RAPID COMMUNICATION / COMMUNICATION RAPIDE

Acoustic seabed classification: improved statistical method

Pierre Legendre, Kari Elsa Ellingsen, Erik Bjørnbom, and Philippe Casgrain

Abstract: Huge amounts of money will be spent by industrialized nations during the next decades to obtain detailed maps of continental shelf seabeds. These maps, which will allow a more rational exploitation of the sea floor, are needed to assess the impact of anthropic activities. The statistical method of analysis of echosounder backscatter data described in this paper presents several improvements over existing techniques. The steps are as follows. (*i*) The backscatter data are decomposed mathematically into a number of quantitative variables, which are subjected to principal component analysis (PCA). (*ii*) Principal components representing 95–99% of the variation are used in a *K*-means partitioning procedure. A statistical criterion indicates what the number of groups is that best reflects the variability of the data. (*iii*) The groups are then plotted on maps of the survey area. Insofar as the mathematical decomposition produces variables that reflect the variations of the physical nature and composition of the seabed, the classes of the partition will correspond to different seabed types. Free software (The Q Package) implementing this method is available at http://www.fas.umontreal.ca/biol/legendre/.

Résumé : Au cours des prochaines décennies, des sommes considérables seront consacrées par les nations industrialisées à la cartographie détaillée des plateaux continentaux. Ces cartes, qui permettront une exploitation plus rationnelle des fonds marins, sont nécessaires pour évaluer l'impacts des activités anthropiques sur ces mêmes fonds. La méthode statistique d'analyse de l'onde réfléchie secondaire des sonars décrite dans cet article propose plusieurs améliorations par rapport aux méthodes actuellement sur le marché. Les étapes sont les suivantes : (*i*) l'onde réfléchie secondaire du sonar est décomposée en une série de variables quantitatives qui sont soumises à l'analyse en composantes principales (ACP). (*ii*) Les composantes principales représentant de 95 à 99 % de la variance sont utilisées pour obtenir une partition des points de sondage en groupes. Un critère statistique permet de déterminer quel est le nombre optimal de groupes pour rendre compte de la variabilité des données. (*iii*) La classification est reportée sur une carte de la région à l'étude. Si la décomposition mathématique de l'onde réfléchie secondaire produit des variables qui reflètent les variations de la nature et de la composition physique du fond, les classes de la partition correspondront à différents types de fond. Un programme d'ordinateur (The Q Package) est gratuitement à la disposition des utilisateurs à l'adresse http://www.fas.umontreal.ca/biol/legendre/.

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 whole embayments, 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.

Classification method

This paper presents a method of statistical analysis of echosounder backscatter data, which includes several improvements over existing techniques. The (free) software implementing this method is described at the end of this paper. Our test data consist in a file of first echosounder returns (Fig. 1) decomposed into 166 variables using the QTC VIEW^{TM} acoustic bottom classification system (Prager et al. 1995). Alternative methods (and software) for decomposing backscatters into sediment-related variables have been proposed, for example, by Chivers et al. (1990) and Clarke and Hamilton (1999). Software for statistical processing of the

