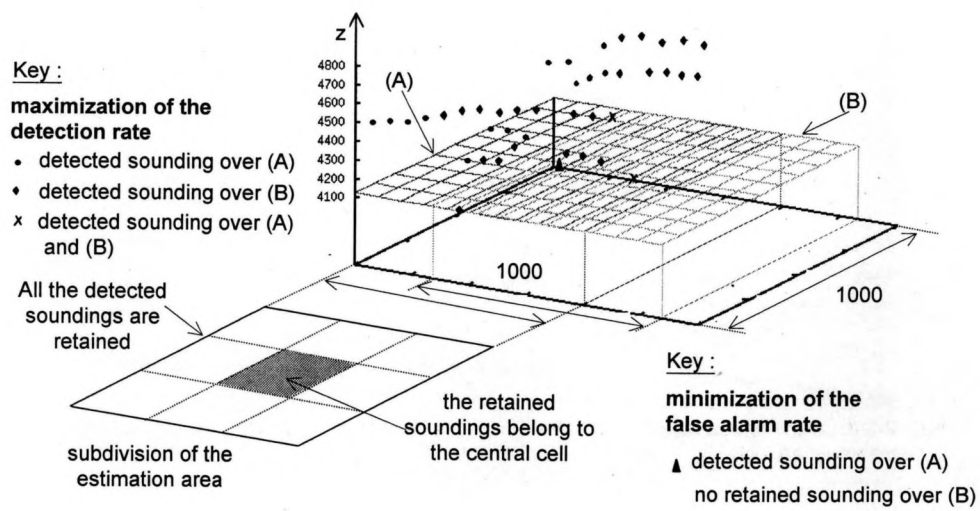

FIG. 2-4 (b).- Cover modes algorithm.

