POPULATION DIVERSITY IN THE RAINBOW SMELT, OSMERUS EPERLANUS MORDAX (MITCHILL, 1814) (SALMONOIDEA: OSMERIDAE) AS REVEALED BY CANONICAL AND DISCRIMINANT FUNCTION ANALYSES ON MORPHOMETRIC, MERISTIC AND ESTERASE DATA

DOUGLAS GORDON COPEMAN

POPULATION DIVERSITY IN THE RAINBOW SMELT, *OSMERUS EPERLANUS MORDAX* (MITCHILL, 1814) (SALMONOIDEA: OSMERIDAE) AS REVEALED BY CANONICAL AND DISCRIMINANT FUNCTION ANALYSES ON MORPHOMETRIC, MERISTIC AND ESTERASE DATA

by

© Douglas Gordon Copeman, B.Sc.

A Thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Department of Biology
Memorial University of Newfoundland

St. John's Newfoundland

April 1973

Codispiece

ACKNOWLEDGEMENTS

I would like to express thanks and appreciation to a number of people who, in many ways, have made contributions, resulting in the final form of this thesis.

Dr. R. G. Ferguson suggested that smelt might exist in separate populations. In addition to freely assisting in the collection of materials, Dr. M. L. Petras indicated that populations might be identified by biochemical methods.

A number of people were of essential assistance in the collection of samples. These were C. Chaisson, M. Copeman, Dr. G. Cowan, Dr. C. Delisle, M. Farwell, T. Lafette, Dr. M. Petras and R. Stevenson. In addition, various officials from Environment Canada's Fisheries Service and fisheries personnel from Maine, Ontario and the federal branch in Wisconsin provided material. Their energies involved both specimen capture and shipment to the M.S.R.L., packed in dry ice. In all cases, assistance or procurement was freely and amicably provided.

Technical assistance in a number of areas was also provided. Dr. G. Cowan was helpful with certain aspects associated with micro starch gel electrophoresis technique. M. Blackmore, M. Copeman and T. Sullivan prepared samples for electrophoresis runs. At various times during the research, photographic assistance was provided by P. York, G. McCurdy, G. Hillier, A. McAllister and R. Ficken. Computing assistance was freely given, especially by S. Andrews of N.L.C.S. and R. Lauzon of University of Windsor. Dr. G. Ross generously supplied Fortran versions of computer programs while L. Lefkovitch of the Statistical

Research Service, Canadian Department of Agriculture processed some data, answered numerous questions and offered valuable criticism. Dr. M. K. Lewis provided many suggestions regarding the application of tests of significance for the esterase data. The final draft was rapidly typed by R. Daggett and her perserverance was appreciated. The contributions from conversations with many people at the M.S.R.L. is noted, as were those with P. Field, W. MacTavish and W. Pohlman. In addition, the efforts of those on my research supervisory committee are acknowledged.

The Board of Regents kindly granted me a two-year leave of absence without pay so that I could complete this thesis. During this time, personal financial support was awarded to me in the form of a Memorial University graduate student fellowship and a fellowship from funds awarded to the university for that purpose by the Fisheries Research Board of Canada. Materials and assistance were made possible by a National Research Council of Canada grant number 5072 (18-655) to Dr. G. Cowan. Further assistance was also permitted by M. Copeman. Various pieces of equipment and chemicals were provided by M.S.R.L.

The form (substantial) of this thesis and possibly its existence have been strongly influenced by two men. Dr. J. R. Strickler introduced me to multivariate analysis and the application of computers to biological research. Without this input, it is dubious whether the problem could have been approached satisfactorily. Further, his enthusiasm, confidence and respect did not appear to falter and was sustaining.

On numerable occasions, Mr. J. Gower was asked a host of questions regarding details associated with multivariate analyses. These, sometimes trivial questions, were responded to promptly, clearly and with an appreciated lack of acridity. Meetings with Mr. Gower proved memorably informative and his assistance surpassed any expected level.

