

REPRODUCIBILITY OF SINGLE-BEAM ACOUSTIC SEABED CLASSIFICATION FOR HABITAT MAPPING

Art Gleason^a, Jon Preston^b, Steve Bloomer^c

^aRosenstiel School of Marine and Atmospheric Science, Univ. Miami, Miami, FL, USA

^bQuester Tangent Corporation, Saanichton, BC, Canada

^cCanadian Marine Acoustic Remote Sensing Facility, Univ. of Victoria, Victoria, BC, Canada

Contact author: Jon Preston, Quester Tangent Corporation, 6582 Bryn Road, Saanichton, BC, Canada V8M 1X6, Tel. 1 250 654 3316, E-mail jpreston@questertangent.com

Abstract: *Single-beam acoustic seabed classification continues to be a popular method for mapping seabeds and their sediments, which are important components of benthic habitat. Modern methods can generate maps of acoustic classes that are useful and reasonably accurate. Research toward improved methods continues. A continuing impediment to this research is ranking the accuracy of maps produced by new methods. Non-acoustic data, or ground truth, is usually sparse compared to the detail of the acoustic survey, which can mean that ranking maps for accuracy can be inconclusive. Here we present a new tool for ranking classification maps, namely the reproducibility of acoustic classes from repeated surveys of the same area on different days. Methods that have high reproducibility achieve that by capturing echo characteristics that are strongly influenced by seabed type while suppressing details that are driven by sea state or the water column. Six surveys, done with a 50 kHz sounder over a pair of transects near Miami, FL, USA, between 1 May and 13 August 2007, were used to evaluate two questions. First, how reproducible were classifications of this dataset using QTC IMPACT™ (Quester Tangent Corporation)? Second, can classification be improved with adjustments to the standard IMPACT processing? Reproducibility was quantified with the overall accuracy and the Kappa statistics, which are both derived from the confusion matrix whose rows and columns are numbers of sites with particular class assignments under distinct circumstances.*

Keywords: *acoustic classification, seabed classification, echo analysis, confusion matrix, reproducibility*

1. INTRODUCTION

A major challenge in developing methods for acoustic seabed classification is comparison of results from competing methods. Since the primary goal is accuracy, comparing acoustic classes with ground truth is the ultimate test. This may not be helpful in ranking the accuracy of class maps made by competing methods, though, because results from acoustic processes are often one hundred or more times more dense than the locations of ground truth. Acoustic class maps can be far more detailed than ground truth (indeed this is one motivation for acoustic seabed classification), so another method is needed to rank classification algorithms. One method is to compare borders between acoustic classes with detailed bathymetry. While this can be accurate for borders of rocky regions, it is not useful in discriminating clastic sediments. Other methods that can be of some value are examining consistency in overlap regions and homogeneity in regions expected to be homogeneous, but they are somewhat subjective.

This paper presents a new method for assessing and ranking maps of acoustic seabed classes, based on reproducibility among a series of surveys of the same area. Methods that have high reproducibility achieve that by capturing echo characteristics that are strongly influenced by seabed type while minimizing the effects of sea state and the water column.

2. SURVEY AND ACOUSTIC CLASSIFICATION

The survey site [1], Fowey Rocks, has geomorphology typical of the major Florida Keys reefs. Two parallel transects were selected, each about 2 km long with depths from 5 to 60 m. A small boat fitted with a Suzuki 2025 echo sounder was used to survey both transects three times on each of six dates: May 1, 2, 9, and 28, and August 3 and 13, 2007. Echo time series were acquired by a QTC5™ data acquisition system, which sampled the amplified echoes at 5 MHz followed by decimation and digital filtering. The sounder operated at 50 kHz and 500 W transmit power, with a rectangular beamwidth $42^\circ \times 16^\circ$ and a pulse length of 0.3 ms.

