Northwest Atlantic Fisheries Organization

Serial No. N611

NAFO SCR Doc. 82/IX/102

FOURTH ANNUAL MEETING - SEPTEMBER 1982

Pattern Recognition: Partitioning in Morphological Hyperspace

by

J.M. McGlade

Dept. of Fisheries & Oceans, Fisheries Research Branch
P.O. Box 1006, Dartmouth, Canada B2Y 4A2

Introduction

The study of fish stocks invariably demands that a sample be partitioned according to preestablished criteria such as reproductive isolation. However, sub-populations and populations often mix during part of their life-cycle, before separating out to spawn. Indeed amongst marine species of fish, where exact locations of spawning are generally unknown, sampling under mixed conditions is likely to be the rule rather than the exception. In essence such a sample is unlabelled. To make an a priori assumption about its composition would therefore be inappropriate.

Under such constraints unsupervized procedures are clearly warranted. These are analyses which make no classificatory assumptions and include procedures such as principal component analysis and maximum likelihood estimation. Based on this type of analysis, a classifier can be designed on a small set of samples, and then "finely tuned" on a large unlabelled set once the classification criteria are established. It is the construction of these classification criteria that will be discussed herein. Beginning with a very restrictive set of assumptions the paper goes on to a reformulation of the problem of classification as one of partioning the data into subgroups or clusters.

Unsupervised Learning

The process of unsupervised learning begins with a set of assumptions as follows:

1) The samples come from a known number of groups (c).

2) The a priori probabilities for each group are known

$P(w_j)$ for $j = 1.....,c.$

3) The group probability densities $p(x|w_j, \theta_j)$ are known for

$j = 1.....,c.$

4) The values for the c parameter vectors $\theta_1...,)c$ are unknown.

Under these conditions a series of samples can be drawn from a mixture to estimate the unknown parameter vector θ. Once θ is known each sample can be broken down into its components. However in fisheries biology a recurring problem is that the number of groups or clusters is unknown.

One solution, although informal, is to extremize a criterion function and repeat the clustering procedure to see how the function changes as c increases. For example a sum-of-squared-error criterion J_e would decrease monotonically as c is increased by transferring a single sample to an original cluster. If n samples are grouped into c discrete clusters J_e would decrease rapidly until $c=\hat{c}$, and then move slowly to reach zero until c=n. Large disparities in the levels at which clusters merge would thus indicate "natural" groupings.

Validation Procedures

From a theoretical standpoint a goodness-of-fit measure, for example a chi-square or Kolmogorov-Smirnov statistic would be an appropriate method of validation. However the dimensionality of most morphometric data sets precludes the use of these measures, and indeed demands a simpler approach such as the criterion function described above.

To discover what constitutes a significant improvement in J(c) a null hypothesis can be set up, involving a decision procedure to accept or reject the sampling distribution $J_{(c+1)}$. It is of course difficult to do anything more than estimate the sampling distribution of $J_{(c+1)}$, however an approximate analysis for a simple sum-of-squared-error criterion has been given by Duda and Hart (1977). They show that for large numbers of samples, the sum of squared error for a partition which minimizes $J_e(2)$ ralative to $J_e(1)$ is approximately normal. A critical value for

$J_e(2)$ can be obtained by assuming that the suboptimal partition is nearly optimal. The variance can be estimated by

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{x \in \mathcal{X}} ||x-m||^2 = \frac{1}{nd} J_e(2)$$

where m is the mean of n samples, $\mathcal{X}$ is a subset of samples and d is the distance between clusters. The null hypothesis is rejected at the p-percent significance level when

$$\frac{J_e(2)}{J_e(1)} < 1 - \frac{2}{\pi d} - \alpha\sqrt{\frac{2(1-8/\pi^2 d)}{nd}}$$

where is given by

$$p = 100 \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

Thi criterion may thus be used as a test to decide whether the partitioning of a sample is justified.

Multivariate Morphometrics

Morphometric characters have long been used in fisheries biology to discriminate between not only species but also populations (Ihssen et al. 1981). Ideally all the information contained within such a set of data should be combined to produce a classification criterion that can produce matching a priori and a posteriori groupings. Clearly multivariate analyses are necessary to achieve this end, yet there are certain logical problems associated with their use. For example, discriminant function analysis, which requires a priori assignment of individuals to groups, produces functions which are highly sample-dependent (Humphries et al. 1981: McGlade 1981).

Principal component analysis however assumes no groups and is thus likely to be a heuristic device in seeking morphological groupings. The eigenvalues can thus be used as classification criteria for partitioning subsequent samples. Yet confounding factors such as size-related changes in shape must be accounted for prior to partitioning. In many studies the first principal component has been considered one of size (Lee 1971; Kuhry and Marcus 1977), whilst the second and subsequent componentns represent

shape (Pimental 1979). This arbitrary division is not clearly justified, and may simply be an artifiact of the onthogonality of the components. An alternative has been put forward by Humphries et al. (1981) in which the loadings on each component are computed from group-free variables i.e. size is the component "whose loadings ... are a linear combination whose coefficients are its own pooled covariances with group", whereas shape is a "linear combination whose coefficients are equal to [the] partial covariance with [the] log-distance measures controlled for intragroup size". If these procedures are adopted shape criteria become readily available as classification criteria which may then be used to test further partitioning of the morphological hyperspace.

Conclusions

Partitioning samples of fish taken at random from a population or number of stocks requires a classification criterion and a test for its validity. Morphological variables, which are traditionally used in fisheries biology, must however be adjusted for size related differences. The subsequent derivation of classification criteria should not be based on a priori assumptions of group-relatedness, but rather on unsupervised learning techniques. One such example is principal component analysis.

Classification criteria, such as the eigenvalues obtained from a principal component analysis can be validated under a null hypothesis in which the sum-of-squared-error is observed in conjunction with successive partitioning.

References

Duda, R.D. and P.E. Hart. 1977. Pattern classification and scene analysis. John Wiley & Sons, New York.

Humphries, J.M., F.L. Bookstein, B. Chernoff, G.R. Smith, P.L. Elder and S.G. Poss. 1981. Multivariate discrimination by shape in relation to size. Syst. Zool. 30: 291-308.

Ihssen, P.E., H.E. Bodre, J.M. Casselman, J.M. McGlade, N.R. Payne and F.M. Ulter. 1981. Stock identification: materials and methods. Can. J. Aquat. Fish. Sci. 38: 1838-1855.

Kuhry, B., and L.F. Marcus. 1977. Bivariate linear models in biometry. Syst. Zool. 26: 201-209.

Lee, P.J. 1971. Multivariate analysis for the fisheries biology. Fish. Res. Board Can. Tech. Report 244: 1-182.

McGlade, J.M. 1981. Genotypic and phenotypic variation in the brook trout, Salvelinus fontinalis (Mitchill). Ph.D Thesis, University of Guelph, Guelph, Ontario.

Pimental, R.A. 1979. Morphometics: the multivariate analysis of biological data. Kendall/Hunt Publishing Col, Dubuque, Iowa.