Finally, I would like to thank Margaret for her participation. Besides assisting in many concrete ways, she was somehow able to maintain calm, in the teapot, throughout the tempest.

ABSTRACT

The rainbow smelt, *Osmerus eperlanus mordax* was first described by Mitchill in 1814. It can live entirely in a freshwater or marine environment. Freshwater smelt may spawn in streams, on beaches or in relatively deep water (Rupp, 1959; Delisle, 1969). Marine smelt are usually anadromous although they may spawn on the bottoms of quiet brackish bays. There has been considerable debate as to whether or not two forms exist, one large and the other small.

There is a great deal of evidence indicating that the smelt exists in rather isolated populations. Therefore, they could be expected to be subjected to different sets of selective factors. These, presumably, would exhibit their different effects by modifications in the characteristics of the populations. Conversely, if populations of smelt from different areas are found to be different, then the populations' proposed attributes (isolation and different selective pressures) are substantiated.

Thirteen smelt populations were sampled during their spawning runs. Morphometric, meristic and biochemical data were recorded for each population. These data were analysed separately and together to obtain a measure of

the degree of difference between the populations. Appropriate statistical methods were employed throughout the study.

It was found that there were differences between the samples on the basis of any type of data and that the differences were real. Individuals could be identified as to population of origin with an average accuracy of 93.4 per cent. The differences between small and normal forms were very large and the marine and freshwater smelt were also distinct. There were also indications that smelt populations vary, not only from place to place but from year to year. Classifications derived from the data indicated that environmental factors may be important in accounting for the observed differences between populations. A considerable amount of redundancy was observed in the data.

The work forms a comprehensive model for the investigation and identification of populations of a species. Methods are proposed for handling data for which the standard analyses of population genetics are not applicable.

TABLE OF CONTENTS

I	INTRODUCTION	1
II	METHODS AND MATERIALS	9
	(A) Description of Sampling Areas and Samples	9
	(B) Electrophoresis - Standard Procedure	13
	(1) Sample Preparation	13
	(2) Electrophoresis	13
	(C) Meristic and Morphometric Methods	14
	(D) Data Processing	16
	(1) Data Preparation	16
	(2) Programmes	16
	(3) Major Analyses	17
III	RESULTS	22
	(A) Data Analysis	22
	(1) Raw Data	22
	(2) Transformed Data	27
	(3) Age Distribution	27
	(B) Discriminant Function Analysis	31
	(1) Latent Roots and Vectors	32
	(C) Summary of the Analyses in the Study, Latent Roots and Variables	36
	(D) Classification Matrices	46
	(1) ALLALL Analysis	46
	(2) Classification Summary of All Analyses	56

(E) Plotting Samples Along Canonical Axes 61
(1) Two Dimensional Representation 62
(2) Three Dimensional Model 65
(F) Confidence Region for Sample Means 73
(G) Calculation of the Intersample Distances 74
(H) Cluster Analysis (ALLALL) 81
(1) Cluster Analyses 81
(2) Three Canonical Axes 83
(3) All Canonical Axes 87
(I) Esterases 92
(1) The General Pattern and Sample Comparisons 92
(2) Frequency of Esterases in Each Sample 95
(3) Joint Occurrence Matrices 99
(4) Chi-Squared Tests 100
(J) Classification of Samples on Esterase Data 111
(1) One Band Considered 111
(2) Two Bands Considered 116
(3) Three Bands Considered 117
(4) All Bands Considered 120
(5) Classification of Each Phenotype 120
(K) Intersample Distances from Esterase Data 125
(1) Between and Within Groups Distances 128
(2) Canonical Analysis 128
(3) Cluster Analysis 131
(L) Combination of Intersample Distance Matrices from ALLALL and ESTALL 136
(1) Cluster Analysis 137

