Acoustic seabed classification methodology: a user's statistical comparison

Pierre Legendre
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec H3C 3J7, Canada

June 2002

E-mail: **Pierre.Legendre@umontreal.ca**

Abstract

Methods for seabed classification using echosounder data have been available for about 10 years. In this paper, we examine aspects of the statistical processing used by the QTC VIEW acoustic bottom classification system, which characterises bottom types through the shape of the first echo. We focused our investigation on 5 questions: (1) How to filter errors in input data? (2) How is principal component analysis (PCA) computed in the QTC software? How many principal components should be kept for classification analysis? Is PCA the only way of eliminating redundancies and noise in the data? (3) Is the method used by QTC for clustering a good method? (4) How to decide on the optimal number of acoustic classes? (5) How does the QTC classification compare to K-means results? While the decomposition of the acoustic signal into shape variables produced by QTC VIEW™ is empirically useful, the QTC IMPACT™ software implements sub-optimal classification procedures. An alternative method (PCA followed by K-means partitioning) is presented which produces statistically better results; software is freely available.

1. Introduction

The future of ecology as a partner for economic development lies in the ability of ecologists to develop means, tools, and methods for rapid assessment of impacts over broad expanses, such as a whole embayment, gulfs, or continental shelves in aquatic ecosystems.

This paper concerns remote sensing of coastal seabed using an acoustic bottom classification system for habitat mapping. Acoustic techniques allow managers to quickly map extensive seabed surfaces; they may eventually be used to map whole continental shelves. This information is urgently needed to assess the impact of coastal urban and industrial developments.

Among the emerging technologies for habitat mapping are acoustic seabed classification techniques, sidescan sonar, interferometric sidescan sonar, multibeam sonar, sediment profiling imagery (SPI camera), and underwater video. In recent years, technology has evolved to allow users to obtain multivariate acoustic data responding differentially to the nature of seabed types such as rock, sediment types, and fixed life forms (algae, seagrass, and invertebrate beds).

To obtain a map of bottom types, the echosounder signal is analysed using a series of computer-based numerical methods. An electronic signal with a given wavelength is sent by an electronic generator to a transducer placed vertically under or on the side of a boat or underwater vehicle. The transducer transforms this signal into a sound wave, or "ping", which is reflected by a portion of the bottom surface (called "footprint" in geostatistics or "grain" in ecology) and captured again by the transducer acting now as receiver. The first part of the reflected incidence signal indicates depth, which can easily be calculated by multiplying the time between emission and reception by the speed of sound in (sea) water.

Depending on the nature of the bottom surface, additional signals, called 'backscatter', are also received by the transducer after the reflected component arrives; they vary in delay and intensity depending on surface roughness and bottom penetration. See Hamilton et al. (1999) for details. The whole response wave is transformed into an analogue electronic signal, which can be sent to a computer for further analysis.

The RoxAnn™ system developed by Marine Micro Systems Ltd was the first commercial software to appear on the market (Chivers et al. 1990). This system uses a multi-echo energy approach. Other systems, e.g., QTC VIEW, characterise the bottom through the shape of the first echo. Seabed classification data are systematically collected along the vessel track and stored together with position (from GPS) and depth data. The RoxAnn and QTC acoustic bottom classification systems have been compared in terms of performance by Hamilton et al. (1999); they found that the QTC classes generally had consistent sediment grain size and texture properties, and gave a better classification than the RoxAnn system.

A few private companies have developed software for analysis of seabed echoes. Research groups in various countries are in the process of developing such software, e.g., the CSIRO in Australia (Pitcher et al. 1999) and the National Institute of Water and Atmospheric Research (NIWA) in New Zealand (Coombs 1994). The Defence Science and Technology Organisation (DSTO) in Australia have been investigating a new "first echo shape" approach, which shows promising results (L. J. Hamilton, pers. comm.).

During the Scale Expert Workshop, we experimented with the QTC VIEW™ software of the Canadian-based *Quester Tangent Corporation* (QTC) (Prager et al. 1995, Collins et al. 1996). The QTC system contains an analogue-to-digital converter which obtains the amplitude envelope of the waveform. It also contains a processor programmed to correct the shape of the response wave to account for depth by transformation to a particular reference depth, normalise the wave by setting peak amplitude of the waveform to unity[1] (which effectively removes energy information), and describe it by 166 shape variables. Waveform analysis is based, in part, on the work of Mayer et al. (1993, 1997) and Prager et al. (1993). The physical or mathematical meaning of the 166 variables is unknown for the user. We know, however, that some variables are based on echo shape and others on spectral characteristics, e.g., spectral moments. Insofar as we can understand them, the QTC variables correspond to:

1. A Fourier analysis (spectral decomposition) of the response wave producing a first set of 64 variables.

2. A wavelet analysis producing a second set of 64 variables.

3. 38 other variables giving information about the shape (distribution of amplitudes of the echoes), kurtosis, surface area, etc. of the response wave in its original and cumulative-transformed forms, and spectral information, e.g., spectral moments.

We are given to understand that the variables are "ranged", i.e., divided by some maximum value, since all values of the 166 QTC variables are in the interval [0,1]; this effectively eliminates the physical dimensions of the variables, making them suitable for statistical analysis without further standardisation. The QTC variables do not represent a universally accepted method of decomposing the response wave, although they may have empirical value for bottom mapping.

We focused our investigation on 5 questions: (1) How to filter errors in input data? (2) How is principal component analysis (PCA) computed in the QTC software? How many principal components should be kept for classification analysis? Is PCA the only way of eliminating redundancies and noise in the data? (3) Is the method used by QTC for clustering

[1] An alternative method, used by Hamilton (pers. comm.), is to set the area under the curve to unity.

a good method? (4) How to decide on the optimal number of acoustic classes? (5) How does the QTC classification compare to K-means results?

