DISCUSSION / DISCUSSION

Reply to the comment by Preston and Kirlin on "Acoustic seabed classification: improved statistical method"[1]

Pierre Legendre

Legendre et al. (2002) described a statistical method for analysing echosounder backscatter data, which consisted of the following steps: the backscatter data were decomposed mathematically into a number of quantitative variables, which were subjected to principal component analysis (PCA). Principal components representing 95–99% of the variation were then used in a *K*-means partitioning procedure. A least-squares statistical criterion indicated the number of groups that best reflected the variability of the data. The groups were then plotted on maps of the survey area. Insofar as the mathematical decomposition of the backscatter echo produced variables that reflected the variation of the physical nature and composition of the seabed, the classes of the partition were likely to correspond to seabed types. This procedure presented several improvements over the Quester Tangent Corporation's QTC VIEW™ acoustic bottom classification method (Prager et al. 1995), and it was described in easy-to-understand terms. Free software implementing this method — *The Q Package* for Windows and *The R Package* for Macintosh — is available on the Web site http://www.fas.umontreal.ca/biol/legendre/.

J.M. Preston, from the Quester Tangent Corporation (QTC), and R.L. Kirlin wrote a comment on that paper, to which I was invited to reply. I will first address some statistical points in Preston and Kirlin's note and then go to more fundamental issues.

Statistical issues

(1) The *K*-means problem was defined by MacQueen (1967) as that of partitioning a data set in Euclidean space into *K* nonoverlapping groups in such a way as to minimize the sum (across the groups) of the within-group sums of squared Euclidean distances to the respective group centroids. The statistical problem had first been stated by Fisher (1958). What QTC seems to be doing, if I understand their description correctly, is implementing a partitioning method based on Mahalanobis distances for solving the mixture problem, like the one described by Demers et al. (1992), for instance. So they are implementing a modified form of *K*-means partitioning. This does not mean that the results produced by such an algorithm are more meaningful that those of a standard *K*-means algorithm.

The description of the classification algorithm implemented by the QTC software, provided by Preston and Kirlin, is wrapped in Bayesian language. A true Bayesian approach would require that prior probabilities be known beforehand; that is not the case in acoustic seabed classification. What is being described is simply a two-step iterative algorithm, preserving Mahalanobis distances, in which objects are assigned to the clusters and the statistical parameters of the clusters are recomputed.

(2) Preston and Kirlin (2003) criticize the closing statement of our 2002 paper, which read: "The report [Legendre, unpublished report, available at http://www.fas.umontreal.ca/biol/legendre/] shows that PCA followed by *K*-means partitioning produces statistically better results than the classification method implemented in the QTC software…". The sentence should have read: "… statistically better results *in the least-squares sense*…". The statement was based on results presented in the unpublished report and summarized in the following paragraphs.

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 echo sounder (Navitronic Systems AS, Hasselager, Denmark) at frequency 50 kHz (transducer beam width 13.5°) connected to the QTC VIEW™ acoustic seabed classification system (CAPS version 3.25, QTC IMPACT™ ver-

Received 14 July 2003. Accepted 16 October 2003. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 12 November 2003.
J17628

P. Legendre. Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, QC H3C 3J7, Canada (e-mail: pierre.legendre@umontreal.ca).

[1] Reply to J.M. Preston and T.L. Kirlin. 2003. Can. J. Fish. Aquat. Sci. **60**: 1299-1300.

doi: 10.1139/F03-132

sion 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, the data set consisted of 1478 data lines (objects or records) and 166 QTC variables, plus geographic positions and depths. Because three of the QTC variables did not vary at all, they were eliminated from the data set, which was thus reduced to 163 variables. These data were used to illustrate the acoustic seabed classification method described in Legendre et al. (2002).

The first three principal components accounted for 96.2% of the variance in the QTC variables. Using seven principal components would have accounted for 99.2% of the variance. For fairness of comparison, I only used the first three principal components in the comparison of partitions, because this is what the QTC software uses. The data were subdivided into groups using the procedure outlined in the QTC manuals (QTC 1999, 2000). The score value was used to determine which class should be split next. As the classes were subdivided, the total score decreased; however, at split level seven (i.e., eight groups), the total score increased again. Results of the partitions into three to seven groups are reported in Table 1*a*. In an a posteriori calculation, the Calinski–Harabasz statistic (described below) selected the partition into five groups as the best one in the least-squares sense.