(M) Transformed Data 144

(1) Classification from SZOTME 145

(2) Canonical Analysis 146

(3) Cluster Analysis 149

(N) Combination of SZOTME and Esterase Data 151

(1) Cluster Analysis 151

(O) Effect of Using a Reduced Data Set 154

IV DISCUSSION 161

SUMMARY 219

BIBLIOGRAPHY 227

LIST OF TABLES

		PAGE
Table 1.	Character, code, overall mean and standard deviation of the group means, for the raw data from thirteen smelt populations	23
Table 2.	Character code, overall mean and standard deviation of the group means, for the transformed data from thirteen smelt populations	28
Table 3.	Percentage age distribution for the various samples	30
Table 4.	Latent roots and accumulative percentage of dispersion for each analysis performed with data from smelt populations	37
Table 5.	Fist five variables incorporated into the discriminant functions for each analysis performed with data from smelt populations and the level of significance for the hypothesis that the groups' means are the same with the inclusion of the fifth variable	42
Table 6.	Classification matrix of smelt samples from the analysis ALLALL, two variables considered	47
Table 7.	Classification matrix of smelt samples from ALLALL, all (32) variables considered	51
Table 8.	Summary of the classification value of the variables in the various analyses conducted with data from smelt samples	57
Table 9.	Standard deviation of the individuals of each smelt sample about their means, for canonical axes I and II (ALLALL)	68
Table 10.	Distances between smelt samples' means computed from the first three canonical variate axes of the analysis ALLALL	77
Table 11.	Generalised distances (D) between smelt samples' means computed from twelve canonical variate axes of the analysis ALLALL	79

PAGE

Table 12. Frequency of each esterase band in each group of smelt and the chi-square value for each band 96

Table 13. Intragroup chi-square values from esterase data and levels of significance, for each sample of smelt 102

Table 14. Chi-square matrix to test the homogeneity between smelt samples, based on the esterase data 107

Table 15. Levels of significance from between smelt samples chi-square matrix (Table 14) 109

Table 16. Average within and between groups chi-square for selected groups of smelt (compiled from Table 14) 112

Table 17. Classification matrices of smelt samples based on esterase data
(a) one band considered
(b) two bands considered 114

Table 18. Classification matrices of smelt samples based on esterase data
(a) three bands considered
(b) all bands considered 118

Table 19. The number of individual smelt from the Chuff Brook sample (C.B.) classified into each group (area) on the basis of the phenotypes 123

Table 20. Distances (D^2) between smelt samples' means derived from esterase data 126

Table 21. Average within and between groups distances based on esterase data for selected groups of smelt samples 129

Table 22. Distance matrix (D^2) derived from combined meristic morphometric and esterase data from thirteen smelt samples 138

Table 23. Effective distance between smelt samples, related to the percentage of correct classification and the total variation between samples for three analyses employing reduced data sets 159

PAGE

Table 24. Summary of the results of transplant studies on *O. eperlanus mordax* in Maine (Rupp and Redmond, 1966) 186

Table 25. Average between samples chi-square values computed from esterase data for different groups of salt water samples of smelt
(a) all 1969 slat water samples, Grand Bank sample, 1970 salt water samples excluding Grand Bank
(b) Notre Dame Bay samples, other salt water samples 198

LIST OF FIGURES

		PAGE
Figure 1.	Diagram of frequency function of two normal populations	33
Figure 2.	Graph showing percentage of individual smelt correctly classified versus the number of variables included in the discriminant functions for the analysis ALLALL	53
Figure 3.	Plot of smelt samples' means along canonical variate axes I and II arms of crosses extend for one standard error of the mean	63
Figure 4.	Three dimensional model of the dispersion of the smelt samples in the analysis ALLALL a. upper left hand view b. view along canonical axes III c. bottom right hand view	71
Figure 5.	a. Plot of smelt samples' means along canonical variate axes I and II and M.S.T. computed from intersample distances derived from the first three canonical axes in the analysis ALLALL b. Dendrogram of smelt samples from U.P.G.M. cluster analysis on intersample distance matrix derived from the first three canonical axes in the analysis ALLALL	84
Figure 6.	a. Plot of smelt samples' means along canonical variate axes I and II and M.S.T. computed from the intersample distances derived from all canonical axes in the analysis ALLALL b. Dendrogram of smelt samples from U.P.G.M. cluster analysis on intersample distance matrix derived from all canonical axes in the analysis ALLALL	88
Figure 7.	Photographs of electropherograms of some smelt biochemical systems	93