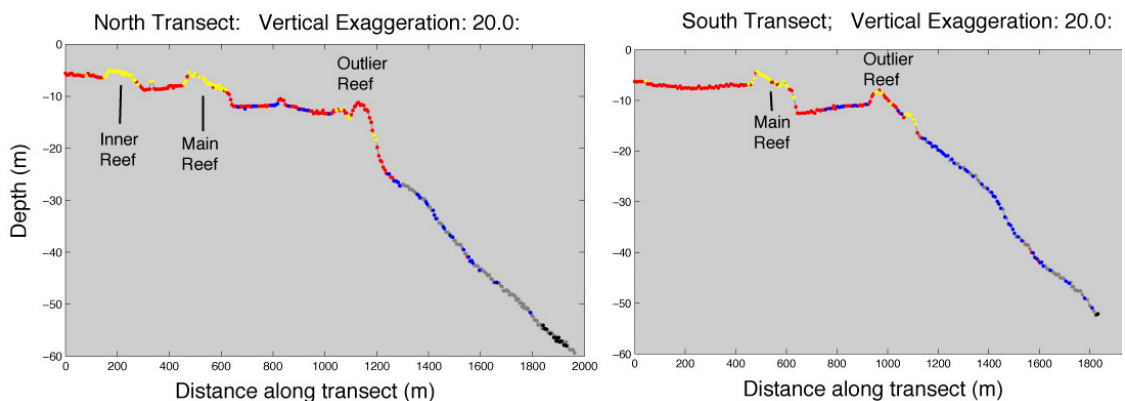


Fig.1: Vertical sections across the north (left) and south (right) transects showing the main bathymetric features, namely a series of linear reefs spaced by belts of relatively flat sediment in shallow water and sloping sediment in deeper water. Colours are acoustic classes from the 1 May data set.

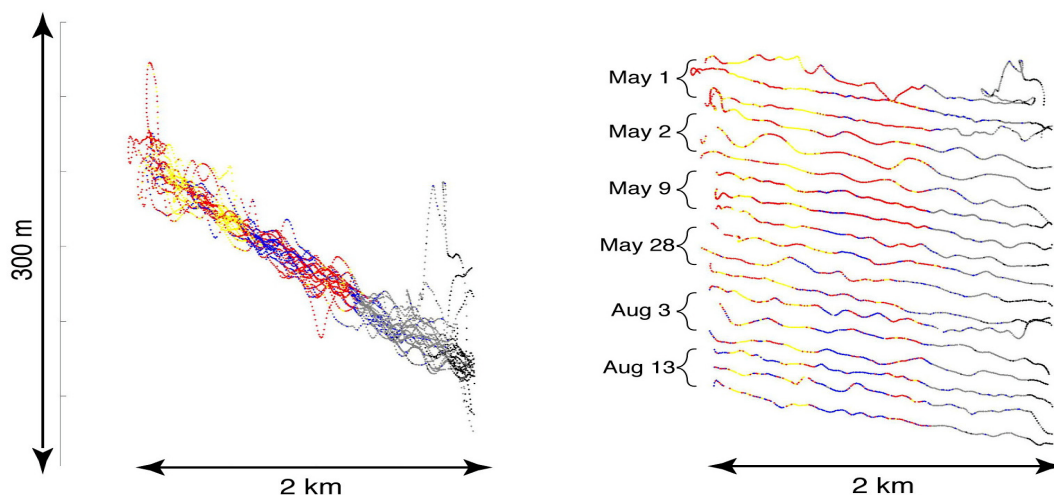


Fig. 2: Acoustic classes on the northern survey line at Fowey Rocks. The two lines were surveyed three times on each of the six days. Left plot shows the 18 northern tran-sects plotted together; right with 50 m artificial spacing between them Class colours (black, grey, blue, red, and yellow) correspond to positions of class centres in feature space.