The same data were partitioned by *K*-means, using the first three principal components (PC1–PC3), as in the QTC procedure. The *K*-means program was asked to produce from ten to two groups; the partitioning was restarted 10 times. The best partitions into $K = 2$ to $K = 7$ groups were retained; some of these results are shown in Table 1*b*. The Calinski–Harabasz statistic (see below) selected the partition into three groups as the best one. We also computed *K*-means partitioning for all 163 QTC variables, without prior filtering by PCA. The Calinski–Harabasz statistic selected again the partition into three groups as the best one (Table 1*c*). This partition is very similar to that obtained by *K*-means on PC1–PC3.

The partitioning procedure described in the QTC manuals (QTC 1999, 2000), which we used in 1999, is not the one described by Preston and Kirlin in their Comment (see Seabed classification issues, subsection 1, below). The partitioning results of QTC IMPACT™ and *K*-means are compared in Table 1 using common measures based on least squares: the sum of within-group sums-of-squares, also called the "sum of squared errors" (SSE), and the Calinski–Harabasz statistic (C–H) (see Legendre et al. 2002 (Classification method, step 3), as well as Seabed classification issues, subsection 4, below). SSE is the global optimality criterion implemented in *K*-means algorithms. C–H is a statistical criterion indicating the best number of groups in the least-squares sense. Least-squares is a widely accepted criterion and has a long history in statistics (Legendre 1805).

By the SSE criterion, Table 1 shows that the QTC partition into three groups is not as good as the *K*-means partitions into three groups based on either the first three principal components or all 163 variables, with respect to either the first three principal components (32% larger SSE) or the 163 QTC variables (27% larger SSE). Likewise for the partitions into five groups: according to SSE, the QTC solution into five groups is much worse than the *K*-means solutions based on either PC1–PC3 or all 163 QTC variables, with respect to either the first three principal components (15% larger SSE) or the 163 variables (10% larger SSE). This shows that the QTC partitions (even the "best one" into five groups) were far from being as good, for this example and in the least-squares sense, as those obtained by *K*-means.

(3) Preston and Kirlin (2003, their paragraph 3) talk about performing an (undescribed) test of statistical significance in their partitioning method. It is not clear to what they are referring. In any case, there is nothing that can be tested for significance in *K*-means or Mahalanobis distance partitioning without invoking an external, independently obtained data set. In particular, the results of a partitioning procedure should not be tested for significance using the same data that were used to produce the partition. This would be a logical mistake, as explained by Milligan (1996, p. 366) and Legendre and Legendre (1998, p. 379).

(4) Preston and Kirlin (2003) state that "Legendre et al. (2002) and QTC both cluster in three-dimensional space". This is not what we wrote; see Legendre et al. (2002, abstract and step 1 of the Classification method). What we recommended was to use as many principal components as were necessary to explain at least 95%, and preferably 99%, of the variance of the data. In the example presented in that paper, three principal components accounted for 96.2% of the variance, so we used three for *K*-means partitioning. We also reported that in the analysis of other acoustic seabed data sets, 3–5 PCs were necessary to reach 95% of the variance and 6–10 to reach 99%. In subsection 2 above, I only used three principal components in the tables to present a fair comparison with the QTC IMPACT™ results, which are limited to three principal components by design of the program. Preston et al. (2001) present that as a feature of the QTC software, but I think it is an objectionable limitation. This point will be revisited below.

(5) Preston and Kirlin (2003, last paragraph) concluded by citing a number of papers that allegedly showed that the QTC classification method "has repeatedly been found to give practical, useful, and accurate classes". It is worth noting that the paper by Morrison et al. (2001), cited by Preston and Kirlin, was directed at developing a technique to identify habitat boundaries. Their analysis compared the accuracy of the transitions predicted by the QTC VIEW™ confidence values with those predicted by the class-dominance Berge–Parker statistic. Morrison et al. (2001) concluded that the Berge–Parker index provided a more consistent transition indicator than the QTC software confidence values.

Seabed classification issues

(1) People have long wondered what was the classification

Table 1. Comparison of partitions (least-squares statistics SSE and C–H) using two different bases: PC1–PC3 (middle) and 163 QTC variables (right).