PAGE

Figure 8. a. plot of smelt samples' means along canonical variate axes I and II and M.S.T. computed from the intersample distances derived from all canonical axes in the analysis ESTALL
b. Dendrogram of smelt samples from U.P.G.M. cluster analysis on intersample distance matrix derived from all canonical axes in the analysis ESTALL 132

Figure 9. a. Plot of smelt samples' means along canonical variate axes I and II from ALLALL and M.S.T. computed from intersample distance matrix derived from the combination of all canonical axes of ALLALL and ESTALL
b. Dendrogram of smelt samples from U.P.G.M. cluster analysis on intersample distance matrix derived from the combination of all canonical axes of ALLALL and ESTALL 140

Figure 10. a. Plot of smelt samples' means along canonical variate axes I and II and M.S.T. computed from intersample distance matrix derived from all canonical axes in the analysis SZOTME
b. Dendrogram of smelt samples from U.P.G.M. cluster analysis on intersample distance matrix derived from all canonical axes in the analysis SZOTME 147

Figure 11. a. Plot of smelt samples' means along canonical variate axes I and II from SZOTME and M.S.T. computed from intersample distance matrix derived from the combination of all canonical axes of SZOTME and ESTALL
b. Dendrogram of smelt samples from U.P.G.M. cluster analysis on intersample distance matrix derived from the combination of all canonical axes of SZOTME and ESTALL 152

Figure 12. a. Plot of smelt samples' means along canonical variate axes I and II from 3RAW01
b. Plot of smelt samples' means along canonical variate axes I and II from 3RAW02
c. Plot of smelt samples' means along canonical variate axes I and II from 3TGN01 155

LIST OF EQUATIONS

EQUATION		PAGE
1	General form of canonical equation	61
2	Radius of the confidence area of the sample mean	73
3	Distance between the i^{th} and j^{th} units	75
4	Similarity coefficient between the i^{th} and j^{th} samples	81
5	Chi-square value for each esterase band	98
6	Within sample chi-square equation	101
7	Overall chi-square equation	104
8	Relationship between α point of chi-square distribution to the α point of the cumulative normal distribution	105
9	Between pairs chi-square equation	106
10	Value of the i^{th} discriminant function	121
11	Combined (total) distance between the i^{th} and j^{th} samples from two distinct analyses	137

INTRODUCTION

This work is concerned with various populations of the rainbow smelt, *Osmerus eperlanus mordax* (Mitchill, 1814), Order Isospondyli, Family Osmeridae (Bigelow and Schroeder, 1963). McAllister (1963) reviewed the family Osmeridae and suggested that *Osmerus mordax* should be considered a subspecies of the European smelt *Osmerus eperlanus* (L). As a result of this revision, it was argued that all the North American, Pacific and Eastern Arctic forms of this species should be referred to as *Osmerus eperlanus mordax* (Mitchill). I have chosen to follow these suggestions when referring to the species on which I have been working.

In the results of a twenty-two year study of the biology of the smelt of the Miramichi River, New Brunswick, McKenzie (1964) concluded the following, based on tagging and recapture studies. Larger smelt moved less in the bay and streams than did the smaller fish. There was no exchange of animals between the early and late spawning grounds and little exchange between the main branches of the river. The smelt from the various spawning grounds exhibited a high degree of homing. This indicated that there was a subdivision of the smelt population of the Miramichi River system, to restricted areas within the system at time of reproduction,

i.e. there was considerable reproductive isolation between subpopulations.