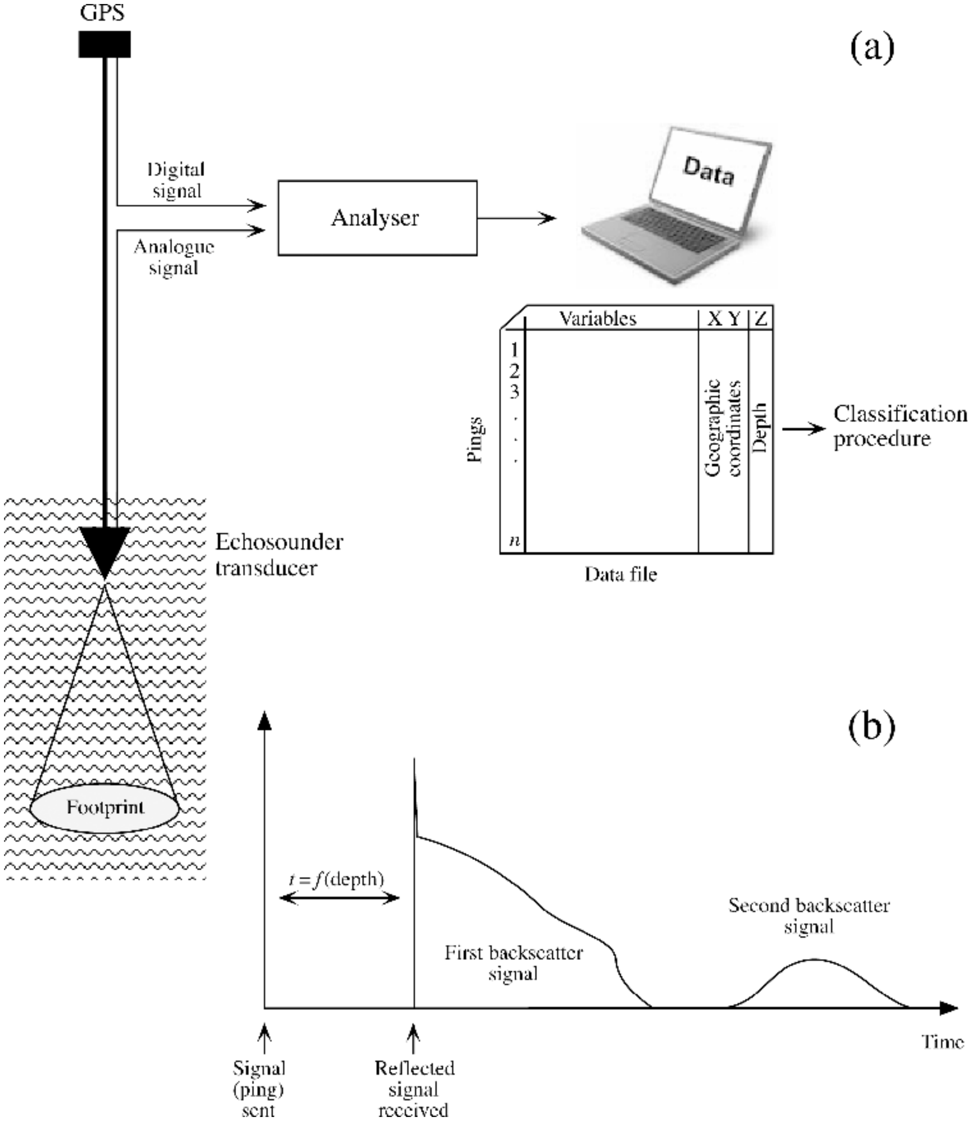
Received 26 April 2002. Accepted 21 June 2002. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 12 July 2002.
J16871

P. Legendre[1] and P. Casgrain. Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, QC H3C 3J7, Canada.
K.E. Ellingsen and E. Bjørnbom. Department of Biology, University of Oslo, P.O. Box 1064, Blindern, 0316 Oslo, Norway.

[1] Corresponding author (e-mail: pierre.legendre@umontreal.ca).

DOI: 10.1139/F02-096

Fig. 1. (*a*) Data acquisition: the echosounder signal is decomposed mathematically into a number of variables that will be used for classification. Each acoustic record is geo-referenced for mapping. (*b*) Analogue signal from the echosounder. The first backscatter portion of each “ping” is analysed in the present paper.



QTC variables is also available from the Quester Tangent Corporation (QTC 1999, 2000). The steps of our analysis follow.

Step 1: reduction of data dimensionality

The 166 QTC variables are very highly collinear; in our example data, the mean of the absolute values of the correlation coefficients was 0.41 with values of *r* ranging from –0.9999 to +0.9999. For highly collinear data, a commonly used method to condense the variance into a small number of variables, prior to classification, is principal component analysis (PCA); this is the method also used in the QTC software. PCA computes a smaller set of new, linearly independent variables, called principal components (PCs), that account for most of the variance in the original data. The remainder of the variance is considered the error portion of the data (noise). We carried out a detailed comparison of classification results based on the whole data set, on the one hand, and on a small number (2–8) of PCs accounting for most for the variance, on the other hand. Comparable *K*-means partitioning results (see below) were obtained by using a number of PCs accounting for 95–99% of the total variance in the data. So, variance condensation into a small number of PCs is a good method if a sufficient number of PCs are used for classification. For the test data, the first three PCs accounted for 96.2% of the total variance. Using seven PCs would have accounted for 99.2% of the variance. For other QTC data sets (J.E. Hewitt, S.F. Thrush, P. Legendre, J. Ellis, and M. Morrison, National Institute of Water and Atmospheric Research (NIWA), P.O. Box 11-115, Hamilton, New Zealand, unpublished data), the first three PCs accounted for 90–97% of the variance of the 166 QTC variables; 3–5 PCs were necessary to reach 95% of the variance, and 6–10 to reach 99%.

