International Hydrographic Review, Monaco, LXXII(1), March 1995

ON THE IDENTIFICATION OF SPIKES IN SOUNDINGS

by Jørgen EEG [1]

"For the Snark's a peculiar creature, that wo'n't
Be caught in a commonplace way.
Do all that you know, and try all that you don't:
Not a chance must be wasted to-day!"

Lewis Carrol: the Hunting of the Snark

Abstract

This paper points out a method of finding blunders in sounding measurements. By postulating that the sea bed enjoys a certain approximation property, inconsistencies in data can be found by a test which allows of an inspection of problem areas in decreasing order of interest. Finally a means to decide if the postulate is in error is presented.

Introduction

This paper describes a method which has been employed at the Royal Danish Administration of Navigation and Hydrography for over a year as an important part of the quality control of soundings. Although the method originally was devised with the multibeam echosounder in mind, it was quickly adopted for use with single beam surveys and, later, used to provide quality control during the digitizing of fair sheets. In both cases, ground truth (echogram and fair sheet) is available and provides an accurate means to verify the measurements by manual inspection. However,this kind of work is tedious and time consuming and therefore prone to errors. The solution which is suggested here is to inspect the potential

[1] Royal Danish Administration of Navigation and Hydrography, P.O.Box 1919, Overgaden O.Vandet 62B, DK 1023 Copenhagen K.

blunders in decreasing order of interest. This is made possible by introducing an ordering into the set of observations so that observation z_i preceeds z_j if and only if z_i is at least as likely as z_j to be a spike.

For the moment, suppose that such a method is available,then how can a quality control procedure benefit from it?

First, it makes it possible to check-up the observations to a preselected level, and thereby achieve a homogeneous quality control for different surveys. For example define the probability of finding a spike as the ratio between the number of observations which are verified as blunders to the total number of inspected observations. As the inspection proceeds this ratio changes, and when the probability of finding a spike drops below a certain limit, the procedure stops.

Second, as the probability that the observation z_i is a spike decreases, it becomes to a greater degree verified by the other observations. This relation may be expressed as a quality measure which, for example, may follow the observation into the database.

It is also worth mentioning, that a simple extension of the method would be to incorporate z_i and z_j as sets of neighbouring measurements, thus allowing a search for features on the sea bed as wrecks etc...

The author wishes to acknowledge the lively interest of the surveyors in the Royal Danish Administration of Navigation and Hydrography in testing this method.

The Model

Surveying with a multibeam echosounder system is a complicated task in which the final result depends on output from several sensors, each having its own caracteristics and therefore prone to errors which are special for its type and which consequently should be modeled separately.

On the other hand, the end product of a survey which is a set of triples

$$(x_i, y_i, z_i) \qquad i = 1, \ldots, n \tag{1.1}$$

of positions (x_i, y_i) and corresponding depths z_i, also offers some means of control which again may give rise to a modeling of the errors in the sensors. For instance, overlapping tracks which are traversed in opposite directions may supply information about a time lag in the system.

Below we shall investigate the special case where we take the positions

$$(x_i, y_i) \qquad i = 1, \ldots, n$$

for granted, and regard the corresponding depths z_i as measurements of a real valued function

$$f : \mathbf{R}^2 \rightarrow \mathbf{R}$$

at the position (x_i, y_i),

$$z_i \; (=) \; f(x_i, y_i) \quad i = 1,2,\ldots,n \tag{1.2}$$

where the parentheses on both sides of the equality sign indicate the errors which are inherent in the measuring process.

Of course, for any set of values of (1.1) we have an infinity of functions g,

$$z_i \; = g(x_i, y_i) \quad i = 1,2,\ldots,n$$

which are continuously differentiable to any degree and interpolate (1.1). This again is a subset of the set of functions which approximate f, but its members may, at least locally, not be good candidates for a representation of the bottom due to gross errors in z_i. Whether or not such a conclusion can be drawn from a given set of triples (1.1) depends on additional statements about the function f which is to be approximated and on the distribution and density of the points (x_i, y_i).