Unsupervised acoustic classification is the segmentation of an area into regions that are acoustically similar [2]. While the basic process is well established, research into improved methods and algorithms continues. In this work, the initial sets of classes were produced in QTC IMPACT™. The steps in IMPACT's process are bottom picking, depth compensation, stacking, feature generation, dimension reduction with principal components analysis, and objective clustering with ACE. Further sets came from a developmental version of IMPACT that both replicates the commercial version and has the flexibility to use a wide variety of other techniques, particularly different feature algorithms.

Averaging a number of consecutive echoes, which is called stacking, is a method of increasing the signal-to-noise ratio. The noise is ping-to-ping variability and the signal is shape and spectral character of echoes. Tables 1-4 contain results based on stacking five echoes, a common practice. Table 5 compares classes with stacks of 1, 5, and 15 echoes.

ACE returns assignments of records into any number of classes, and recommends an optimal number of classes [2]. These data sets, from the six survey days and processed in various ways, almost always had an optimal class number from 4 to 6. Results that are to be compared for reproducibility need to have the same number of classes, so in each case the five-class solution was selected. This means that every classified point was assigned to a class numbered from 1 to 5. Figure 1 shows bathymetric cross-sections of the survey lines overlaid with classes from QTC IMPACT using data from 1 May. Figure 2 shows all 18 replicates (three on each of the six days) of the northern of the two transects coloured by the classes derived for each day.

Features are values calculated from echo time series that capture information that discriminates among echoes from different seabeds. Normalized cumulative integrals discriminate well by capturing details of echo shape. These features are values of the cumulative integral at fixed fractions of the number of samples being processed. Another set of shape features, evaluated recently [3], is the echo centre of gravity, energy spread, and skewness. These three are abbreviated to cg_es_skew in Table 4.

The commercial version of QTC IMPACT uses only features derived from echo shape and spectral character, not from amplitude. Amplitude features had been avoided because, if not corrected carefully for depth, they can produce depth-dependent artifacts. This is particularly true at high sonar frequencies where sound absorption can be important but water temperature and salinity are often unknown. This risk was minimized in the developmental version of IMPACT by compiling tables of amplitude against depth in a first pass through the echo data, and calculating from that a fitted trim-TVG correction curve to use in preparing each echo for feature generation. This is similar to Quester Tangent's method for range and angle compensation of multibeam images [4].

3. MEASURES OF REPRODUCIBILITY

Classification processes are reproducible if they consistently assign nearby locations to the same geological class. Acoustic classes from QTC IMPACT and most other methods, though, are maps of acoustic diversity, showing areas that are acoustically homogeneous and distinct from other areas. Non-acoustic data are needed to attach geological names. In this work we wish to measure reproducibility based on acoustic classes alone.

Each day's data were classified independently of the others, with its five classes randomly numbered. Tables are therefore needed showing the correspondence of class numbers on any particular day to the numbers on any other day. Matching classes could be based on proximity in feature space or geographically. Matching in feature space (Q space) is the more pure test: match in feature space then measure reproducibility of geographical neighbours. Figure 3 illustrates matching in feature space. Class centres are shown in a plot of the first two dimensions (Q space is three-dimensional) with colours and shapes denoting dates and class numbers respectively. Elliptical borders separate the obvious groups of six. Each of the five groups has one class from each of the six days, and these five groups are