There is also parasitological evidence that smelt do not move far from their spawning areas. Templeman *et al.* (1957) have shown a high incidence of nematode (Genus *Porrocaecum*) infection in smelts when they and the definitive host of the nematode, seals, inhabit the same territory. However, there was little to no infection in smelts living only ten miles from a seal colony. This indicates that smelts do not range far or they would have shown a heavier infection from the nearby seals. The tagging data from McKenzie (1964) corroborates the above.

Physical factors also play an important role in the isolation of smelt populations. Marine smelts are usually anadromous and confined to a narrow zone of less than a mile off shore, in no more than two to three fathoms (4-6 m) of water (Bigelow and Schroder, 1953). Further, only certain types of rivers are suitable for smelt spawning. Gravel or coarse sandy stream beds seem to be a necessity and even small impediments (small falls, rapids or brushpiles) can prevent the smelt from reaching otherwise suitable spawning areas (Templeman, 1966). Therefore, it appears that there are a number of factors which indicate that marine smelt populations tend to be rather isolated from each other. It should also be noted that although smelts are found at numerous sites around the Newfoundland coast, there are often

long stretches between the associated rivers which support spawning populations.

The evidence for the isolation of the lake (freshwater) populations is more obvious. If a lake is landlocked, the resident population is isolated. If a lake is part of a system (i.e. the Great Lakes), the individual lakes are often separated by rather long fast flowing rivers which would serve as deterrents for the up stream migration of fish. Down-system migration is possible, as exemplified by the distribution of smelts in the Upper Great Lakes (Creaser, 1925). Although smelts have moved up-system into Lake Superior during their original dispersal in the Great Lakes, this is probably not a general phenomenon. Once the smelt become established in the Great Lakes, they developed preferences for certain spawning areas. Thus, for large lakes, the isolating mechanisms which apply to marine smelts have a parallel for the populations of these lakes.

Therefore, the evidence indicates that smelts, whether marine or freshwater forms, exist in populations which are more or less isolated, especially at time of reproduction.

Sewall Wright (1949) came to a number of conclusions with regard to the role of populations in general in the evolutionary process. A certain degree of subdivision of a species into partially isolated groups provides the largest store of variability both locally and within the

species as a whole. A continual shifting of the statistical characteristics of the populations is to be expected within any species which occupies a reasonably large range, at low or moderate density. This shifting of characters eventually becomes so great in certain populations that new species must eventually be recognised. In the splitting of one species to form two or more, the critical event is necessarily a complete or nearly complete interruption of gene flow between the populations. Once reproductive isolation has occurred, the accumulation of random and selective differences in genetic composition is increased; or, if migration into unoccupied territory occurs, extensive changes should follow as the result of new selective pressures. In either case, this should result in the exploitation of the various special niches made available by either the modified genetic composition or the mere absence of competition. Therefore, smelts represent favourable animals in which to study divergence caused by genetic isolation, in that the characteristics of this species satisfy the preconditions set by Wright, which favour diversity amongst the populations of a species. Isolated populations are subjected to different selective pressures. The marine forms appear to be reproductively isolated from each other and the Great Lakes populations were introduced into the upper Great Lakes as the result of a single transplant into Crystal Lake, Michigan, in 1912.

As smelt occur in more or less isolated populations a major objective would be to determine the degree of divergence that exists between samples (of smelt) from various populations. Divergence can be measured in many ways, however, a primary consideration in any measurement is that the greater the divergence between any two forms, the greater the ease in distinguishing between them. Therefore, the problem becomes one of establishing the degree to which various samples can be distinguished from each other. The use of discriminant function analysis (Fisher, 1936), allows one to identify an individual as belonging to one of a number of groups, with the least chance of making a mistake, given that the individual does in fact come from one of the groups. The number of available characters which can be used as potential discriminators between samples is finite, although very large. In this study, selected meristic and morphometric characters as well as "biochemical" (esterase and muscle myogen) characters, revealed by means of electrophoretic procedures, were used as discriminators. Each of these types of data can be tested individually by determining the extent to which it permits individuals to be correctly assigned to their appropriate groups and thus their powers as discriminators determined. Furthermore, it is possible to determine which variables in each type (morphometric, meristic, and biochemical), or which combinations of these types allow for the maximum

discrimination between the samples.