Below is investigated some consequences for the set (1.1) which may be drawn, when the function f enjoys the following property:

For the given sample the function f can locally be approximated homogeneously by a polynomial of fixed order in x and y.

For the set (1.1) of triples this means, that for any measurement z with neighbours z_k, $k = 1,2,\ldots,m$, the effect on the approximation of leaving out z from a least squares fit of a surface of this fixed order should be of the same order of magnitude as the average effect of leaving out any one of its neighbours.

For any function which enjoys such a property it is possible to check a set of triples (1.1), once the order of the polynomial in x and y is fixed and the definition of which measurements z_k we shall denote neighbours of z is chosen: *For every measurement z the quotient q(z) of the above effects is calculated, and this quotient introduces an ordering in a subsequent check-up of the observations for gross errors.*

THE *SPIKES* ALGORITHM

As an example I shall in some detail work through the *spikes* detection algorithm as it is currently implemented for multibeam measurements at the Royal Danish Administration of Navigation and Hydrography.

In the shallow Danish waters the density of measurements is very high, mostly there is less than 2 metres between neighbouring measurements in a track, and as the seabed furthermore is rather smooth it is reasonable to postulate that it, at this sampling density, can be approximated homogeneously by a polynomial of order zero in x and y, i.e. by a constant.

Let z have the neighbours z_k, $k = 1,2,\ldots,m$, then, applying a least squares adjustment with equal weights, the least squares estimate of the above mentioned constant is the average of the observations

$$\bar{z} = \frac{1}{m+1}(z + \sum_{k=1}^{m} z_k) \tag{2.1}$$

and the least squares fit of this model is the square sum *SSD* of deviations from the average,

$$SSD = (z-\bar{z})^2 + \sum_{k=1}^{m} (z_k - \bar{z})^2$$

If z is left out of this adjustment, one gets a new estimate $\hat{z}$ of the sea bed

$$\hat{z} = \frac{1}{m} \sum_{k=1}^{m} z_k \tag{2.2}$$

and of the least squares fit *ssd*

$$ssd = \sum_{k=1}^{m} (z_k - \hat{z})^2$$

According to the postulate, the difference

$$SSD - ssd = \frac{(z - \bar{z})^2}{\frac{m}{m+1}} \tag{2.3}$$

and the mean s^2 of *ssd*

$$s^2 = \frac{1}{m-1} \sum_{k=1}^{m} (z_k - \hat{z})^2$$

should be of the same order of magnitude for the neighbourhood of each measurement in the swath. Therefore it follows, that the larger the quotient

$$q(z) = \frac{(z - \bar{z})^2}{\frac{m}{m+1} \cdot \frac{1}{m-1} \sum_{k=1}^{m} (z_k - \hat{z})^2} \tag{2.4}$$

is, the more z separates from its neighbours, or, put in another way, as the quotient decreases the observation becomes to a greater degree verified by its neighbours. This fact makes rational spike detection feasible: *For each observation in the track the* *corresponding quotient is calculated. Using the size of the quotient as a key the problem areas of the track can be inspected in increasing order of interest.*

The choice of neighbours to a given observation still has to be accounted for. For the moment, regard the positions of the measurements in the track

$$(x_i, y_i) \qquad i = 1,2,\ldots,n \tag{2.5}$$

as points in $\mathbb{R}^2$. Then for any point (x_j, y_j), the pair

$$(x_j, y_j) \text{ and } (x_i, y_i) \qquad i = 1,..,j\text{-}1,j+1,..,n$$

splits $\mathbb{R}^2$ into two halfplanes, each consisting of members of $\mathbb{R}^2$ which are closer to one of the points than to the other. The intersection of these halfplanes as the index i runs through the set $1,..,j\text{-}1,j+1,..n$ forms a polygon P_j around (x_j, y_j) consisting of the elements of $\mathbb{R}^2$ which are closer to (x_j, y_j) than to any of the other measurements in the track.

