NOT TO BE CITED WITHOUT PRIOR
REFERENCE TO THE AUTHOR(S)

Northwest Atlantic Fisheries Organization

Serial No. N616

NAFO SCR Doc. 82/IX/112

FOURTH ANNUAL MEETING - SEPTEMBER 1982

Introducing a new discriminant analysis (with covariance) and multivariate analysis of covariance, with a study of beaked redfishes Sebastes mentella and S. fasciatus

by

R. K. Misra
Fisheries Research Branch, Department of Fisheries and Oceans,
P. O. Box 550, Halifax, Nova Scotia B3J 2S7

I-H. Ni
Fisheries Research Branch, Department of Fisheries and Oceans,
P. O. Box 5667, St. John's, Newfoundland A1C 5X1

ABSTRACT

The analysis of morphometric data for fish species differentiation and stock discrimination has frequently been unsatisfactory due to sampling bias associated with the varying size of specimens and the large overlapping of characters. These difficulties may be overcome by employing discriminant function with covariance and multivariate analysis of covariance. In this paper, (1) these methodologies are introduced in a classification study of beaked redfishes, in which the specimens of Sebastes fasciatus are smaller than those of S. mentella. Discriminant function with covariance provided a more effective discrimination between species/populations than one without covariance. (2) It is demonstrated that employing a large number of characters in discriminant analysis may not be appropriate. (3) It is explained why expressing morphometric measurements as ratios, proportions, or percentages of body length may not be an appropriate way of reducing variation owing to size differences. Presentation of analysis includes discussion of intermediate results, which are not easily accessible even though these details are often of interest to users. Seven morphometric characters were identified as pertinent discriminators between S. fasciatus and S. mentella. Discriminant function separated the two species remarkably well; as much as 89% of the total variation in the sample was accounted for by the discriminant function and only 8 out of 198 individuals (i.e. 4%) were in the zone of uncertainty.

STOCK DISCRIMINATION SYMPOSIUM

Introduction

In morphological studies of beaked redfishes, Ni (1981a and b) pointed out that the effective management of redfish resources in the Northwest Atlantic requires a clear understanding of Sebastes spp. composition and stock units. However, for decades the distinction between S. mentella and S. fasciatus has not been clearly established. Although the distinction can be sustained on the basis of one anatomical character, the extrinsic gasbladder musculature (Ni 1981a), examination of this character is time-consuming and not practical in field studies. In his discriminant analysis, Ni (1981b) employed data on meristic and nominal characters only and reported that discriminant analysis was remarkably effective in separating the two species and for identifying good discriminators to be utilized in field studies. However, he did not apply discriminant analysis to morphometric data because the specimens of S. fasciatus were smaller than those of S. mentella and the range of characters overlapped broadly between groups. This difficultly is frequently encountered in morphological studies of species differentiation and stock discrimination in fish. ᵀt is, therefore, submitted that use of discriminant function with covariance will overcome this difficultly.

A survey of the applied research in social, behavioural, business and medical sciences would indicate that the use of Fisher's discriminant function has been extensive by the most conservative standards (Goldstein and Dillon 1978). "Even when two similar species can be identified with a single measurement, a combined criterion of two or more may increase the separation between them" (Bliss 1970). In this analysis of redfish morphometrics it was observed that a single character would not separate species effectively. But a compound criterion (discriminant function) of several characters separated the species effectively and made identification possible from morphometric data. Discriminant analysis would be particularly appropriate when the existence of reference samples can be assumed on the basis of an external criterion (Kendall and Stuart 1976), as in this study of redfish data where reference samples were formed on the basis of the extrinsic gasbladder musculature (Ni 1981a). Multivariate normal distribution is a required condition for Fisher's linear discriminant function (LDF) to yield optimal assignment rule (Dillon 1979). Performance of LDF in non-normal situations can be very misleading (Lachenbruch et al. 1973, Dillon 1979). Morphometric measurements are taken on "continuous" variables and are far more appropriate for discriminant analysis as multivariate normality is closely approximated by their logarithms (Piementel 1979, p. 57; Bliss 1967, p. 115). Individuals vary within populations, as in standard length, to warrant its correction and to employ a discriminant function which is adjusted by covariance (Bliss 1970). The methodology is documented in statistical literature (see e.g. Bliss 1970) but, as far as the authors are aware, it has not previously been applied to fisheries data. A valid discriminant analysis must be preceded by a significant difference between population mean vectors (Piementel 1979). This may be tested by multivariate analysis of covariance (MANCOVA). As far as the authors are aware this methodology also has not been used for comparing fish populations, although the utility of ANCOVA at the univariate level has long been recognized. For example, Marr (1955) remarked that the analysis of ratios is inefficient as opposed to regression analysis of original variates and Royce (1964) preferred regression analyis to ratios or indices in order to control the effect of size of fish in his comparisons.