If individuals and thus samples can be distinguished, it should then be possible to graphically present the relationships between the samples. One method of achieving this end is by means of canonical variate analysis. This method has been used to represent the dispersion between different samples of species (Delany and Healy, 1964, 1966). The procedure results in a series of axes which represent the different modes of variation amongst the samples. The first axis represents the largest mode of variation, the second axis represents the second largest mode and so on. It is therefore possible to position the individuals of the samples with respect to the canonical axes. Thus each individual can be assigned a position in a multidimensional space defined by the canonical axes. Once the samples have been dispersed, it is then possible to analyse the relationships between them on the basis of the distances between the samples. Cluster analysis and Minimum Spanning Tree analysis may be used in conjunction with the plot of the samples along the major canonical axes to show the relationships which exist between the samples on the basis of the different data sets considered.

The foregoing analytical procedures do not operate from any *a priori* position regarding the relationships between the groups, except that the groups exist. There is no prior weighting of variables and the eventual variable

weightings in the analyses are assigned as the result of the values and properties of the variables themselves. Therefore, it is of interest to see if the final variable weightings would produce results which reflect known differences. That is, would the values of the variables themselves permit the distinction between large and small forms, salt water and freshwater groups, or various year classes?

Finally, it has been a standard technique in ichthyology to use ratios of the various measurements in the comparison of samples. This is done on the rationale that by converting raw measurements to ratios the size factor is eliminated, resulting in the establishment of a more reliable measure by which the samples can be distinguished. The use of ratios in multivariate studies has been discussed by Sokal and Sneath (1963), Sokal and Rohlf (1969) and Blackith and Reyment (1971). However, it is possible to ask the same questions using data that have been converted to ratios as with unconverted (raw) data and to evaluate the relative merits of raw data and ratios with regard to their discriminatory value. Comparisons can also be made between the two forms of the variables when the dispersion patterns resulting from appropriately matched analyses are examined.

As the rainbow smelt exists as a number of more or less isolated populations, it was hoped that this work would elucidate the level of variation (degree of wobble) that exists within this species. Large variations would suggest

the independence of the populations whereas little variation would indicate a more homogeneous group. With these general questions in mind, the study was undertaken.

MATERIALS AND METHODS

The following is a description of the sampling areas and the methods by which the samples were obtained. Also, included is the code which will be used to refer to each of the samples in the remainder of the text.

(A) Description of Sampling Areas, Samples and Sample Code

(1) Norris Arm 1969 (N.A. 69)

Bottom Brook runs into Northeast Arm of the Bay of Exploits. The sampling area was just west of a bridge on the North Norris Arm Road, approximately 0.9 miles north of the turn off from the Trans Canada Highway. According to local observers, in 1969 the smelt had first appeared in the brook on April 12th. On April 16, they were being fished during the day by truant children and a number of adults. Approximately 120 males and 10 females were taken from the pool below the culvert-type bridge between 1230 and 1300 hrs. A second sample was taken at Bottom Brook between 2330 and 2400 hrs., on April 17. Approximately 80 females were selected. The water temperature was not taken at the actual time of sampling but on June 7 it was 15.4°C.

(2) Chuff Brook (C.B.)

Chuff Brook is between four and six feet wide. It is

located about ½ mile south of Sandringham on the road to Eastport, Bonavista Bay. The smelt had started to run on the evening of June 5, 1969. A few males were seen trapped in pools at 1500 hrs. The fish started entering the stream about midnight and the sample was taken between 100 and 130 hrs., June 7, 1969. The water temperature, at time of capture, was 12.5°C.

(3) Little River (L.R.)

Little River or Benoit Brook, as it is also known, is the first river north of Fox Island River, Port au Port Bay. The sample was taken approximately ¼ mile from the mouth of the river at 1330 hrs., June 15, 1969. The water temperature was 12°C. There were a large number of fish in the river. They were very active and the majority were able to avoid the seine net.