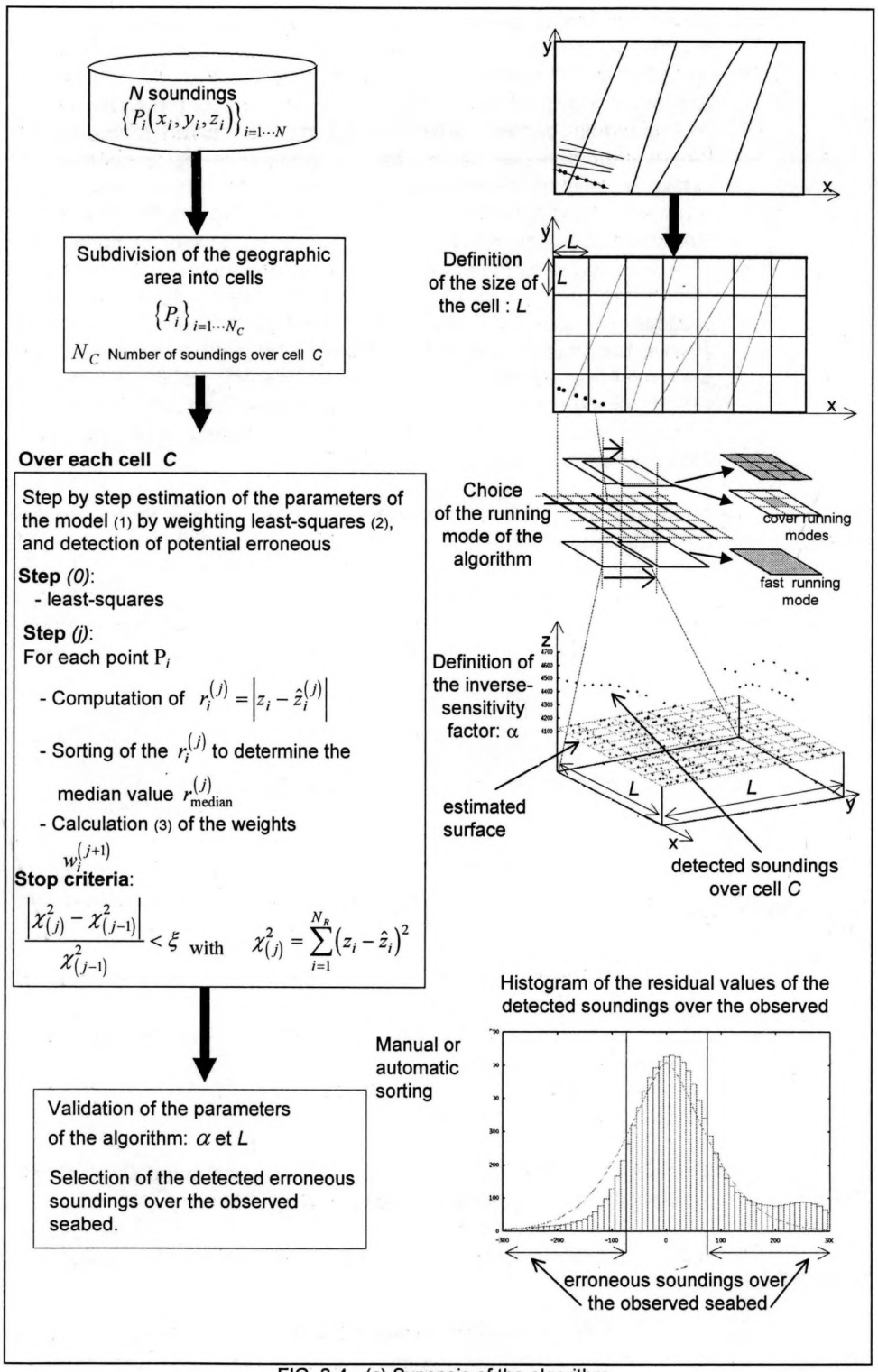

FIG. 2-4.- (c) Synopsis of the algorithm.

- In the cover mode, the succession of the *LxL* cells is drawn up so that they partially overlap (*i.e.* 2/3 along x-axis and 2/3 along y-axis). Each sounding is observed several times, each time in a different neighbourhood. This mode partially compensates for the weakness of the quadratic model when local seabed shapes may have necessitated a higher order model. Accordingly, the probability of finding a seabed shape which includes most of the soundings increases, in other words, the probability of fitting (to the seabed) a local quadratic surface. Consequently, the probability of finding an erroneous sounding in bumpy subareas increases too. Moreover, the cover mode gives a score to each possible outlier. This score is used to convey the doubtful feature of a sounding. The greatest score is given to the soundings observed N times and pointed out N times as outliers by the estimator. In practice, the implementation of this idea is to subdivide the observed area. Each cell, whose size defines the estimated area, is created by bringing together nine sub-cells obtained from a thinner grid (*Cf.* FIG- 2-4 (b)).

Several alternatives of the *cover* mode are proposed. They depend on the weighting scheme assigned to each cell. By taking into account only the outliers of the central cell, the false alarm rate is minimised (*Cf.* definition in 3) : only the soundings for which an isotropic distribution of the information is available are retained. In contrast by taking into account all the soundings detected over the nine cells, the detection rate is maximized because the observed erroneous soundings increase.

3. DESCRIPTION OF THE DATASETS

The algorithm was evaluated on five swaths (*i.e.* set of bathymetric data acquired along a vessel course) selected in a reference set of data acquired with three different **MBES**:

- deep water (200m - 12000m) *SIMRAD EM12-dual* : with 162 beams and an opening angle of 128°,
- shallow water (5m - 300m) *Thomson-Lennermor* with 16 beams and an opening angle of 75°,
- very shallow water (0 - 150m) *SIMRAD EM3000* : with 127 beams and an opening angle of 140°.

These data sets of soundings were previously cleaned manually to obtain the reference data set : all soundings were systematically visualized.

Figure 3 represents the error rates versus the beam index, for each of the three MBES.

Concerning the *EM12-dual* and *EM3000* swaths (Cf. *Fig. 5*), errors mainly belong to the central beams where the detection mode changes from amplitude to phase detection.