	No. groups	Base: PC1–PC3		Base: 163 QTC variables	
Software and variables	*K*	SSE	C–H	SSE	C–H
(*a*)					
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
(*b*)					
K-means (PC1–PC3)	3*	41.86	2904*	49.67	2453*
K-means (PC1–PC3)	5	25.97	2561	33.53	1992
(*c*)					
K-means (163 QTC var.)	3*	41.87	2903*	49.67	2453*
K-means (163 QTC var.)	5	25.97	2563	33.51	1993

Note: SSE, sum of within-group sums-of-squares (small is best, for a given number of groups); C–H, Calinski–Harabasz statistic (high is best among partitions obtained using the same data). Asterisk (*) indicates the best number of groups for that classification, according to C–H.

method used by QTC. We asked questions to that effect to the Quester Tangent Corporation in 1999 but received no answer. When they use a method, scientists need to know, and be able to describe in scientific papers, how the data are processed. However, there seems to be greater openness and QTC is now providing more details of their statistical procedure.

The Comment by Preston and Kirlin (2003) is, to my knowledge, the first published paper describing the fact that the QTC software now uses a two-step iterative algorithm preserving Mahalanobis distances. Only now are we learning from Preston and Kirlin (2003) that the QTC software is using Mahalanobis distances instead of Euclidean distances. It is not clear from Preston and Kirlin's Comment whether they state that this approach has been implemented all along the development of the QTC proprietary software, or not. In any case, the procedure used in the 1999 CLUSTER™ and the 2000 QTC IMPACT™ programs and described in the software manuals (QTC 1999, 2000) was one-dimensional (the clusters were split along the first principal component in earlier versions of the software and along a single principal component chosen by the user in the 2000 version). An evolution of the software is summarized in Preston et al. (2001), who simply described it as “an automated variant of the k-means clustering method. Clustering is done in the space of the three principal components and is iterative and stable.” What is now described in the Preston and Kirlin (2003) Comment seems to be a new evolution of the QTC software, recently released or perhaps still in the testing phase. In earlier manuals provided by QTC with their statistical software (QTC 1999, 2000), there was no indication about the nature of the calculations leading to partitioning, except for the description of how to use it, which implied a lot of fiddling and left room for personal, unreplicable decisions on the part of the user.

I also applaud the fact that QTC has recently (Preston et al. 2001) released some information about the mathematical nature of the 166 variables produced by the QTC VIEW™ software. The information found in Preston et al. (2001) provides a general qualitative overview of the variables generated by QTC VIEW™ from the backscatter, but the methods by which they are derived and the quantitative nature of the information remain unexplained.

(2) In the summer of 1999, Hewitt et al. (J.E. Hewitt, National Institute of Water and Atmospheric Research, P.O. Box 11-115, Hamilton, New Zealand, unpublished data) carried out a multiresolution nested survey in Kawau Bay, located on the northeast coast of North Island, New Zealand. The spatial distribution of epibenthic communities was studied using side-scan sonar, single-beam sonar, and video. The objective was to find relationships between assemblages visible from the video and the single-beam and (or) side-scan data that would enable the researchers to use these devices to both interpolate between and extrapolate from the restricted video survey. The substrate was soft sediment in all eight 1-km^2 sites investigated. There were reasonably dense but patchy epibenthic communities. At each site, six pairs of 1-km-long transects were sampled with single-beam sonar. The transects ran down the depth gradients. Three of the eight sites could not be videoed because of the presence of shoals and subsea cables. However at the other five sites, three 1-km-long video transects were run in approximately the same positions as three of the single-beam transects. The sonar was a Simrad EA501P hydrographic sounder (Simrad AS, Horten, Norway), attached to the boat, and operated at 200 kHz, 250 W transmit power, with a ping rate of 5 s^{-1}, and a fixed beam width of 7°. This was connected to a QTC VIEW™ series 4 (Collins et al. 1996) data acquisition system. Settings for the QTC VIEW™ system were a reference depth of 14 m and a base gain of 15 dB. Sampling resolution varied from 0.37 to 3.0 m^2, depending on depth, although more generally the range was from 1.22 to 2.44 m^2. As QTC

VIEW™ uses a stack of five consecutive pings for each record, at our speed of about 5 knots (2.6 m·s^{-1}), a ping stack (generally covering 6–12 m) was processed approximately every 8 m.