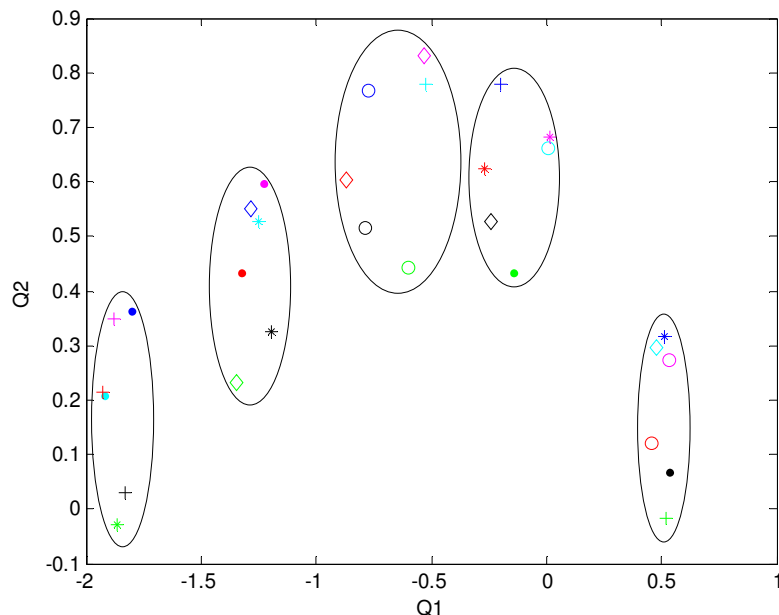


Fig.3: Groups of class centres in feature space (Q space). Each colour represents one of the six days and each symbol one of the five classes. Grouping is needed because clustering assigns numbers to classes randomly. In these data, which are from the first row in Table 1, the single acoustic class at high Q1 and low Q2 was assigned numbers 2, 1, 3, 4, 2, and 5 in the independent classification processes of each of the six days.

the correspondences we seek. To form these groups in less obvious cases, one could calculate the net intra-group distances for all possible arrangements and choose its minimum. However with six days and five classes there are almost 25 billion possible groups. To identify groups in a reasonable time, two groups of six were identified by eye, leaving distances to be calculated for only 7776 candidate groups.

A complication in making these matches is that principal components analysis within QTC IMPACT gives random signs to the component axes. Because all the situations studied in this work were at least reasonably reproducible, it was usually obvious which axes required their sign to be reversed to coincide with the others. The indicated sign reversals were done before matching.

Geographical interpolation was the next step in calculating reproducibility. While the repeated surveys followed the same transects accurately, the locations in these surveys to which classes were assigned did not coincide precisely. Taking one survey as the base set, all the classified points in another set that lie within a search radius, 10 m, were sought and a modal class identified using an occurrence histogram. This is the method of categorical interpolation in QTC CLAMS™ [2]. Let us take, as an example, a point in Class 1 in the first day's survey. Nearby points of the second day's survey might be predominantly in class 3, giving one correspondence between Day 1 Class 1 and Day 2 Class 3. Repeat for 500 points of the base set, and tabulate the correspondences in a 5x5 matrix with the base day's classes as columns and the other day's classes as rows. Table 1 is an example. Finally, rearrange the rows according to the feature-space matches found as described above. If reproducibility were perfect, all the off-diagonal elements of this matrix would be zero. A perfect situation is very rare, which is why these matrices are called confusion matrices.

The statistic called overall accuracy [5] is the trace of a confusion matrix divided by its sum, thus the fraction of assignments that agree. Some correct matches occur randomly, and the statistic Kappa, K, corrects for this [5].

$$K = \frac{OA - P_r}{1 - P_r} \quad (1)$$

Here OA is the overall accuracy and P_r is the random probability of a correct match in the confusion matrix. One way to calculate P_r is to assume all classes are equally probable, in which case it equals the reciprocal of the number of classes. Other methods include class populations.

In the literature, OA stands for Overall Accuracy and is a measure of a set of class assignments against a set of validation data that is known to be correct, or, at least, trustworthy. As used here, however, it is a measure of reproducibility among class maps, with no assumptions about which map has the highest fidelity. In this context, OA should perhaps stand for Overall Agreement and all references to accuracy should read agreement, even though accuracy is often used in this looser sense by other authors.

4. RESULTS AND DISCUSSION

The thesis of this work is that reproducibility among repeated surveys of the same area can indicate the relative merits of current and developmental methods for acoustic seabed classification. Methods that have high reproducibility achieve that by capturing echo

characteristics that are strongly influenced by seabed type while suppressing details that are driven by sea state or the water column.