Table 3
Description of the swaths of the reference data set obtained from manual cleaning.

	MBES	Period	Characteristics of the observed seabed	Depth range (meters)	Number of soundings	Number of errors - manual data cleaning -
1	EM12-dual	4h 40 mn	Abyssal plain	3500 to 700	108 257	768 (101 doubtful)
2	EM12-dual	1h 10 mn	Sea mount	2400 to 3900	46 592	142 (39 doubtful)
3	EM12-dual	1h 40 mn	Undulated shape	1800 to 3000	88 574	410 (238 doubtful)
4	Lennermor	5mn 20s	Sand dune	30 to 35	38 764	327
5	EM 3000	1mn 52s	Levee and channel	3 to 9	178 287	774

For the *Lennermor* MBES, the error rate is, on the contrary, at its maximum for the lateral beams. Automatic and manual detection were carried out over a set of five parallel swaths with overlapping areas, due to the undersampled data across the swath.

The evaluation of the algorithm is based on the following criteria:

- the *detection rate* is defined as the ratio of the number of detected erroneous soundings to the total number of erroneous soundings.
- the *false alarm* rate is defined as the ratio of the number of valid detected soundings to the total number of detected soundings.

These definitions are relative to the results of the manual data cleaning which is considered as the reference.

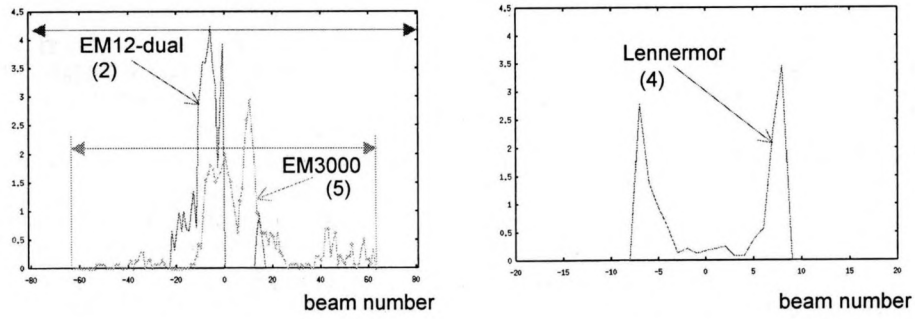

FIG - 3 Error rates *versus* beam number.
Numbers in brackets refer to swath number used to compute these error rates.

4. EVALUATION OF THE ALGORITHM

4.1 Parameters fine tuning

The optimum size of the cells (*i.e.* the maximum scale for which the seabed can be modelized with a quadratic surface) was fixed for each of the data sets. Figure 4-1(a) shows that a 1000 meter size cell for swath 2 is statistically correct. In fact, for this size, the histogram mode of the residual values of the detected soundings is approximately zero-centered and sharper than the others. In practice, the trained operator can take into account the diversity of the seabeds tested in this study, presented in table 4, to set the *a priori* size of the cells.

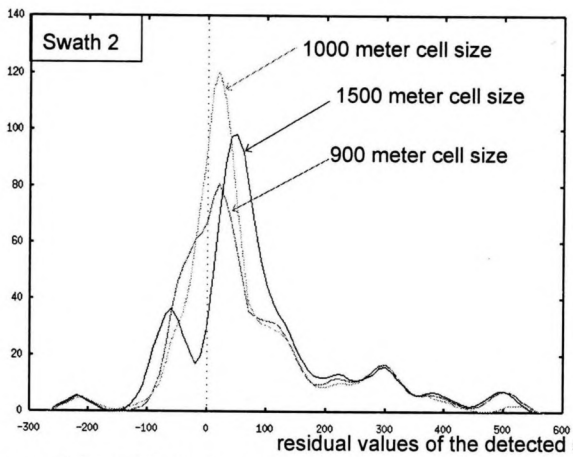

FIG. 4-1(a).- Histogram of the residual values of the soundings detected by the estimator for several cell sizes.