The single-beam sonar data were processed as follows: on the one hand, using the first three principal components computed by the proprietary principal component analysis procedures available in the program QTC IMPACT™ (QTC 2000), cluster splits were made in the principal component data scatter until further splits failed to reduce the overall variance in an important way. Splitting decisions were made as detailed in Morrison et al. (2001) using inflexion points of the total scores and the QTC cluster performance index. On the other hand, principal component analysis was applied to the 166 variables produced by the QTC VIEW™ system using the data from all five sites; the number of principal components required to explain 95% of the variance was five. Ping scores along those five axes were used in the *K*-means partitioning procedure of *The Q Package* freeware. The Calinski–Harabasz statistic was used as a stopping criterion to determine the best number of groups for each data set, in the least-squares sense.

The QTC IMPACT™ classification into seven groups was more related to depth than was the six-group *K*-means classification. Discriminant analysis identified that 53% of the points could be allocated correctly to the QTC IMPACT™ groups based on depth alone, compared with 34% for the *K*-means classification. As a consequence, the rest of the analyses reported by Hewitt et al. (J.E. Hewitt, National Institute of Water and Atmospheric Research, P.O. Box 11-115, Hamilton, New Zealand, unpublished data) were based on the *K*-means classification, on which depth had less of an influence.

The first principal component of the 166 QTC variables was highly correlated with depth (Spearman's $r = -0.91$). The relationship found between depth and partition was not totally avoidable in that study because the transects ran down depth gradients and the size of the sonar footprint is a function of depth. Obviously, the fact that QTC IMPACT™ only uses three axes in determining its partition makes it especially sensitive to depth. By opposition, Legendre et al.'s (2002) *K*-means partitioning uses the number of axes necessary to explain 95% (or 99% in other studies) of the variability in the data; that explains the differences between the results of the two methods. There is certainly an advantage in using more than three principal components as the basis for classification.

(3) Preston and Kirlin (2003) argue that elongated (hyperellipsoidal) clusters, produced by their Mahalanobis-based clustering method, are more natural than, and thus preferable to, the hyperspherical clusters produced by *K*-means. There is no particular reason why the data points (sonar backscatters, decomposed into QTC-generated variables) should be structured in any particular way in multivariate space, or in a reduced space of principal components. Within the range of variation of the 166 QTC variables, any intermediate value is possible, so that observations may be found anywhere within the convex envelope surrounding the data points in multivariate space. Natural separation of clusters is predicted, for instance, by the theory of biological evolution, which was the starting point for the development of many of the methods of numerical classification (Sokal and Sneath 1963), but I do not think any theory predicts the existence of regions occupied by points, in the space of acoustic variables, separated by regions where no observations are possible. Nor do we have a theory that predicts that the clusters should have any particular shape. We only want to empirically divide the sonar backscatters into groups, to simplify their multivariate description. These groups will be useful if they are found to correspond to characteristics of the seabed. A division of the space into multivariate boxes of equal sizes would produce a perfectly good classification of the seabed. This can easily be done, for example, in the two-dimensional space of RoxAnn™ variables E1 and E2 developed by Marine Micro Systems Ltd. (Chivers et al. 1990; E1 and E2 are often referred to as "hardness" and "roughness"), but it would be impractical to attempt doing it in a space with 166 dimensions. The reason that we resort to partitioning methods in that space is because we only need to define boxes that are occupied by a sufficient number of points; in particular, we do not want boxes (or classes) containing no point at all. So it is perfectly reasonable to look for spherical or hyperspherical *K*-means clusters in that space. Because they have borders forming flat interfaces, their shape is actually hyperpolyhedric rather than hyperspherical. On these grounds, any other cluster shape is also perfectly acceptable, including those produced by the Mahalanobis-based QTC algorithm. In summary, I believe that the particular metric used for partitioning is not a key point to insure a useful partition of the data points. The choice of the variables derived from the sonar backscatter, and the number of principal components retained for partitioning (see above), are much more important.