(4) Smelt Brook (S.B.)

Smelt Brook is located 4.4 miles north of Port au Port on the way to Port au Mal and Fox Island River. The sample was taken between 50 and 100 yards upstream from the mouth of the brook, by seine. The run had begun on June 8 and the sample was taken at 1530 hrs. a week later, June 15, 1969. The water temperature at time of capture was 9.0°C.

(5) Indian River (I.R.)

Indian River is located approximately one mile, by foot, from Boyd's Cove, Notre Dame Bay. At 2030 hrs. when

the sample was taken, the smelts in the stream were predominantly males. There were eggs attached to the rocks on the bottom of the stream indicating that the spawning run was already in progress. The sample was taken by seine on May 13, 1970, from water at 9.5°C.

(6) Norris Arm 1970 (N.A. 70)

The 1970 Bottom Brook sample was taken from the same location as the 1969 sample. The collection was made between 12 midnight and 100 hrs., May 14 by seine and was composed primarily of males. The water temperature was 9.0°C.

(7) Grand Bank (G.B.)

The Garnish River is located just north of the town of Garnish on the Burin Peninsula. This sample was collected on October 16, 1970, by officers of the Federal Department of Fisheries, Grand Bank. The sample was then transported to Fortune, where it was deep frozen and stored. The water temperature was 5.6°C.

(8) Green Lake (G.L.)

Green Lake, Maine, is a landlocked lake which contains the parental stock of the smelts of the upper Great Lakes (Creaser, 1925). The sample was taken by officials of the Department of Inland Fisheries and Game of The State of Maine. Approximately 100 smelts were collected, between 2200

and 2400 hrs. on April 27, 1970, from the first part of the season's spawning run. The sample was composed predominantly of males. They were frozen immediately after capture and shipped to St. John's on dry ice.

(9) Lake Superior (L.S.)

The Lake Superior sample was taken from the west shore of Chequanagon Bay near Ashland, Wisconsin, by officers of United States Bureau of Commercial Fisheries. The fish were frozen immediately after capture and air-expressed to St. John's on dry ice. The sample was taken on April 23, 1970, from water of 5.0°C.

(10) Lake Huron (L.H.)

The Lake Huron sample was collected by officers of the Ontario Department of Lands and Forests. These fish were from the spawning run of April 29, 1970. The collecting site was McKim Creek, South Bay, Manitoulin Island, Ontario.

(11) Lake Erie (L.E.)

The sample from Lake Erie was beach seined near the town of Morpeth, Ontario, on April 23, 1970. The smelt were spawning between 2100 and 2130 hrs. in water at 12.5°C.

(12, 13) Lake Heney Giant (L.H.G.) and Lake Heney Stunted (L.H.S.)

The samples of "giant" and "stunted" smelt were collected by Dr. C. Delisle from Lake Heney, Gatineau County, Quebec. They were seined in May, 1967 and preserved in 10% formalin. No fresh specimens were available.

(B) Electrophoresis - Standard Procedure

(1) Sample Preparation

Upon capture, all fish were quick frozen in water and stored at -20°C until used. A sample of approximately 0.5 grams of skeletal muscle and a second sample consisting of most of the liver in small fish and up to one gram of liver in larger specimens were taken from each fish. Each tissue sample was mixed 1:1 (wt./vol.) with distilled water and ground in a Virtis 45 tissue homogenizer with micro cup. The ground samples were then frozen and thawed to increase cell disruption and centrifuged at 54,000 g for one hour at 2°C. The supernatants were decanted off and recentrifuged at 2°C and 7,500 g for at least one hour for muscle extracts and at least twelve hours for liver extracts.