As a visual analysis of the histogram allows *a posteriori* check and validation of the cell size, a wider choice of seabed types than the one presented in this paper, can be envisaged. Figure 4-1(a) represents the results of the shape of the histogram mode of an unadapted cell size. A smaller cell size increases the probability of retaining gaussian noise errors. In contrast, a too large cell size has a smoothing effect on the reliefs (*i.e.* clipping the peaks and partially filling in the valleys). A subdivision of the central mode of the histogram shows an inadequate mesh. Nevertheless, it is important to emphasize that this heuristic criterion is useless when the MBES is the *Lennermor* because of data undersampling along the transversal axis.

The second parameter of the algorithm is the inverse-sensitivity factor of the *Tukey* estimator. Figure 4-1(b) shows the detection rate *versus* the false alarm rate. The three represented curves, one per running mode, were obtained for different values of the inverse-sensitivity factor, α, (*i.e.* α varies from 6 to 14 with a step of 2). The lower the inverse-sensitivity factor, the higher the false alarm (*i.e.* the cleaning process becomes sensitive to small relative perturbation). In practice, the value of the inverse-sensitivity is between 6 (for shallow water) and 10 (for deeper water).

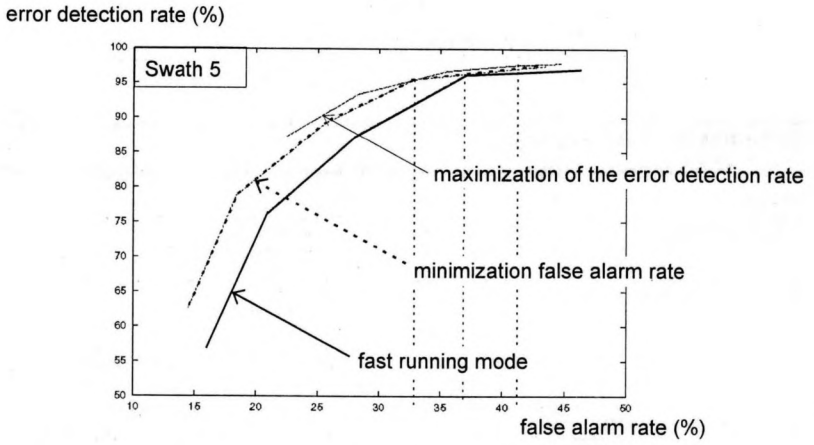

FIG. 4-1(b).- Detection rate *versus* false alarm rate. Vertical dotted lines correspond to results obtained from swath 5 (*i.e.* EM3000 data), numerical values are given in table 4.

The third parameter is a global threshold applied to the magnitude error of the detected soundings.

4.2 Results

Table 4 shows the algorithm detection rates for the five swaths of the reference dataset. Thresholds applied for the residual values of the detected soundings were automatically determined from the central-mode of each histogram (*i.e.* the width is set to 2/3 of the height of the lobe); thresholds obtained are nearly equivalent to those deduced by visualizing the lobe of the histogram mode.

As regards the EM3000 and EM12 MBES, the choice of the running mode only depends on imposed operator constraints.

For swaths 1 to 3, the use of the cover mode increases the detection rate by 5%, by keeping a false alarm rate of approximately 25 % (except for swath 3). The data set of swath 1 contains 108 000 soundings which were acquired by the EM12 over a period of four hours. The algorithm only takes 2mn 25s (on a Sun Ultra-sparc Station) to detect 93% of the errors of this swath in the cover running mode.

For swath 5, the detection rates are the same. In the fast running mode, our algorithm takes less than 30s (on a Sun Ultra-sparc Station) to detect 96 % of the errors included in this data set of approximately 178 000 soundings.

However, for the *Lennermor* MBES it is crucial to use the cover mode to obtain 88% of the errors. In this way, we compensate for the spatial disparity of the data.