Step 2: *K*-means partitioning

A (crisp) partition is a division of the “objects” under study into nonoverlapping subsets. Agglomerative clustering methods produce nested partitions, whereas partitioning methods produce partitions into a predetermined number of groups (*K*). For *n* objects, most agglomerative clustering algorithms require the computation of a (*n* × *n*) similarity or distance matrix; this is impractical for large data sets like sonar data. Hence, we turned to partitioning methods. *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 (across the groups) of the sums of squared residual distances to the group centroids; this statistic is also called the sum of within-group sums-of-squares, the error sum-of-squares, or the sum of squared errors (SSE). SSE is the global optimality criterion, or objective function, implemented in *K*-means algorithms. Hundreds of algorithms have been proposed in the literature to solve the *K*-means problem.

We implemented the following two-step iterative least-squares algorithm: (*i*) compute cluster centroids and use them as new cluster seeds; and (*ii*) assign each object to the nearest cluster seed. This algorithm is described in several books; for example, Legendre and Legendre (1998).

Since *K*-means is a NP-hard problem (a category of very hard problems in computer 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. (*i*) The program was made to proceed in a cascade, finding first a partition into a number of groups larger than what is needed (e.g., starting at 10 groups). It is easier to find the best partition for a large number than for a smaller number of groups. When this partition has been found, the two groups whose centroids are the closest in multivariate space 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 two groups is found). (*ii*) The whole classification process (e.g., from 10 to two groups) can be repeated a number of times (e.g., 25 or 50 times, as specified by the user) using different random starting configurations. For each number of groups (e.g., for $K = 10$, $K = 9,\ldots,K = 2$ groups), the solution where SSE_K is minimum is retained and written to the output file.

Step 3: how many acoustic classes?

How to decide on the optimal number of acoustic classes? 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 and 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, J.E. Hewitt, S.F. Thrush, P. Legendre, J. Ellis, and M. Morrison (NIWA, P.O. Box 11-115, Hamilton, New Zealand, unpublished data) found cases where the best number of groups was from $K = 2$ to $K = 19$, depending on the data set.

Step 4: other computation modules

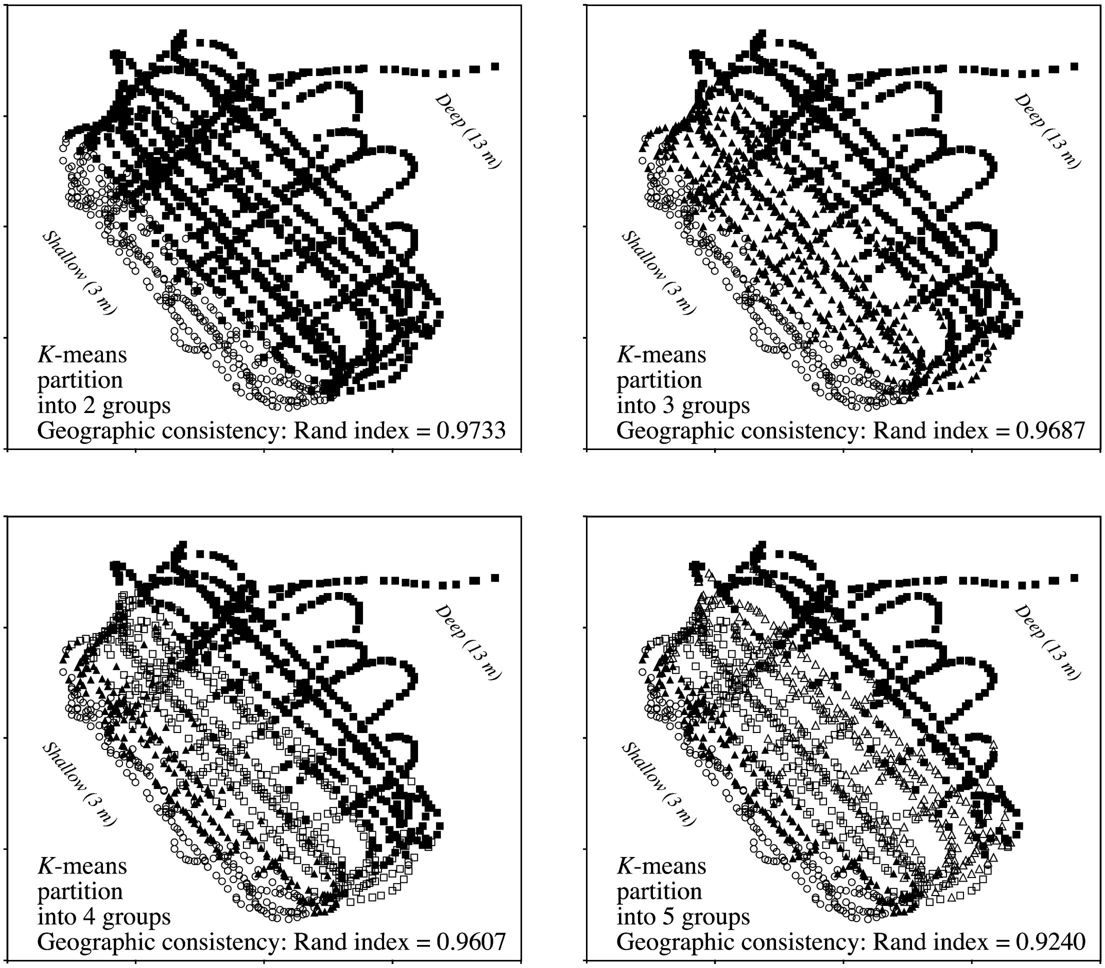
A drawing module allows users to produce simple maps from the *K*-means partitioning results and the geographic coordinates of the individual acoustic records. Figure 2 presents examples of these maps (printed here in black only); they may include colour, symbols, 95% confidence ellipses around groups, etc. The maps can be copied and pasted in one's favourite drawing program and saved as standard EMF (Enhanced MetaFile) format.