Seasonal variations often occur in coastal waters and seabeds, due to physical and biological processes in those habitats. At Fowey Rocks there were seasonal changes between the four surveys in May and the two in August, as shown in Table 3. The overall accuracy of reproducibility was less between May-August pairs of days than among the four days in May. While these changes were real, their nature is not known. This seasonal variation does not interfere with assessing the merits of classification methods because each of the columns in Table 3 has the same pattern and would serve for this purpose.

For the transects at Fowey Rocks, surveyed three times on each of six days, reproducibility of classes among the days can be improved by using different families of

		Reference set, May 1				
		Class 1	Class 2	Class 3	Class 4	Class 5
Test set, May 2	Class 3	85	0	4	16	2
	Class 4	0	39	0	0	32
	Class 5	37	0	2	0	0
	Class 1	13	0	0	8	5
	Class 2	2	12	0	20	132

Table 1: A confusion matrix for 409 locations in the 2 May data set that were close enough to 500 random locations in the 1 May reference set to have a reference class assigned. For this matrix, the overall accuracy and Kappa are 0.65 and 0.51. The values in Table 2 were averages from three of these matrices for each pair of dates.

	May 1	May 2	May 9	May 28	Aug 3	Aug 13
May 1	100%	54%	51%	58%	40%	40%
May 2	62%	100%	53%	58%	41%	39%
May 9	54%	50%	100%	48%	49%	56%
May 28	55%	58%	54%	100%	40%	40%
Aug 3	48%	46%	52%	38%	100%	56%
Aug 13	42%	46%	58%	44%	53%	100%

Table 2: The Kappa measure of reproducibility for unsupervised classification with QTC IMPACT, with each of the six days treated as both reference and test data set.

Feature Families	Within May	May to August	All pairs of days
1	67%	57%	61%
2	59%	61%	61%
3	55%	46%	51%
4	55%	67%	63%
5	80%	70%	74%
6	80%	70%	73%

Table 3: Averaged Overall Accuracy with various feature sets considered seasonally.

Six day-to-day comparisons were made among the four days in May, and eight between May and August dates. Overall accuracies for the one pair of dates within August were omitted from this table because they are erratic.

features, particularly features that explicitly use echo amplitudes. Table 4 has values of overall accuracy, both raw and corrected for random assignments (Kappa), for the set 1, the current feature set of QTC IMPACT, and several developmental feature sets, 2-6. These new features will soon be commercially available in a real-time version of QTC IMPACT and in a new release of the post-processing software suite.

The first three families of features in set 1 of Table 4 (cumulative integral, quantile, and histogram) capture echo shape, while the FFT and wavelet features were designed to capture spectral content. Figure 2 shows that they produce useful and realistic maps of acoustic class even though the statistical values are less than impressive. Switching to an alternate method of depth compensation, in which the echo time series is not resampled to a fixed number of samples but rather the feature algorithms are modified to compensate as needed, does not affect reproducibility (wavelet features were dropped because they do not fit the alternate method). Replacing some of these shape features with others does not improve reproducibility, nor does calculating some fractal-based features in place of Fourier transforms. However, these results do show that a smaller number of features can be as effective as the original 166 features. Adding features based on echo amplitudes are, for this data set, a significant improvement. Kappa values as high as 80% were obtained with feature sets that include amplitude and with averaging over the best-matching five of the six days. Overall, this indicates that smaller feature sets with well-chosen shape and spectral features and with amplitude features significantly improves reproducibility and thus the quality of class maps.