2. The QTC method of acoustic seabed classification

The QTC software uses the 166 variables to obtain a classification of the acoustic bottom records into a number of groups; these groups can then be plotted on geographic maps or on depth-by-transect-position plots. See Quester Tangent Corporation (1999: QTC VIEW™ CAPS; 2000: QTC IMPACT™). The manual states that the QTC approach to acoustic seabed classification "was built on the most recent approaches to digital signal processing available" (Quester Tangent Corporation 2000). We will scrutinize this statement. Examination of the statistical procedure implemented in the QTC software indicates that it proceeds as follows:

• The 166 QTC variables are very highly collinear; for our example data described below, the mean of the absolute values of the correlation coefficients was 0.41 with values or r ranging from –0.9999 to +0.9999. The filtering method used by the QTC software to extract pertinent information is to carry out a principal component analysis (PCA) of the 166 variables. PCA acts as a filter sorting out the relevant portion of the information and discarding noise. In the QTC software, the classification is based upon the first three principal components only. The remainder of the information is not used to obtain the classification. The QTC IMPACT™ manual (Quester Tangent Corporation 2000) states that "QTC research has determined that three principal components effectively describe each echo." Indeed, three principal components commonly account for 90% or more of the total variation in a data set. The QTC software allows users to plot the data points in a pseudo-3-dimensional graph along these three principal components, which are referred to as variables Q1, Q2 and Q3.

• The user asks the program to split the observations in two parts along one of the principal components. This is done as follows: (1) the user chooses a principal component (1, 2 or 3)[2]; (2) the data are split, presumably at the middle point of that component; (3) the mean of each of the two classes is calculated (the manual does not say if the mean is calculated along that component only, or if the 3-dimensional centroid is computed in the (Q1, Q2, Q3) space; (4) the points are reassigned to the closest mean. Three statistics are provided for each group in the "Class Statistics" window: the number of observations in the group, a statistic called "Chi2" and another one called "Score" which is the product of the previous two numbers. The total score, also given in the "Class Statistics" window, is the sum of the individual class scores for the given level of splitting. We have not been able to figure out how "Chi2" is calculated; the QTC CAPS™ manual states that "the Chi2 is a measure of the extent of the cluster in Q-space" while the QTC IMPACT™ manual says that it is "a measure of the clumpiness of each cluster", although it actually is the reverse, i.e., some measure of the dispersion of the points around their centroid.

• For the next divisions, the manuals (Quester Tangent Corporation 1999, 2000) are vague about how to proceed; they only state that "usually, the class with the highest score is the candidate for the next cluster split". It seems that one can use any of the three group statistics of the previous paragraph to determine which class should be split next: the number of class members, the "Chi2" or the "Score". We are told that "the means [are] adjusted with each iteration", which presumably means that steps 2 to 4 of the previous paragraph are completed after a given principal component has been selected by the user for splitting (step 1). "The process continues as long as the results of splitting improve the overall statistical description of the clusters [this presumably refers to the three group statistics] or until the class

[2] This option has become available in QTC IMPACT™ 2.0. In QTC CAPS™ and in previous versions of QTC IMPACT™, principal component 1 was always used for splitting the data.

memberships become stable [which cannot occur if one keeps on splitting groups]". As the classes are subdivided, the total score decreases as a result of tighter clusters (small chi-square) with smaller number of records. Beyond some point, further splits have little impact on the total score; this indicates the most appropriate number of distinct acoustic classes within a data set.

• At any step of the division procedure, the result may be plotted on a map, i.e., a graph whose axes are the X and Y geographic coordinates of the sampling points. Each group obtained in the classification is represented by a different colour. Another graph can be produced where the axes are the sampling sequence (in abscissa) and depth (in ordinate).

• At the end of the clustering procedure, extra data records that have not been used in the classification procedure can now be classified. For very large data sets, reference clusters may be generated through PCA and clustering of some regularly spaced subset of the data from the survey area (e.g., one record out of every ten). Following that, the positions of the remaining data records can be found on the PCs and each point attributed to the nearest cluster centroid.

• With or without extra records, one more iteration of steps 3 and 4 of the classification algorithm is done before the classification results are written to an output file. So, the number of members of each class usually does not fit the memberships displayed in the last step of the "Class Statistics" window. A confidence index, based upon the distance of each observation to its cluster's centroid, is also written out to this file.

3. Example data set: Forty Baskets Beach (FB), Sydney Harbour

On 16 August 1999, acoustic data were collected in Forty Baskets Beach area of Sydney Harbour (henceforth called "FB data"), Australia. We used a Navisound 50 echo sounder at frequency 50 kHz (transducer beam width 13.5°) connected to the QTC VIEW™ acoustic seabed classification system (CAPS version 3.25, QTC IMPACT™ version 1.0 Beta). The transducer was mounted on an over-the-side strut on the survey vessel. The positioning equipment was a differential GPS (Global Positioning System). After validation using our *Parser* program (described below), the data set consisted of 1478 data lines (objects or records) and 166 variables, plus geographic positions and depths. Since three variables did not vary at all, they were eliminated from the data set which was thus reduced to 163 variables.

The data were subdivided into groups using the procedure outlined in the QTC manual and described in Section 2. The score value was used to determine which class should be split next. As the classes were subdivided, the total score decreased (results not given here). However, at split level seven (i.e., 8 groups), the total score increased again. Results of the partitions into 2 to 7 groups are reported in Table 1 (top). The Calinski-Harabasz statistic (see below) selected the partition into 5 groups as the best one.

The same data were also analysed by K-means partitioning using the first three 3 PCs (PC1-PC3), as in the QTC procedure. For this example, they accounted for 96.2% of the total variance in the 163 variables. The partitioning program was asked to produce from 10 to 2 groups; the partitioning procedure was restarted 10 times. The best partitions into K=2 to K=7 groups were retained (Table 1, centre). The Calinski-Harabasz statistic (see below) selected the partition into 3 groups as the best one (Fig. 1).