In this paper, (1) methodologies of MANCOVA and discriminant analysis with covariance are introduced and applied to a classification study of beaked redfishes based on morphometric data. (2) It is demonstrated that parsimony in the number of characters to be included in discriminant analysis is desirable. (3) It is explained why expressing morphometric measurements of characters as ratios, proportions, or percentages of body length may not be an appropriate way of reducing variation owing to size differences. The presentation also includes discussion of intermediate results which are not easily accessible even though these details are often of interest to users.

Materials and Methods

Morphometric data of the 200 beaked redfish specimens described by Ni (1981a and b) were employed in the present study. These specimens were separated into two groups on the basis of the extrinsic gasbladder musculature prior to discriminant analysis. Each S. fasciatus or S. mentella group consisted of 100 specimens. All specimens were frozen after capture and thawed prior to measurement. There were twelve morphometric characters examined (Table 1), most of which were suggested by Barsukov and Zakharov (1972) and Barsukov (1972). All mensural characters were measured with calipers to the nearest 0.1 mm except standard length which was rounded to the nearest 1 mm. Body weight was recorded to the nearest gm. Head length and preanal length were measured from anterior part of upper jaw to the relevant posterior point on a line parallel to the main axis of the fish.

Statistical methodology was organized and computer program was written in FORTRAN by one of us (RKM).

Statistical Analysis and Results

It is often believed that the effect of size differences in population comparisons can be eliminated by expressing measurements as ratios, proportions, or percentages of body length. For examples, in their examination of morphometric data for the evidence of stock discreteness, Casselman et al (1981) expressed body measurements as ratios of body legth to reduce variations of fish size within each sample. Such relative values were also employed with a similar objective in a number of stock discrimination studies presented at the fourth annual meeting, September 1982, of NAFO. This use of ratios has been criticized (see e.g. Blackith and Reyment 1971, p. 27). The following would demonstrate that this may not be an appropriate procedure: Consider, e.g. two variables X (body size such as standard length) and Y (a morphometric character) related by an equation of simple allometry,

$$Y = aX^b \qquad (1)$$

where a and b are constants. Equation (1) shows that Y is functionally related to X and requires adjustment in its value for the effect of X. Functional relationship fcr the ratio, Y/X, would then be $Y/X = aX^{(b-1)}$ which is of the same form as (1). Thus, ratio of Y (or percentage, which is only a constant, viz. 100, times the ratio) is affected by X just as Y itself is, barring the special case when $b = 1$. In fact, a statistical analysis of ratio of Y would very likely be more questionable than the analysis of Y itself, since additional problems prevail with ratio data. For example, ratios have unusual distributions and are subject to various errors (Pimental 1979, p. 60). The argument against appropriateness of analyzing ratio data would always hold when X and Y are correlated, even if not related by allometry. For example, a simple linear regression $Y = a+bX$ for Y leads to the equation $Y/X = a/X + b$ for its ratio, which shows that the ratio is still not independent of X.

All measurements were transformed to common logarithms for MANCOVA and discriminant analysis for the following reasons. (1) Multivariate normality is more closely approximated by logarithms than by the original variables

(Pimentel 1979, p. 57; Bliss 1967, p. 115). (2) MANCOVA adjustments generally assume linear relationships and this assumption is also made in the present analysis. Logarithmic transformation should satisfy test of linearity (Pimentel 1979, p. 60, p. 182). (3) The convention is to use common logarithm (Pimentel 1957, p. 57). Only complete specimens i.e., specimens for which all twelve measurements were available, were used in statistical analyses. "Missing observations virtually destroy morphometrics" (Pimentel 1979, p. 191). Samples from S. mentella and S. fasciatus had 97 and 99 complete specimens, respectively. Morphometric characters listed were designated Y_i, i = 1, 2, . . ., 12. Table 1 gives means and ranges of Y_i.