The polygons P_j, $i = 1,\ldots,n$ form the Voronoi diagram of (2.5). As neighbours to the point (x_j, y_j) I choose those points (x_k, y_k) in the track for which the corresponding polygons P_k share a side with P_j. It is well known, that if points defined as neighbours in this way are connected by line segments, then one gets the dual of the Voronoi diagram namely the Delauney triangulation of the point set. Due to its importance for applied interpolation in $\mathbb{R}^2$ the Delauney triangulation has been treated extensively in the litterature. I shall only give one reference here [1], which deals with the aspect of finding an efficient algorithm to produce a Delauney triangulation, and otherwise refer to the Internet where the source text is available in public domain.

Using (2.1) and (2.2) one gets

$$\begin{aligned}(m + 1)(z - \bar{z}) &= (m + 1)z - z - \sum_{k=1}^{m} z_k \\ &= m(z - \hat{z})\end{aligned}$$

so that (2.4) can be written

$$q(z) = \frac{(z - \hat{z})^2}{\dfrac{m+1}{m} \cdot \dfrac{1}{m-1} \displaystyle\sum_{k=1}^{m} (z_k - \hat{z})^2}$$

which shows, that the quotient at z (apart from the number) depends on the variance of the neighbours

$$\frac{1}{m-1} \sum_{k=1}^{m} (z_k - \hat{z})^2$$

and their ability to interpolate z,

$$(z - \hat{z})^2$$

So, when the set of neighbours to *z* is chosen, it is important that the choice reflects the variation of the seabed at the position where *z* is measured and that it consists of measurements which surround *z*.

For the spikes detection algorithm this means that measurements located on the boundary of the convex hull of the track, and measurements which are more than a fixed distance from any one of its neighbours are not checked by the algorithm, but flagged as boundary measurements.

Also notice, that the quotient is invariant with respect to a joint change of scale for *z* and its neighbours. Depending on the accuracy demands on the survey, a constant *c* is chosen, so that only areas of the track for which

$$(z - \hat{z})^2 > c$$

are inspected.

Below an extract of the ASCII output from the *spikes* algorithm is depicted. The track contained in this case about 65000 measurements which on a HP720 were processed in less than 30 seconds. The output can be read as follows: Every prospective spike has associated two lines of information. On the first line is placed facts about the suspected observation: quotient, depth, easting, northing and beam angle, while the second line contains information on the neighbours, as an example: 6.48,0 means that the observation 7.80 m has a neighbour of 6.48 m for which the maximum distance in northing or easting is between 0 and 1 metre.

```
quot. depth   easting    northing   date      time       transducer angle
 116   7.80  731543.02  6172844.99  17/10/92  13:22:10.47       12.4
               6.48,0  6.49,0  6.43,1  6.29,0  6.21,1  6.18,1
 106   4.73  731716.44  6172335.66  17/10/92  13:19:29.01      -60.3
               6.44,0  6.59,0  6.53,1  6.46,1  6.20,0  6.65,1
  91  12.00  730887.02  6174941.15  17/10/92  13:31:22.99      -34.6
              13.26,2  13.14,1  13.32,2  13.26,1  13.41,0
  66  14.48  730819.60  6175077.53  17/10/92  13:32:16.57       53.3
              13.44,2  13.40,0  13.43,2  13.41,2  13.26,2
  57  12.32  730872.57  6174910.78  17/10/92  13:31:14.82       57.0
              13.38,2  13.30,2  13.33,0  13.24,2  13.41,1
 etc...
```

Other outputs from the programme are:

- a copy of the input where the observations are flagged according to whether they are verified by the algorithm, boundary points or prospective spikes. Furthermore the value of the quotient and the number of neighbours are registered here for use as a quality measure in the database.

- a file containing a 25 x 25 m extracted neighbourhood of each prospective spike for evaluation by hydrographer.

- a file containing information to draw a circle around each of the prospective spikes.

The last two files are used as input to AlliedSignal-Elac's postprocessing system, but it is planned to develop an interactive graphical interface which allows the user to acces the information which is necessary in order to decide the fate of the prospective spikes, that is whether to

- remeasure for confirmation

- flag as a spike

- flag as not a spike, confirmed by hydrographer.

A STATISTICAL INTERPRETATION

Above I have investigated some consequences which may be drawn for the behaviour of the measurements of a function f which enjoys the property that it locally can be approximated homogeneously by a polynomial of some fixed order in x and y.