We also computed K-means partitioning for the 163 variables in the data set, without prior filtering by PCA. The Calinski-Harabasz statistic (see below) selected the partition into 3 groups as the best one (Table 1, bottom). This partition is very similar to that obtained by K-means on PC1-PC3 (Table 3).

4. Questions about the QTC method of data analysis

4.1. Error in input data

The QTC software may produce several types of errors when recording the data and matching the coordinates with the vector containing the 166 variables. The QTC data file can contain erroneous data due to interrupt failure in the hardware. Excessive vessel speed and manoeuvring (e.g., sharp turns) can affect the acoustic data quality through the presence of highly reflective bubbles. The presence of fish or water column stratification can act to reflect the acoustic pulse which may yields a false indication of the bottom (Collins and Rhynas 1998). The QTC software (CAPS version 3.25) that we used in Sydney Harbour did not provide any kind of validation of the data. Truncated and mismatched lines of data have to be found by manual examination, and outliers or other aberrant data vectors have to be detected by users before carrying out the analysis. Users of later versions of the QTC software encounter these problems less frequently.

The Parser program. — The input of error-laden data in a computer program will fail, since the data do not conform to the expectations of the program. A parser is an algorithm capable of interpreting and modifying data in ways that will make it acceptable to the program. Modifications may consist of additions (also known as fillers of missing values), deletions, or type modifications (e.g., integer to decimal), based on criteria built into the parsing algorithm.

Users could check the data stream manually for errors and validate the QTC data, but with lengthy data streams (thousands of acoustic records) this task soon becomes tedious and error-prone.

The errors generated by the QTC hardware are somewhat stereotypic, however, allowing automation of much of the grunt work. A *Parser* program has been written for error correction and validation of the QTC data. Our *Parser* reads the data, decomposes them into different kinds of strings, analyses them, reports the errors it finds, and corrects them if possible. It outputs a clean data stream, suitable for import by other computer programs. It deals specifically with the following errors:

• Step 1: Ensure that any $ sign is preceded by an end-of-line. The QTC hardware sometimes omits to insert a “newline” character in the data stream just before the beginning of a new line, which always starts with the “$” character.

• Step 2: Remove lines that do not begin with one of the known tokens. Each line of the data stream begins with a keyword, or “token”, which specifies the kind of data on this line. Recognized tokens include:

$PFVEC, followed by depth and acoustic data (backscatter signal decomposed into 166 variables);

$GPGGA or $GPGLL[3], followed by latitude and longitude (GPS) record data;

All other tokens are ignored.

• Step 3: Remove single $PFVEC records. An acoustic data record ($PFVEC) is only useful if it has a known geographic location, which means that it must be followed by a $GPGGA or $GPGLL record. If it is not, the $PFVEC record is eliminated.

[3] Different GPS systems or settings may generate position records with “tokens” that differ from $GPGLL or $GPGGA.

• Step 4: Remove multiple $GPGGA/$GPGLL records. Multiple locations ($GPGGA or $GPGLL) following an acoustic data record are not useful. Only the first location record is kept, all others are ignored.

• Step 5: Validate the data. Validation is done differently for acoustic and location data. With the former, the Parser verifies that each $PFVEC record contains 167 fields: depth, followed by the 166 QTC backscatter variables. Location data is validated by ensuring that each $GPGGA record contains 5 fields, or 4 fields in the case of $GPGLL records: latitude, north or south, longitude, east or west.

• Step 6: Ensure that each $PFVEC record with unknown depth is removed. During our work, we found that when depth was lost, it was most often indicated as "0.45" in the $PFVEC record. Since we only want complete records, records with unknown depths, given as 0.45, are ignored.

When these steps are completed, a report and a clean data stream are generated, ready for use by other programs in the chain of analysis.

4.2. QTC is using 3 PCs only for classification

While PCA is certainly an efficient method for filtering noise, we will ask the following questions: (a) How is PCA computed in the QTC software? (b) How many principal components should be kept for classification analysis? (c) Is PCA the only way of eliminating redundancies and noise in the data? If not, how do the classification results compare with an alternative method of differential variable selection?

4.2.1. How is principal component analysis (PCA) computed in the QTC software? — There are various ways of computing and scaling principal component in PCA, but the manual does not explain how they are computed by the QTC software. So, we computed PCA for the FB data described above (1478 objects x 163 variables), without and with standardization of the variables. Using different scalings of the eigenvectors, we compared our results to the principal components produced by the QTC software. We found that the principal components (Q1, Q2, Q3) of the QTC software are computed from the covariance matrix (i.e., non-standardized data), as they should for variables that have already been ranged by some maximum value and are thus without physical dimensions. The eigenvectors are scaled to unit length by the QTC software before the principal components are computed, which is also correct. In this way, the principal components preserve the distances among objects (records) in multidimensional space. Distances among records computed from the first three principal components only are thus slight underestimates of the true distances in 166-dimensional space if the first three principal components account for most of the variation in the data. This was indeed the case for the FB data set; see "Example data set" above.

The QTC software adds an arbitrarily-chosen constant to each principal component, perhaps in order for the pseudo-three-dimensional ordination diagrams to display all records on the same side of each axis. How this constant is chosen is unclear. It is not a round number, and it does not bring the extreme points to a round-number coordinate value. In any case, this does not matter for the classification phase of the calculations.

4.2.2. How many principal components should be kept for classification analysis? — For the FB data, the first three principal components accounted for 96.2% of the total variance. Using 7 principal components would account, however, for 99.2% of the variance, which would be better. There may be useful classificatory information left in the unused principal components. For other data sets (Hewitt, Legendre et al., in prep.), the following results were obtained:

1. Proportion of the variance explained by the first 3 PCs: 90% to 97%. So it appears that the statement that "with only 3 out of 166 eigenvectors we can typically account for over 95% of the covariance produced from several thousand pings spanning a wide variety of seabed types" (Prager et al. 1995) does not apply to all cases.