Another process that improves reproducibility is to stack a larger number of echoes. Usually five echoes are stacked, that is, their time series are averaged to reduce ping-to-ping variability, which is a form of noise reduction. Table 5 shows that reproducibility is much lower without stacking but is usefully improved by stacking three times as many echoes. In

Feature Families	<i>n</i>	<i>OA</i>	<i>K</i>
1. Cumulative integral, quantile, histogram, wavelet, FFT	166	61%	49%
2. Cumulative integral, quantile, histogram, FFT	104	61%	48%
3. Cumulative integral and histogram, quantile, histogram, FFT, cg_es_skew, fractal	117	51%	38%
4. Cumulative integral, cg_es_skew, cumulative histogram, fractal	30	63%	53%
5. Cumulative integral, cg_es_skew, cumulative histogram, fractal, amplitude	35	74%	63%
6. Cumulative integral, cg_es_skew, fractal, amplitude	25	73%	63%

Table 4: Averaged Overall Accuracy (OA) and Kappa (K) with various feature sets. The number of features is n. Row 1 is the QTC IMPACT process with its usual depth compensation. A simpler compensation method was used for the other rows. Averages are from tables such as Table 2, omitting the 100% values on the diagonal. Values of Kappa over 0.80 were observed with amplitude features included and excluding the one day that matched least well.

Number of consecutive echoes stacked	<i>OA</i>	<i>K</i>
1 (no averaging)	0.36	0.19
5	0.61	0.49
15	0.69	0.60

Table 5: Effect of stacking various numbers of consecutive echoes. Overall Accuracy (OA) and Kappa (K) were averaged as for Table 4. Classification was done with IMPACT's present features and usual depth compensation.

Angle Filter	OA	K
No stacking and no filtering by angle	0.36	0.19
No stacking and pitch within 2° of horizontal	0.60	0.41
No stacking and beam angle within 5° of vertical	0.60	0.40

Table 6: Effect of filtering by beam angle and ship attitude with stacks of one (that is, no stacking). Overall Accuracy (OA) and Kappa (K) were averaged as for Table 4. Classification was done with IMPACT's present features and its usual depth compensation.

planning a survey, choosing the number to stack is a compromise between classification accuracy and spatial resolution.

A third process that improves reproducibility is to ignore echoes that were recorded with the boat pitched or rolled beyond some angle, or were from a steeply sloped seabed. Table 6 shows how reproducibility is improved by angle filtering, in an evaluation done without stacking. One filter was simply vessel pitch of more than 2°. Filtering by angle of incidence is more complicated because seabed slope and vessel attitude both have to be considered. The fraction of echoes that had angles of incidence above 5°, and were then set aside by the filter, ranged from 33% for the data from 3 Aug to 70% for 1 May. Both filters produce substantial improvements in reproducibility, as would be expected with such small angular tolerances and substantial fractions of the echoes being ignored. Further work is needed to define maximum allowable angles, which are expected to vary with beamwidth.

For each of these three processes, reproducibility has been shown to improve as the classification system is modified to emphasize seabed effects on the echoes. Developmental feature sets, especially those with features that are based on echo amplitudes, improve reproducibility, as does stacking a larger number of echoes, and as does using only echoes with angles of incidence near zero. This supports the thesis that improved reproducibility among repeated surveys of the same area is a reliable indication that the quality and accuracy of acoustic seabed classification has also been improved.

REFERENCES

- [1] **A.C.R. Gleason**, "Single-beam acoustic seabed classification in coral reef environments with application to the assessment of grouper and snapper habitat in the upper Florida Keys, USA", PhD thesis, Univ. Miami, 2009.
- [2] **J.M. Preston, A.C. Christney, L.S. Beran, and W.T. Collins**, "Statistical seabed segmentation", In Seventh European Conf. On Underwater Acoustics, Delft, Netherlands, Proceedings. Editor. D.G. Simons, pp 813-818, 2004.
- [3] **P.A. van Walree, J. Tegowski, C. Laban, and D.G. Simons**, "Acoustic seafloor discrimination with echo shape parameters: A comparison with the ground truth", Continental Shelf Rsrch 25 pp 2273-2293, 2005.
- [4] **A.C. Christney and J.M. Preston**, "Compensation of sonar image data primarily for seabed classification", US patent 6868041 (2004), UK patent 2403013 (2005).
- [5] **R.G. Congalton and K. Green**, "Assessing the accuracy of remotely sensed data: principles and practice", Boca Raton, Lewis Publishers, 1999.