Another module of the package computes the "geographic consistency" of the *K*-means solutions. We want to know if the groups obtained by partitioning consist of geographic neighbours; if they do not, we want to know how close they are to a "geographically consistent" solution in which each group would only contain points that are contiguous in space. First, one computes a matrix of geographic contiguity among points, using one of a number of connection networks described, for instance, in Legendre and Legendre (1998). The type of connection most often used is the Delaunay triangulation. Our "Links" module, which can plot the connection network on a map of the data points, is based upon a Delaunay algorithm by Shewchuk (1996). Then, one employs the "GeoConsist" module: using the list of connections between geographic neighbours, this program subdivides each group obtained by *K*-means partitioning into geographically connected subsets of points, using a simplified constrained clustering algorithm (Legendre and Legendre 1998). One obtains a new partition into a larger number of groups that are nested into the groups of the *K*-means partition. The Rand index (Rand 1971) between the original and spatially constrained partitions is computed as an index of geographic consistency. The closer this index is to 1, the greater is the geographic consistency of the original *K*-means solution. Membership of the points in the geographically constrained groups is also available for mapping.

Example data

On 16 August 1999, acoustic data were collected in the Forty Baskets Beach area of Sydney Harbour, Australia (33°48′S, 151°16′E). We used a Navisound 50 echosounder 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 of Quester Tangent Corporation), which was used to decompose the backscatter waves mathematically into 166 variables (Fourier analysis of the response wave, 64 variables; wavelet analysis, 64 variables; 38 other variables describing the shape of the first acoustic backscatter based upon the original and cumulative forms) (Fig. 1). The transducer was mounted on an over-the-side strut on the survey vessel. The positioning equipment was a differential GPS (Global Positioning System). The recorded data were corrected and validated using a "Parser" procedure, which is part of our software. The test data set consisted of 1478 data

Fig. 2. Map of the Forty Baskets Beach sampling area (Sydney Harbour, Australia: 33°48′S, 151°16′E) showing the *K*-means partition of the acoustic records into 2–5 groups (symbols) based upon the first three principal components (96.2% of the variance in the data). These partitions explain 58.4, 76.7, 81.4, and 84.1%, respectively, of the variance in the data. The partition into three groups is the one for which the Calinski-Harabasz (C-H) criterion is maximum.