2. Number of PCs necessary to reach 95% of the variance: 3 to 5.

3. Number of PCs necessary to reach 99% of the variance: 6 to 10.

4. The number of groups is not a monotonic function of the number of records. Using the number of PCs necessary to reach 95% of the variance in the K-means partitioning procedure, 6 groups were found, using the C-H criterion, in an analysis of 1571 records; 2 groups for 4011 records; 14 groups for 6406 records; and 18 groups for 24497 records. Ground truthing was done for some of these data sets using underwater video; it indicated that the groups identified by K-means partitioning were ecologically meaningful (Hewitt, Legendre et al., in prep.).

4.2.3. Is PCA the only way of eliminating redundancies and noise in the data? If not, how do the classification results compare with an alternative method of differential variable selection? — PCA is not the only way of eliminating redundancies and noise from data. An alternative method would be to eliminate the variables that are highly responsible for collinearity, followed by classification using only the remaining variables.

We used the following approach to variable selection for the FB example data. After performing a preliminary K-means analysis using all 163 variables, we used discriminant analysis to discard collinear and noisy variables. We actually used the StepDisc procedure of the SAS Package (SAS Institute Inc. 1995) which implements a backward elimination procedure for selection of variables in discriminant analysis. For the 2-group solution produced by K-means (K=2), StepDisc suggested to eliminate 114 of the 163 variables. For K=3, 4 and 5, respectively, 89, 95 and 91 variables were eliminated. The results, not reported in detail, were very similar to the partitions obtained by K-means using PC1 to PC3, displayed in the lower portion of Table 1. The similarities between the K-means PC1-PC3 partition for K=3 and the K-means partition for K=3 after elimination of 89 variables were, for instance: Rand = 0.96, modified Rand = 0.92 (see Section 4.5.2 for these two forms of the Rand statistic). So, using the first three principal components, which accounted for 95% of the variance, gave approximately the same results in K-means as the more complex procedure which involved three steps: (1) a first partitioning of the data, (2) variable elimination by stepwise discriminant analysis, (3) partitioning again using only the remaining variables. Furthermore, the K-means PC1-PC3 partition is more similar to the K-means partition based upon 163 variables than the latter is to the K-means partition obtained after variable selection. (They both differed greatly, however, from the classification obtained from the QTC software.) The procedure involving variable selection is thus not recommended for QTC data.

4.3. Is the method used by QTC for clustering a good method?

The QTC algorithm stops after a single iteration of a procedure that could lead to an optimal solution. Thus, it is unlikely to find at least a locally optimal solution, let alone the global optimum. If the process of computing group centroids and reassigning the observations to the nearest centroid were continued in an iterative way till convergence (i.e., until no further improvement of the split could be obtained), the algorithm would find at least a local optimum in the least-squares sense. In partitioning analysis, a least-squares optimal solution is one where the sum (across the groups) of the sums of squared residual distances to the group centroids (called SSE below) is minimum. Finding a partition of *n* objects into K groups following this criterion is the classical K-means problem of cluster analysis (Jain and Dubes 1988; Legendre and Legendre 1998).

K-means partitioning. — K-means is the most widely used numerical method for partitioning data. The K-means problem consists of dividing a set of multivariate data into nonoverlapping groups in such a way as to minimize the sum of within-group sums-of-squares, also called the "sum of squared errors" (SSE). SSE is the global optimality criterion, or objective function, implemented in K-means algorithms. There are hundreds of algorithms that have been proposed in the literature to solve the K-means problem. Many are easy to program. None of them is implemented in the QTC software.

The program developed by the lead author during the Scale Expert Workshop is a simple two-step iterative least-squares algorithm:

• Compute cluster centroids and use them as new cluster seeds.

• Assign each object to the nearest cluster seed.

This algorithm is described in several books, e.g., Legendre and Legendre (1998). It was modified to incorporate weights. Since K-means is a NP-hard problem (a category of very hard problems in computing science), no algorithm can guarantee that it will find the optimum partition every time. To increase the likelihood of finding this partition, two features have been added to the basic algorithm:

(1) The program proceeds in a cascade, finding first a partition into a number of groups larger than that which is needed (e.g. starting at 10 groups). The reason is that it is easier to find the best partition for a large number of groups than for a smaller number of groups. When this partition has been found, the two groups whose centroids are the closest are fused and the algorithm iterates again to optimize the SSE function. This is repeated as far as the user wants it to go (e.g., until a partition into 2 groups is found).

(2) The whole classification process (e.g., from 10 to 2 groups) can be repeated a number of times (e.g., 10, 25, or 50 times, as specified by the user) using different random starting configurations. For each number of groups (e.g., for K = 10, K = 9, …, K = 2 groups), the solution where SSE_K is minimum is retained and written to the output file.

Following K-means analysis, we will be able (Section 4.5) to compare the QTC partitions to partitions into the same number of groups obtained by K-means.

4.4. How to decide on the optimal number of acoustic classes?

The QTC software offers no indication as to the partition having the "best" number of groups in some statistical sense. The user is supposed to decide on the optimal number of classes using the total score value, given in the "Class Statistics" window, as described in Section 2. This decision may thus be subjective or difficult to make. For the FB data set, for example, the total score value decreased, then started increasing again at split level seven (i.e., 8 groups).

A large number of criteria have been proposed in the statistical literature to decide on the correct number of groups in cluster analysis. A simulation study by Milligan and Cooper (1985) compared 30 of these criteria. The best one turned out to be the Calinski-Harabasz (1974) criterion, called C-H in the present paper. C-H is simply the *F*-statistic of multivariate analysis of variance and canonical analysis. *F* is the ratio of the mean square for the given partition divided by the mean square for the residuals. To help users decide on the best number of groups present in a data set, our K-means program computes the C-H criterion; the number of classes for which C-H is maximum is the best one in the least-squares sense.