Once the order of the polynomial is fixed and the definition of the local term for a given approximation problem is chosen, the question arises whether there is a means to ascertain when these assumptions break down.

Assume that, for the neighbourhood of a measurement z , f really is a polynomial of the said order in x and y, and that the measurements $z, z_1, z_2, \dots, z_m$ are independent with errors which follow a normal distribution with zero mean and variance σ^2.

Then it is proven in [2] that (2.3) may be regarded as the Pythagorean theorem for a right angled triangle in the space of observations, making *ssd* independent of *SSD - ssd*, and each of these expressions follows a chi-square distribution with m-1 and 1 degree of freedom respectively. The quotient (2.4)

$$q(z) = \frac{SSD - ssd}{\frac{1}{m - 1} \cdot ssd} \tag{3.1}$$

then is (Fisher) $\upsilon^2(1,m\text{-}1)$ distributed, as the numerator and the denominator are independent estimates of the variance σ^2. Now, the quotient is independent of σ^2, so if the value of the quotient is calculated for several disjoint sets of observations, each of which consists of an observation with m neighbours, then, even though each set may be supposed to have its own variance, the quotients can be pooled together and compared to a $\upsilon^2(1,m\text{-}1)$ distribution.

With one minor modification I shall propose to use the above procedure to control if the assumptions about the function f are wrong. It is well known, that the square of the t-distribution with m-1 degrees of freedom is a $\upsilon^2(1,m-1)$ distribution. So, instead of using (3.1) one ought to use its signed square root, which sometimes in statistical litterature is called the jack-knife residual or the externally studentized residual. This quantity measures the effect on prediction which stems from deleting an observation, see [3].

When the polynomial is a constant this means that we, instead of using (2.4), compare the distribution of

$$r(z) = \frac{z - \bar{z}}{\sqrt{\frac{m}{m+1}} \cdot \sqrt{\frac{1}{m-1} \sum_{k=1}^{m} (z_k - \bar{z})^2}} \qquad (3.2)$$

to a t-distribution with m-1 degrees of freedom. In doing so we take advantage of the symmetry of the t-distribution to sharpen the comparison, which in a case like this, usually is carried out by employing a chi-square test. However, when evaluating the outcome of the test it has to be taken into account, that the assumption about the distribution of the errors is rather arbitrary. What still makes it worthwhile to consider the statistical angle of the approximation is, that for the small values of the degrees of freedom that we encounter in multibeam surveys (usually less than seven when we use a zero degree polynomial and a Delauney triangulation), the t-distribution of (3.2) is robust in the sense, that it essentially depends on symmetry and finite variance of the distribution of the errors of the measurements. While we, for any measurement of the sea bed, are able to guarantee the latter, the former may be sensitive to the degree of the polynomial or to systematic errors, for example in the outermost beams.

Conclusion

Above I have treated the problem of finding isolated gross errors in multibeam measurements. It is important to bear in mind, that the deviation in the numerator of the quotient (2.4) is normalized by the denominator. The effect of this is, that when one of the neighbouring measurements catches the build up of a structure, the quotient becomes small. The price one pays is, of course, that one does not find gross errors which are neighbours using the above model. From the expression (3.1) for the quotient it follows, that it is straightforward to generalize the above procedure to leave out two or more observations. As this number grows, the number of combinations and therefore the computing time makes it impractical to proceed along this line. Some small modification may, however, give a usable result. For example, in the variance plot which is widely used in post processing systems for multibeam measurements, it would be simple to use (3.1) combined with a display to indicate where the disposal of 2, 3 or more observations in a cell reduces the variance of the remaining observations drastically, and therefore makes a further inspection of the area necessary.

References

[1] Preparata & Shamos: Computational Geometry. Springer Verlag 1985.

[2] J. EEG: On the Adjustment of Observations in the Presence of Blunders. Geodætisk Institut, Technical Report no.1, 1986.

[3] A.C. ATKINSON: Plots, Transformations, and Regression. Oxford Statistical Science Series, Clarendon Press, 1985.