Deep (13 m)
Shallow (3 m)
K-means partition into 2 groups
Geographic consistency: Rand index = 0.9733
Deep (13 m)
Shallow (3 m)
K-means partition into 3 groups
Geographic consistency: Rand index = 0.9687
Deep (13 m)
Shallow (3 m)
K-means partition into 4 groups
Geographic consistency: Rand index = 0.9607
Deep (13 m)
Shallow (3 m)
K-means partition into 5 groups
Geographic consistency: Rand index = 0.9240

lines (objects, or records) and 166 QTC variables, plus geographic positions and depths. Since three of the QTC variables did not vary at all, they were eliminated from the data set, which was thus reduced to 163 variables.

K-means divided the acoustic data into a series of bands that follow the depth gradient (Fig. 2). Unfortunately, we do not have geographically localized visual observations to validate the classification results, but divers reported that the sediment changed along this gradient and that seagrass formed a bed parallel to the coast. The C-H criterion indicated that the partition into three groups was the best one in the least-squares sense. As a statistical model, this partition explained 79.8% of the variance in the first three PCs, or 76.7% of the variance in the 163 original QTC variables. Acoustic classification results should be subjected to ground truthing, which consists of relating the acoustic classes to visually observed data describing the seabed. J.E. Hewitt, S.F. Thrush, P. Legendre, J. Ellis, and M. Morrison (NIWA, P.O. Box 11-115, Hamilton, New Zealand, unpublished data) have done such a validation study, using underwater video data, of an acoustic seabed classification obtained from QTC variables analysed by our software.

Program and report

A computer package (The Q Package) has been developed, with the financial help of NIWA of New Zealand, to implement the seabed classification method described in this paper and analyze large data sets. In its present state of development, it can handle 10 000 data points in real time and 100 000 points with a small delay, using a recent Windows-based operating system. Any computer capable of running Microsoft Windows 95 or later versions (including Windows NT and Windows 2000) can be used to run the Q Package. A low-end Pentium with 32 Mb of RAM and Windows 95 is powerful enough to run the program, and is perfectly adequate in most cases. The package, which comes complete with a user's manual, is available free of charge at http://www.fas.umontreal.ca/biol/legendre/.

A report, available from the first author, presents a user's comparison of the method described in this paper with that of the QTC VIEW™ CAPS and QTC IMPACT™ software of Quester Tangent Corporation. The report shows that PCA followed by *K*-means partitioning produces statistically better results than the classification method implemented in the QTC software with which we experimented during the SCALE EX-

PERT workshop (Spatial Comparisons Across Large Estuaries: EXPerimental Evaluation of Recent Technologies) organized and hosted by Professor A.J. Underwood at the University of Sydney, Australia, 2–22 August 1999.

Acknowledgements

We are grateful to Professor Underwood for having organized the SCALE EXPERT workshop and for his hospitality. We extend thanks to Jim Drury, Fisheries Biologist at NIWA–Auckland, New Zealand, for demonstration of and help in understanding the QTC IMPACT™ classification program, version 2.00, in March 2001, and to Judi Hewitt, Benthic Ecologist at NIWA–Hamilton (P.O. Box 11-115, Hamilton, New Zealand), who performed numerous comparisons of the QTC IMPACT™ program results with 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 kindly provided comments on a draft of the report.

References

Calinski, T., and Harabasz, J. 1974. A dendrite method for cluster analysis. Commun. Stat. **3**: 1–27.

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

Clarke, P.A., and Hamilton, L.J. 1999. The ABCS program for the analysis of echo sounder returns for acoustic bottom classification. Report DSTO-GD-0215, Maritime Operations Division, Aeronautical and Maritime Research Laboratory (Defence Science & Technology Organisation, Commonwealth of Australia), Melbourne, Australia.

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

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

Prager, B.T., Caughey, D.A., and Poeckert, R.H. 1995. Bottom classification: operational results from QTC View. *In* Oceans 95. MTS/IEEE. Challenges of Our Changing Global Environment Conference Proceedings. Vol. 3. 9–12 Oct. 1995, San Diego, Calif. pp. 1827–1835.

Quester Tangent Corporation (QTC). 1999. CLUSTER operator’s manual. 24 March 1999. Quester Tangent Corporation, Sidney, B.C., Canada.

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

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

Shewchuk, J.R. 1996. Triangle: engineering a 2D quality mesh generator and Delaunay triangulator. *In* First ACM Workshop on Applied Computational Geometry, 27–28 May 1996, Philadelphia, Pa. Association for Computing Machinery, 1515 Broadway, New York, NY 10036, U.S.A.