One cannot assume that the best number of groups is small in acoustic sediment classification: using the C-H criterion, Hewitt, Legendre et al. (in prep.) found cases where the best number of groups was from K=2 to K=19, depending on the data set.

4.5. How does the QTC classification compare to K-means results?

The QTC software does not compute the Calinski-Harabasz (C-H) or any other statistical optimality criterion for the partitions that it produces. To compare the QTC to our K-means results, it was necessary to write a program designed to compare different partitions using objective criteria.

The ComparePart program. — This program serves two purposes: (1) determine how well each partition reflects a given multivariate data set, and (2) compare several partitions of the same data among themselves.

4.5.1. To determine how well a partition reflects a given multivariate data set (either the original 166 variables or the first three principal components derived from them), the error sum-of-squares (SSE) and Calinski-Harabasz (C-H) statistics are computed. Among two partitions that have the same number of groups, the best one in the least-squares sense is that which has the lowest value of SSE. For a group of partitions into different numbers of groups (for instance K=2 to K=7), the best one in the least-squares sense is that which has the highest value of C-H; the SSE statistic cannot be used to compare partitions with different numbers of groups.

For the FB data, we have seen that the best QTC classification for PC1-PC3 was into 5 groups, based upon the C-H statistics (Table 1), whereas the best one obtained by K-means for either PC1-PC3 or all 163 variables was into 3 groups. A fair comparison requires computing the SSE statistics for all partitions into 3 groups, based upon PC1-PC1 and also the 163 variables, and the same for all partitions into 5 groups. For a given number of groups, the partition that has the lowest value for SSE is the best one.

Table 2 shows that the QTC partition into 3 groups is not as good as the K-means partitions into 3 groups based upon either the first three principal components or all 163 variables, with respect to either the first 3 three principal components (32% larger SSE) or the 163 variables (27% larger SSE). Likewise for the partitions into 5 groups: the QTC solution into 5 groups is much worse than the K-means solutions based upon either PC1-PC3 or all 163 variables, with respect to either the first three principal components (15% larger SSE) or the 163 variables (10% larger SSE). This clearly shows that the QTC partitions (even the "best one" into 5 groups) are far from being as good, for this example, as those obtained by K-means. Similar results were obtained with all other data sets that we investigated.

4.5.2. A number of partitions can be compared using indices that are appropriate for the comparison of partitions. We used the Rand (1971) index and the modified Rand index (Hubert and Arabie 1985) to compare partitions. The Rand coefficient produces values in the interval [0, 1] with 0 meaning no similarity. The modified Rand index is the most widely used coefficients to compare partitions; if the relationship between two partitions is comparable to that of partitions picked at random, the corrected Rand index returns a value near 0, which can be slightly negative or positive; similar partitions have indices near 1. Matrices of any one of these two coefficients can be used to produce ordination diagrams representing graphically the similarities among partitions.

Results for the FB example are reported in Table 3 (Rand and modified Rand indices) and Fig. 2 (principal coordinate ordination of the partitions based upon the Rand indices). They show that the partitions into 3 groups based upon all 163 variables or the first 3

principal components are very similar to each other, and that they differ greatly from any one of the partitions obtained from the QTC software.

5. Discussion

The decomposition of acoustic signals into 166 variables produced by the QTC VIEW™ system is an acceptable empirical method, although the physical and mathematical nature of the individual variables remains unknown to the users. We have shown, however, that the QTC acoustic class partitions are not as good, statistically speaking, as those obtained by the K-means procedure presented in this paper. (1) The QTC manual does not provide clear information as to the calculation procedures (it took a lot of guess work and experimenting before we came up with the description of the QTC method of acoustic seabed classification presented above) and the QTC splitting process is subjective. (2) The QTC software does not produce classifications where groups of records are best separated in the least-squares sense. It should be modified to optimize a global statistical criterion, like the sum of within-group sums-of-squares (SSE). The classification procedure presently implemented corresponds to the first two iterations of our K-means algorithm; the problem is that it stops there instead of going on to find an optimum of the least-squares criterion (SSE). (3) Some statistical stopping rule, from the many that have been suggested in the literature, should be computed by the software to help users decide on the correct or most interesting number of acoustic classes.

The procedure that we recommend for classification of acoustic data, represented by QTC or other variables, is the following: (1) Compute a principal component analysis of the data. If the variables have already been ranged with respect to some maximum value, no further standardization by variable is necessary. Scale the eigenvectors to length 1 in order for the principal components to reflect the Euclidean distances among records. (2) Select enough principal components to represent from 95% to 99% of the variance in the data. For the data sets that we experimented with, 95% of the variance was represented by 3 to 5 principal components; to reach 99% of the variance required 6 to 10 principal components. (3) Compute K-means partitioning of the records for a range of number of groups K (e.g., from K=10 or 20 to K=2); for each number of groups K, this partitioning method optimizes the separation of the groups by minimizing the sum of within-group sums-of-squares (SSE). Repeat the analysis a number of times (e.g., 10 to 50 times) to increase the chance of finding the optimal solution for each value of K. (4) Use some statistical criterion to decide what the best number of acoustic classes is. One can use, for instance, the Calinski-Harabasz criterion; the number of classes for which this statistic is maximum is the best one in the least-squares sense. (5) Chart the records using symbols and/or colour to differentiate the acoustic classes. Class overlaps and inconsistencies may then be checked for.

Acoustic classification results should be subjected to ground truthing, which consists in relating the acoustic classes to observed data describing the seabed. Hewitt, Legendre et al. (in prep.) have used underwater video data to validate the acoustic classification of sites in New Zealand.

The QTC software always averages the backscatter signal of 5 pings. It may be too few or too many for the site under study, depending on the speed of the boat and the size of the seabed structures to be identified, compared to the sum of the footprints of five pings. Pings are averaged before statistics are produced; the objective is to increase acoustic stability. The size of the footprint represented by the average of 5 pings poses a problem when QTC data are to be matched to observational data of other natures, e.g., underwater video data: are we dealing with a footprint of the size of 1 or 5 pings? Averaging 5 pings is not equivalent to measuring a single ping since averaging should reduce the variance. Is it the equivalent of measuring a 5-ping footprint? It is up to QTC to demonstrate it.