(4) There are computational and statistical reasons to prefer a solution implementing a least-squares criterion. The first one is that it runs faster than a Mahalanobis-based algorithm. So, the available computing time can be used to try more random starts of the algorithm; this, in turn, increases the chances of finding an optimal partition (in the least-squares sense). The second reason is that least squares go well with least squares: because *K*-means is a partitioning method optimizing the least-squares criterion, as in multiple regression, it allows the application of a criterion based on least squares to determine the "best" number of clusters. In the *K*-means algorithm incorporated in *The Q Package* and *The R Package*, we are using the Calinski–Harabasz (1974) criterion (C–H), which is a least-squares criterion. This criterion is simply an *F* statistic of multivariate analysis of variance; the partition displaying the highest value of this criterion is the best one in the least-squares sense. Actually, the partitions corresponding to the various maxima of the graph of the C–H statistics against the number of

groups may be worth examining and mapping. There is no direct equivalent of this criterion in Mahalanobis space, at least none that I know of. I encourage the Quester Tangent Corporation to offer our procedure as an option in their computer package. We published our procedure in the scientific literature (Legendre et al. 2002), so it is not proprietary.

(5) The relative usefulness of the partitions produced by QTC VIEW™ and our software should be judged by ground truthing. Because our *K*-means software is freely available, users in governmental and private research institutions, as well as universities, should be encouraged to analyse QTC data using the QTC VIEW™ classification software (based on three PCA axes) and our *K*-means software (using a sufficient number of PCA axes to account for 95% or 99% of the variation in the QTC data) and to compare the results to ground-truthing data, as was done by Hewitt et al. (unpublished data; see subsection 2 above). Comparisons of this kind are part of the scientific process needed to determine the usefulness of acoustics for mapping features of the seabed and, in particular, the horizontal distribution of ecological communities.

Conclusion

This debate does point to a more general dilemma, where technological innovation leads to proprietary products that are used and should be scrutinized by the scientific community. Although science is supposed to be based on free and open communication and debate, companies may choose to act differently in commercial activities. There is clearly a need for new and cost-effective surveying devices that enable us to image and map large areas of the seafloor in a routine and timely fashion. Every surveying device has its limitation and it is important that we recognize them so that technologies can continue to be developed and improved. This note is meant to contribute to that process. I also hope that our paper (Legendre et al. 2002) and freeware (*The Q Package* for Windows and *The R Package* for Macintosh) will be used for seafloor management, especially by researchers in underdeveloped countries, who can benefit from freely available software implementing sound statistical procedures.

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.

Collins, W., Gregory, R., and Anderson, J. 1996. A digital approach to seabed classification. Sea Technol. **37**: 83–87.

Demers, S., Kim, J., Legendre, P., and Legendre, L. 1992. Analyzing multivariate flow cytometric data in aquatic sciences. Cytometry, **13**: 291–298.

Fisher, W.D. 1958. On grouping for maximum homogeneity. J. Am. Statist. Assoc. **53**: 789–798.

Legendre, A.M. 1805. Nouvelles méthodes pour la détermination des orbites des comètes. Courcier, Paris.

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

Legendre, P., Ellingsen, K.E., Bjørnbom, E., and Casgrain, P. 2002. Acoustic seabed classification: improved statistical method. Can. J. Fish. Aquat. Sci. **59**: 1085–1089.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *In* Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. *Edited by* L.M. Le Cam and J. Neyman. University of California Press, Berkeley. pp. 281–297.

Milligan, G.W. 1996. Clustering validation — results and implications for applied analyses. *In* Clustering and classification. *Edited by* P. Arabie, L.J. Hubert, and G. De Soete. World Scientific Publ. Co., River Edge, N.J. pp. 341–375.

Morrison, M.A., Thrush, S.F., and Budd, R. 2001. Detection of acoustic class boundaries in soft sediment systems using the seafloor acoustic discrimination system QTC VIEW. J. Sea Res. **46**: 233–243.

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

Preston, J.M., and Kirlin, T.L. 2003. Comment on “Acoustic seabed classification: improved statistical method”. Can. J. Fish. Aquat. Sci. **60**: 1299–1300.

Preston, J.M., Christney, A.C., Bloomer, S.F., and Beaudet, I.L. 2001. Seabed classification of multibeam sonar images. *In* Proceedings of MTS/IEEE Oceans 2001: An Ocean Odyssey, Honolulu, Hawaii, Nov. 2001. Holland Publications, Escondido, Calif.

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

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.

Sokal, R.R., and Sneath, P.H.A. 1963. Principles of numerical taxonomy. W.H. Freeman, San Francisco, Calif.