Table 4
Detection rates of the algorithm

Running mode	Detecttion rates (%) errors	false alarms
Swath 1 : EM12 - Abyssal plain cell size : 1000 meters - inverse-sensitivity factor : 10 - magnitude threshold : 50 meters		
Fast running mode	**88.1**	**21.3**
Max. of the error rate	93.3	25.6
Min. of the alarm rate	87.8	11.6
Swath 2 : EM12 - Sea mount cell size : 1000 meters - inverse-sensitivity factor : 10 - magnitude threshold : 50 meters		
Fast running mode	**81.6**	**19.5**
Max. of the error rate	98.1	25.7
Min. of the alarm rate	94.2	8.8
Swath 3 : EM12 - Undulated shape cell size : 600 meters - inverse-sensitivity factor : 10 - magnitude threshold : 40 meters		
Fast running mode	**86**	**58.2**
Max. of the error rate	91.9	51.9
Min. of the alarm rate	86.1	28
Swath 4 : Lennermor - Sand dune cell size : 10 meters - inverse-sensitivity factor : 6 - magnitude threshold : 3 meters		
Fast running mode	**12.5**	**29.3**
Max. of the error rate	88.1	29.4
Min. of the alarm rate	7.6	10.7
Swath 5 : EM3000 - Levee and channel cell size : 2 meters - inverse-sensitivity factor : 8 - magnitude threshold : 0.25 meters		
Fast running mode	**96.1**	**36.9**
Max. of the error rate	97.6	40.8
Min. of the alarm rate	95.5	32.8

5. CONCLUSION

The proposed algorithm is based on the *Tukey* robust estimator to detect erroneous soundings in bathymetric data.

Its advantage is that it only needs two parameter settings, one of which, the cell size, can be controlled *a posteriori*.

The second advantage is its different running modes. With the fast running mode, data from shallow water MBES are cleaned four times faster than the acquisition. With the cover running mode, it is able to detect a 98 % error rate with a 25% false alarm rate. While minimizing the false alarm rate, more than 94 % of

errors present in deep water data are detected with less than a 10 % false alarm rate. Finally, data from the Lennermor MBES can be cleaned with a 88% error detection rate.

Consequently, it has been recently adopted and used in the SHOM post-processing software.

Acknowledgements

Special thanks to Andre GAUDIN from the Canadian Hydrographic Service (Laurentian region), who has provided us with *EM3000* data.

References

DEBESE N., Typologie des erreurs les plus courantes du sondeur EM12-dual. *Rapport d'étude du SHOM* n° 007/97 (1997).

DEBESE N., Application d'un estimateur robuste à la détection des erreurs ponctuelles présentes dans les données bathymétriques multifaisceaux : l'estimateur de Tukey. *Rapport d'étude du SHOM* n° 002/98 (1998).

DEBESE N., MEVEL Ch., FREULON X., Application d'un estimateur robuste à la détection des erreurs ponctuelles dans les données bathymétriques multifaisceaux. *Conférence Hydrographique du Canada*, Victoria 10 -12 mars (1998).

GAUDIN A., The calibration of Shallow Water Multibeam Echo-sounding Systems. *Conférence Hydrographique du Canada*, Halifax 3-5 juin (1996).

HAMPEL F. R, RONCHETTI E. M, ROUSSEEUW P. J, STAHEL W. A., Robust Statistics : The Approach Based on Influence Functions. *Wiley Series in Probability* (1986).

YAN HUANG, PALANIAPPAN KANNAPPAN, ZHUANG XINHUA, CAVANAUGH J. E., Optic Flow Field Segmentation and Motion Estimation Using a Robust Genetic Partitioning Algorithm. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol 17, n°12, pp. 1177-1190 (1995).

ROUSSEEUW Peter J, LEROY Annick M., Robust Regression and Outlier Detection. *Wiley series in probability* (1987).

SOMOGYL J., ZAVOTI J., A Comparison of Weight-functions in Robust Regression using Iteratively Reweighted Least-Squares. *Acta. Geod. Geoph.*, Vol 31(1-2), pp. 11-24 (1996).