The footprint of individual pings increases in size with depth if the transducer is fixed to a boat or raft. This is not the case if the transducer is carried by a remote underwater vehicle operating at constant distance from the sea bottom. On the other hand, the ecology of environmental gradients tells us that communities of organisms inhabiting the sediment often vary with water depth, and so is the sediment structure. If the groups of records are found to be related to depth, how much of this effect is due to a real difference in sediment structure or in the communities inhabiting it, and how much is a simple geostatistical effect since it is well-known that the variance of spatially-structured variables varies with footprint (grain size)? The acoustic technology may only be applicable to the classification of sediment types within narrow ranges of depths, until transducers with beam width that automatically varies with depth (providing constant footprint) become available, or the use of remote underwater vehicles becomes generalized.

Acknowledgements

The work reported in this paper was conducted in great part during the SCALE EXPERT workshop (Spatial Comparisons Across Large Estuaries: EXPerimental Evaluation of Recent Technologies) organised and hosted by Prof. A. J. Underwood at the University of Sydney, Australia, 2-22 August 1999. We are grateful to Jim Drury, Fisheries Biologist at the National Institute of Water and Atmospheric Research (NIWA) in Auckland, New Zealand, for demonstration of and help in understanding the QTC IMPACT™ Acoustic Seabed Classification program, version 2.00, and to Judi Hewitt, Benthic Ecologist at NIWA-Hamilton, who performed numerous comparisons of the QTC IMPACT™ program results to those of the K-means partitioning program described in the present paper. Les Hamilton, Defence Science & Technology Organisation (DSTO) in Australia, helped us understand the QTC waveform analysis and provided comments on a draft of the manuscript. Programs for principal component analysis and K-means partitioning of large data sets are available free of charge from the WWWeb site <http://www.fas.umontreal.ca/biol/legendre/>.

References

Calinski, T. and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3: 1-27.

Casgrain, P. and P. Legendre. 2001. *The R Package for Multivariate and Spatial Analysis, version 4.0 d3 – User's Manual.* Département de sciences biologiques, Université de Montréal. Software and user's manual available on the WWWeb site <http://www.fas.umontreal.ca/BIOL/legendre/>.

Chivers, R. C., N. Emerson and D. R. Burns. 1990. New acoustic processing for underway surveying. *The Hydrographic Journal* 56: 9-17.

Collins, W., R. Gregory and J. Anderson. 1996. A digital approach to seabeb classification. Habitat assessment for juvenile cod is just one application of this acoustic method. *Sea Technology* 37: 83-87.

Collins, W. T. and K. P. Rhynas. 1998. Acoustic seabed classification using echo sounders: operational considerations and strategies. Canadian Hydrographic Conference '98, CHS, Victoria, British Columbia, Canada, March 1998.

Coombs, R. 1994. An adaptable acoustic data acquisition system for fish stock assessment. *International Conference on Underwater Acoustics*, Australian Acoustical Society, December 1994.

Hamilton, L. J., P. J. Mulhearn and R. Poeckert. 1999. Comparison of RoxAnn and QTC-View acoustic bottom classification system performance for the Cairns area, Great Barrier Reef, Australia. *Continental Shelf Research* 19: 1577-1597.

Hubert, L. J. and P. Arabie. 1985. Comparing partitions. *J. Classif.* 2: 193-218.

Jain, A. K. and R. C. Dubes. 1988. Algorithms for clustering data. Prentice Hall, Englewood Cliffs, New Jersey.

Legendre, P. and L. Legendre. 1998. *Numerical Ecology. 2nd English edition.* Elsevier Science BV, Amsterdam.

Mayer, L. A., S. Dijkstra, J. E. Hughes Clarke, M. Paton and C. Ware. 1997. Interactive tools for the exploration and analysis of multibeam and other seafloor acoustic data. In: N. G. Pace, E. Pouliquen, O. Bergem and A. Lyons (eds.), High frequency acoustics in shallow water. SACLANT Conference Proceedings Series, CP-45, NATO SACTLANT Research Centre, La Spezia, Italy, pp. 355-362.

Mayer, L. A., J. E. Hughes Clarke, D. E. Wells, and the HYGRO-92 Team. 1993. A multi-faceted acoustic ground-truthing experiment in the Bay of Fundy. In: N.G. Pace and D.N. Langhorne (eds.), *Acoustic classification and mapping of the seabed*. Proceedings of the Institute of Acoustics, v. 15, pt. 2, pp. 203-220. Bath University Press, Bath, U.K.

Milligan, G. W. and M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159-179.

Pitcher, C. R., S. R. Gordon, R. J. Kloser and P. N. Jones. 1999. *Development of an acoustic system for remote sensing of benthic fisheries habitat for mapping, monitoring and impact assessment.* Project No. 93/058, CSIRO Division of Marine Research, Cleveland, Queensland, Australia.

Prager, B. T., D. A. Caughey and R. H. Poeckert. 1995. Bottom classification: Operational results from QTC View. *Oceans-95, Challenges of Our Changing Global Environment* Conference, San Diego, CA, USA, October 1995.

Prager, B. T., R. Inkster, P. Lacroix and L. A. Mayer. 1993. ISAH-S bottom classification – Preliminary results. *Oceans-93* III: 202-207.

Quester Tangent Corporation 1999. *CLUSTER Operator's Manual*. 24 March 1999. Quester Tangent Corporation, Canada.

Quester Tangent Corporation 2000. *QTC IMPACT™ acoustic seabed classification, user guide version 2.00. Integrated mapping, processing and classification toolkit. Revision 2.* Quester Tangent Corporation, Canada.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* 66: 846-850.