"A series of univariate statistical analyses carried out separately for each of the variables is, in general, not adequate as it ignores the correlations among variables" (Kshirsagar 1972). Following Bliss (1970, p. 329 and 332), the following was noted: Ranges (Table 1) of characters overlapped between species, from 39% (Y_2) to 60% (Y_{11}) in S. mentella and from 61% (Y_4) to 98% (Y_{11}) in S. fasciatus. With these large overlaps no single character would separate species effectively. The probability of misclassification based on Lubischew's coefficient of separation was large, varying from 7.5% (for Y_6) to 18.2% (for Y_9). Yet univariate analysis of variance (ANOVA) to test null hypotheses of equality of means indicated that the difference between species in means was highly significant (probability level at $p<0.001$) for each character. It was, therefore, desirable to find a compound criterion (discriminant function) of characters which would make identification possible from several measurements.

However, discriminant analysis is valid only if populations differ significantly in their means (Pimentel 1979, p. 188). Discriminant analysis was therefore preceded by MANCOVA. Standard length (Y_{12}) was employed as covariate. The following were noted (Bliss 1970, Morrison 1976). In the general linear model of MANCOVA the within-sample linear regression of each variate Y_i on the covariate is incorporated. Sample mean vectors $\underline{Y}$ are thereby adjusted for inequalities in standard length. Incidentally, large overlaps of 46% in S. mentella and 67% in S. fasciatus in the ranges (Table 1) of Y_{12} and its inadequacy to discriminate between the two species effectively, remarked earlier in the text, provided additional support for qualifying standard length as a reasonable covariate (Snedecor and Cochran 1967, p. 430). MANCOVA model assumes that populations do not differ in their regression model and hence, utilizes the matrix B of computed regression coefficients from "within-sample" sums of squares and products (SS and SP) matrices.

The null hypothesis of no difference between species in slopes (weighted by SS of the covariate) of individual variates was, therefore, tested and accepted ($p>0.05$). When the combined slope of a Y_i differs significantly from zero, the residual variation in Y_1 about each sample regression will be less than that around the respective means of individual samples with the covariate ignored, thereby leading to a more effective comparison of means and discriminant function based on covariance procedure. The null hypothesis $B = 0$ of no linear regression of variates on the covariate was tested by the union intersection procedure. The test statistic Θ was 0.9889 with values 1, 4.5 and 188.5 of parameters s, m, and n, respectively (Morrison 1976, Section 5.4), leading to the rejection of the null hypothesis ($p\tilde{}0.001$). The null hypothsis of equal vectors of adjusted means was next tested (Morrison 1976, Section 5.4). The test statistic Θ was 0.7243 with the same values of s, m, and n. The null hypothsis was rejected in favour of the conclusion that the two species differ in mean values of one or more variates independent of the difference between them in standard length.

Before proceeding with discriminant analysis it was considered desirable to investigate which of the eleven variates, if any, did not contribute to the difference between species in the MANCOVA, with the aim of omitting them from discriminant analysis, for the following reason. Discriminant analysis has a close analogy to multiple regression with many stages of calculation parallel to those for a multiple regression but with X and Y reversed (Bliss 1970, Kshirsagar 1972). The expected value of R^2, coefficient of multiple correlation squared, is proportional to the number of variates (Morrison 1976, p. 108). This implies that for samples of limited sizes choosing a large number of characters in discriminant function would artificially inflate its discriminatory power. Parsimony in the number of variates should, therefore, be exercised. Needless to say that working with a smaller number of discriminators also makes discriminant function that much more convenient to employ in field studies. Following Morrison (1976, Section 5.5), 95% simultaneous confidence intervals for characters were estimated. The hypothesis of no significant difference between two species in the adjusted means of Y_3 (snout length) and Y_{10} (width of caudal peduncle) was accepted, as their confidence intervals included zero. Discriminant analysis was therefore done for variates Y_1, Y_2, Y_4, Y_5, Y_6, Y_7, Y_8, Y_9, and Y_{11} only, with Y_{12}

as the covariate. The discriminant analysis methodology employed here was taken mainly from Bliss (1970 Chapters 18 and 20). The methodology combines discriminant analysis with analysis of covariance by adjusting variates by means of their within-sample regressions on the covariates and then finding a compound response (Z) of adjusted variates which would measure best the difference between two species. Species were qualified by values +1 and -1 of "dummy variate" X and discriminant coefficients computed so as to maximize ratio of Z to its standard error (SE). This Z may be expressed as

$$Z = \sum_i L_i Y_i - \sum_i L_i b_{i,12} d \text{ , } i=1, 2, 4, 5, 6, 7, 8, 9, 11$$