SAS Institute Inc. 1985. *SAS user's guide: statistics. Version 5 edition.* SAS Institute Inc., Cary, North Carolina.

Shewchuk, J. R. 1996. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. *First Workshop on Applied Computational Geometry*, ACM, May 1996.

Table 1. Comparison of partitions into K=2 to K=7 groups obtained by QTC (using PC1-PC3), K-means (using PC1-PC3) and K-means (using all 163 variables). The total sum-of-squares was 206.69 for the first three principal components and 214.91 for all 163 variables. SSE: sum of within-group sums-of-squares. C-H: Calinski-Harabasz (C-H) statistic; the maximum of C-H (*) indicates the best partition in the least-squares sense.

Software and variables	No. groups K	SSE	C-H (high is better)	Group membership
QTC (PC1-PC3)	2	86.70	2043	1029 449
QTC (PC1-PC3)	3	55.16	2026	804 375 299
QTC (PC1-PC3)	4	41.28	1969	584 345 345 204
QTC (PC1-PC3)	5	29.85	2182 *	537 341 216 204 180
QTC (PC1-PC3)	6	25.93	2053	430 345 206 179 177 141
QTC (PC1-PC3)	7	23.39	1921	347 339 205 177 143 151 116
K-means (PC1-PC3)	2	81.29	2276	1074 404
K-means (PC1-PC3)	3	41.86	2904 *	611 587 280
K-means (PC1-PC3)	4	31.81	2701	525 513 231 209
K-means (PC1-PC3)	5	25.97	2561	381 362 361 188 186
K-means (PC1-PC3)	6	22.92	2360	382 338 324 199 148 87
K-means (PC1-PC3)	7	20.25	2257	301 296 274 196 175 132 104
K-means (163 var.)	2	89.48	2269	1123 355
K-means (163 var.)	3	49.67	2453 *	619 580 279
K-means (163 var.)	4	39.49	2183	526 515 237 200
K-means (163 var.)	5	33.51	1993	364 361 361 198 194
K-means (163 var.)	6	30.07	1810	422 341 273 207 145 90
K-means (163 var.)	7	27.27	1687	317 308 232 221 185 138 77

Table 2. Comparison of partitions (SSE and C-H statistics) using two different bases: the PC1-PC3 (middle) and 163 variables (right). SSE: sum of within-group sums-of-squares (small is best, for a given number of groups). C-H: Calinski-Harabasz (C-H) statistic (high is best among partitions obtained using the same data). The stars (*) indicate the best number of groups for that classification, according to C-H; see Table 1.

Software	No. groups	Base: PC1-PC3		Base: 163 variables	
and variables	K	SSE	C-H	SSE	C-H
K-means (163 var.)	3 *	41.87	2903	49.67	2453
K-means (163 var.)	5	25.97	2563	33.51	1993
K-means (PC1-PC3)	3 *	41.86	2904	49.67	2453
K-means (PC1-PC3)	5	25.97	2561	33.53	1992
QTC (PC1-PC3)	2		2043		1867
QTC (PC1-PC3)	3	55.16	2026	62.94	1781
QTC (PC1-PC3)	4		1969		1678
QTC (PC1-PC3)	5 *	29.85	2182	37.00	1771
QTC (PC1-PC3)	6		2053		1620
QTC (PC1-PC3)	7		1921		1484

Table 3. Comparison of partitions using the Rand and modified Rand indices. The stars (*) indicate the best number of groups for that classification, according to the Calinski-Harabasz (C-H) criterion; see Table 1.

Software and variables	No. groups K	Rand index K-means (163 var.)	Rand index K-means (PC1-PC3)	Modified Rand index K-means (163 var.)	Modified Rand index K-means (PC1-PC3)
K-means (163 var.)	3 *	1.00000	0.99154	1.00000	0.98174
K-means (PC1-PC3)	3 *	0.99154	1.00000	0.98174	1.00000
QTC (PC1-PC3)	2	0.64701	0.64950	0.32218	0.32706
QTC (PC1-PC3)	3	0.71784	0.71568	0.40365	0.39894
QTC (PC1-PC3)	4	0.83388	0.82998	0.62367	0.61464
QTC (PC1-PC3)	5 *	0.80751	0.80234	0.55191	0.53960
QTC (PC1-PC3)	6	0.78892	0.78514	0.49466	0.48525
QTC (PC1-PC3)	7	0.76623	0.76359	0.42947	0.42256