where

L_i are discriminant coefficients,

$b_{i,\ 12}$ is within-sample coefficient of regression of Y_i on Y_{12}, and d is the difference between the observed value and a selected level of the covariate Y_{12}, the selected level in this analysis was its overall mean (Bliss 1970). For the redfish data computed $Z = 10.5866 + 0.8935Y_1 - 9.6952Y_2 - 1.5965Y_4 + 5.7151Y_5 + 12.3990Y_6 - 4.7132Y_7 + 6.5272Y_8 - 7.7912Y_9 - 1.3918Y_{11} - 4.5100Y_{12}$. Simultaneous equations determining L_i (Bliss 1970, P. 335) were based on "total" SS and SP in order to facilitate the ANOVA of Z in terms of X. ANOVA showed that 89% of the total SS in X was attributable to the discriminant function. In an attempt to reduce the number of variates further, this ANOVA was extended to test the significance of each discriminant coefficient the same way as a partial regression coefficient is tested (Bliss 1970). The null hypothesis that each L_i has true or population value of zero was accepted ($p>0.05$) for Y_1 (F =1.36) Y_4 (F =1.89) and Y_{11} (F = 3.55), each F with degrees of freedom (df) = 1, 185, and rejected for every other variate. Following were noted (Bliss 1970, Section 18.3): (1) If coefficients for two or more variates are non-significant, the one with the smallest F is omitted first. Omitting a variate with a non-significant coefficient reduces the error of the remaining recomputed coefficients, especially if the variate is highly correlated with one or more of the other variates. Stability of the discriminant function based on remaining variates is also increased by its omission. For the redfish data paired within-sample coefficients of correlation were all high (in the range of 0.76 to 0.98). (2) Deletion of variates is continued, one at a time and starting with the one which yields smallest F value, until each remaining variable has a significant effect. Following this procedure, the discriminant function was recomputed with Y_1 (body weight) omitted. The recomputed value of the disciminant coefficient for Y_4 was still non-significant. Therefore, the discriminant function was recomputed with Y_4 (inter-orbital width) omitted. As discriminant coefficients of all the remaining variates were then significant ($p<0.05$), computation was stcpped at this point. ANOVA of the discriminant function for the reduced set of variates indicated that percentage of the total SS in X attributable to the discriminant function was still the same, viz. 89%. Thus, it was sufficient to employ seven variates in the discriminant function. It was, however, noted that the combined effect of all partial regression coefficients cannot be partitioned orthogonally when the variates are correlated with one another (Bliss 1970, Section 18.3). Interpretations based on individual discriminant coefficients may, therefore, be of restricted scope. Discriminant analysis uses (rather than removes) intercorrelations among variates (Pimentel 1979).

For discriminant analysis with seven variates Y_2, Y_5, Y_6, Y_7, Y_8, Y_9, and Y_{11} and the covariate Y_{12}, the number of "complete" specimens was 99 in each sample. The discriminant function computed was $Z = 7.0682 - 10.2039Y_2 + 5.6028Y_5 + 12.8670Y_6 - 5.0213Y_7 + 7.3811Y_8 - 8.2966Y_9 - 1.5160Y_{11} - 3.0098Y_{12}$. A larger discriminant coefficient does not necessarily indicate a measure of greater importance than a smaller discriminant coefficient (Bliss 1970). ANOVA of discriminant coefficients by the partial regression approach indicated that characters did not contribute equally and were, in fact, placed in the following order of decreasing importance Y_6 ($F = 115.94$), Y_8, Y_2, Y_9, Y_7, Y_5, and Y_{11} ($F = 4.26$), each F with df = 1,189. In other words, the effective discriminators are: pectoral fin base, length of longest pelvic ray, head length, length of longest pectoral ray, anal fin base, preanal length, and dorsal length of caudal peduncle. Variance of a single Z was estimated as 0.086105. When the discriminant function was computed with Y_{12} (standard length) included as an additional discriminator (rather than as a covariate), variance of a single Z was 0.103184 which is as much as 20% higher than 0.086105. This indicated, in yet another way, that a discriminant function with covariance provided more effective discrimination than one without it. Mean Z values were -4.1697 and -2.77552 for *S. mentella* and *S. fasciauts*, respectively. Difference between these means was highly significant (p`0.001). The zone of uncertainty or wrong identifications of individuals at each end (at $p = 0.05$) was small viz. -3.6530 to -3.2883. Only 8 out of 198 individuals (i.e. 4%) were in this zone of uncertainty.

Acknowledgements

We are grateful to Mr L. S. Parsons and Mr E. M. LeGrow for the privilege of using previously collected data. Dr. G. R. Lilly and Mr. B. Gellately kindly reviewed the manuscript.