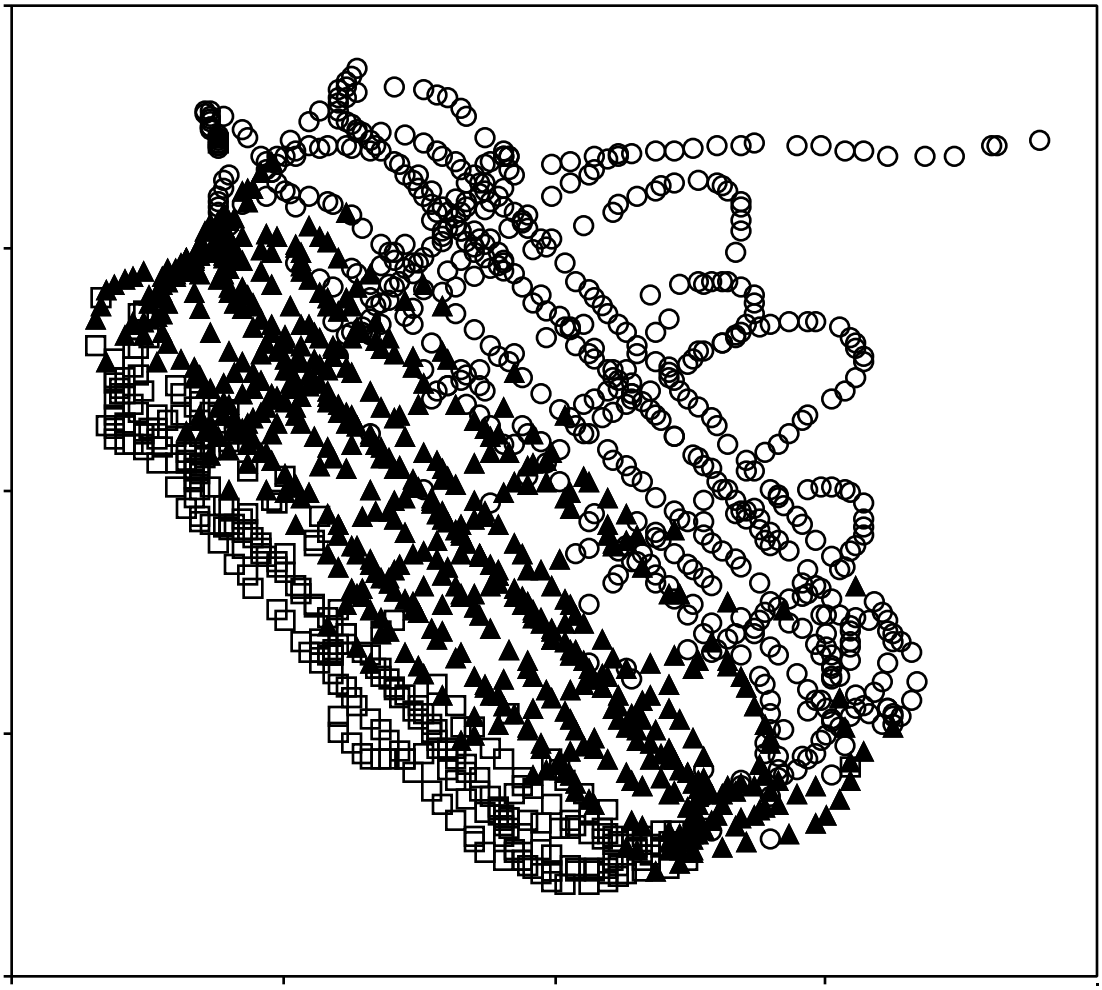

Fig. 1. Map of the Forty Baskets Beach sampling area showing the K-means partition of the records into 3 groups (symbols) based upon PC1-PC3.

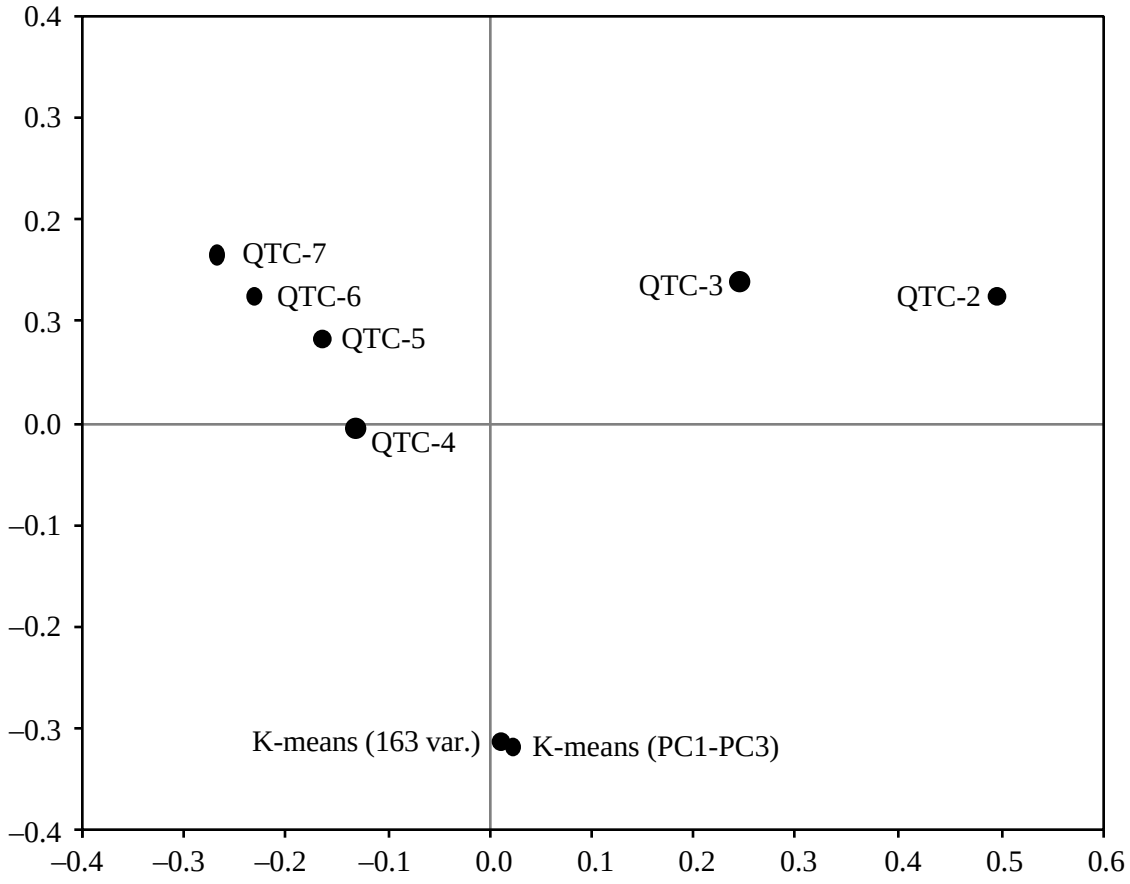

Fig. 2. Principal coordinate ordination of 8 partitions, based upon the matrix of Rand similarity indices (Table 2). Partitions that are closer in the graph are more similar. Ordination axes I (abscissa) and II (ordinate) account together for 79% of the variation in the matrix. The modified Rand index ordination (not shown) is nearly identical.