References

Barsukov, V. V. 1972. Systematics of the Atlantic redfishes. Trundy, PINRO 28: 128- 142 (Transl. from Russian for Fish. Res. Board Can. Ser No. 2531, 1973).

Barsukov, V. V., and G. P. Zakharov. 1972. Morphologial and biological characteristics of the American redfish. Trudy, PINRO 28: 143-173. (Transl. from Russian for Fish.'Res. Board Can. Ser. No. 2488, 1973).

Blackish, R. E., and R. A. Reyment. 1971. Multivariate morphometrics. Academic Press London and New. York. 412 p.

Bliss, C. I. 1967. Statistics in biology. Vol. 1. McGraw-Hill Company, New York, 558 p.

1970. Statistics in biology. Vol. 2. McGraw-Hill Book Company, New York, 639 p.

Casselman, J. M., J. J. Collins, E. J. Crossman, P. E. Ihssen, and G. R. Spangler. 1981. Lake whitefish (Coregonus clupeaformis) stocks of the Ontario waters of Lake Huron. Can. J. Fish. Aquat. Sci. 38: 1771-1789.

Dillon, W. R. 1979. The performance of the linear discriminat function in nonoptimal situations and the estimation of classification error rates; a review of recent findings. J. Marketing Res. XVI: 370-381.

Goldstein, M., and W. R. Dillon. 1978. Discrete discriminant analysis. John Wiley and Sons, New York, 186 p.

Kendall, M. G., and A. Stuart. 1976. The advanced theory of statistics. Vol. 3. Hafner Publishing Company, New York, 585 p.

Kshirsagar, A. M. 1972. Multivariate analysis. Marcel Dekker, Inc. New York, 534 p.

Lachenbruch, P. A., C. Sneeringer, and L. T. Revo. 1973. Robustness of the linear and quadratic discriminant functions to certain types of non-normality. Communications in statistics 1: 39-56.

Marr, J. C. 1955. The use of morphometric data in systematic, racial and relative growth studies in fishes. Copeia 1955(1): 23-31.

Ni, I-H. 1981a. Separation of sharp-beaked redfishes, *Sebastes fasciatus* and *S. mentella* from northeastern Grand Bank by morphology of extrnsic musculature. J. Northw. Atl. Fish. Sci. 2: 7-12.

1981b. Numerical classification of sharp-beaked redfishes, *Sebastes mentella* and *S. fasciatus* from northeastern Grand Bank. Can. J. Fish. Aquat. Sci. 38: 873-879.

Pimentel, R. A. 1979. Morphometrics. Kendall/Hunt Publishing Company, Dubuque, Iowa, 276 p.

Royce, W. F. 1964. A morphometric study of yellowfin tuna *Thunnus albacares* (Bonnaterre). Fishery Bull. 63(2): 395-443.

Snedecor, G. W., and W. G. Cochran. 1967. Statistical methods. The Iowa State Univ. Press, Annes, Iowa, 593 p.

Table 1. Means and ranges of body weight and eleven morphometric characters Y_i (i = 1,...12) for *S. mentella* (n=97) and *S. fasciatus* (n=99). All measurements were transformed to common logarithms.

Character	Character	*S. mentella*		*S. fasciatus*	
No.	description	Mean	Range	Mean	Range
1	Body weight	2.7000	2.1703-3.2423	2.2324	2.0374-2.6830
2	Head length	2.0011	1.8325-2.1772	1.8211	1.7634-1.9685
3	Snout length	1.3564	1.1584-1.5428	1.1586	1.0682-1.3365
4	Inter orbital width	1.2421	1.0755-1.4099	1.0486	0.9638-1.2529
5	Preanal length	2.2466	2.0846-2.4190	2.0991	2.0158-2.2480
6	Pectoral fin base	1.3460	1.1790-1.5198	1.2184	1.1335-1.3856
7	Anal fin base	1.6124	1.4330-1.7528	1.4438	1.3522-1.6107
8	Length of longest pelvic ray	1.6723	1.4914-1.8156	1.5486	1.4265-1.6776
9	Length of longest pectoral ray	1.8380	1.6484-1.9845	1.6712	1.6085-1.8312
10	Width of caudal peduncle	1.3522	1.1875-1.5340	1.2103	1.1335-1.3579
11	Dorsal length of caudal peduncle	1.5954	1.3324-1.7679	1.4351	1.3263-1.5888
12	Standard length	2.4238	2.2742-2.5763	2.2725	2.2041-2.4133