(2) Electrophoresis

Muscle myogens and liver esterases were investigated using the following procedure. Electrophoresis was performed in a micro starch gel apparatus (Tsuyski *et al.*, 1966). A borate buffer system at pH 8.5 was used with 12% (wt./vol.) starch gels (Connaught Laboratories). Ionic strength of the bridge and gel buffers were 0.30 M and 0.023 M respectively. Electrophoresis proceeded for 90 minutes at 210 volts and 4°C. After electrophoresis was completed, the gel strips to be examined for myogens were stained in Amido Black 10B (British Drug House) at 0.10% (wt./vol.) in gel wash (1 water:

1 methanol: 0.2 acetic acid), for 5 minutes. The stain was then poured off and the gel strips were placed in a gel washer to remove the excess background stain. When this had been satisfactorily completed, the gel strips were wrapped in plastic (Handi Wrap), labelled and stored at 5°C until they were to be read.

Gels to be examined for the esterase activity of the liver extracts were developed in a solution containing 0.10 gm Fast Blue RR, 188 ml water, 8 ml 2M Tris-HCL buffer at pH 7.0 and 4 ml of 1% vol./vol. alpha napthyl-butyrate (Sigma). Development continued in the dark for 10 to 20 minutes at room temperature. The gel strips were then washed with water and placed in gel wash to harden. The gels were read using a Kodak colour transparency illuminator.

(C) Meristic and Morphometric Methods

After the samples for electrophesis were taken, the fish were fixed in 10% neutralized formalin. Samples of 20 fish were taken from each population. Fifty-eight data points were recorded from each of these fish. These included two identification points, twenty morphometric characters, thirteen meristic counts, four binary decisions and twenty present/absent decisions from electrophoresis bands. For the remainder of each sample, identification, weight, sex, sexual maturity and the electrophoresis bands were recorded.

The morphometric and meristic characters used in

this study are seen in Table 1,(Results). These measurements were made according to the methods described by Hubbs and Lagler (1947), except where marked with asterisks (Table 1). In these cases, the measurements were as described by Berg (1948). Age was determined using the "shiny line" method (McKenzie, 1958). Measurements were taken to the nearest 1/10 mm, using Helios dial calipers. The tip of the lower jaw was taken as the most anterior point on the animal and it was used as the anterior reference point as it seemed more stable than the upper jaw. Vertebral counts, dorsal and anal fin ray counts were made from X-ray plates. Kodak Industrial Type M, X-ray film was used throughout the study. The X-rays were produced by Picker Hotshot X-ray Machine with a 0.5 mm aperature and a beryllium window. A Kodak colour transparency illuminator was used to read the resulting X-ray plates.

The four binary decisions were: male/female, ripe/spent, gill rakers--normal/abnormal, and pored lateral line scales--continuous/discontinuous. The determination of the first two of these measurements was done when the tissue samples for the electrophoresis experiments were taken. At this time, the gonads could be easily studied. A gill raker was scored as abnormal if it curved to an angle of 90° or greater with its long axis or if there was any branching seen on the structure. The pored scales of the lateral line were scored as discontinuous if there were any

non-pored scales flanked by two pored scales.

(D) Data Processing

(1) Data Preparation

When the measurements were taken, they were recorded on eighty column general coding form data sheets. Two formats were used. All the binary data were recorded in one format (I2) and the remainder of the data were recorded in the other (F6.1). Computer cards were then keypunched. A printout of the punched cards was obtained and checked against the original data sheets to detect any errors that might have occurred. Errors were corrected and the cards for the various samples were grouped together in order. Four cards were required for the data on each fish. Each card also contained information about the sample area and which fish the card represented. The sequence number of the card (1-4) was also included. With the data in this form, the subsequent analyses of the data were facilitated.

(2) Programmes

The data were analysed with the aid of a number of computer programmes. Frequency and joint frequency counts were done using the BMDO8D programme. The BMDO7M programme was used for investigations into the classification value of the different meristic and morphometric characters, the discriminant function analysis and the dispersion of the

populations. The conversion of the raw data to indices and logarithmic functions was accomplished by means of the BMO9S programme. The Fortran IV versions of the Minimum Spanning Tree and Single Linkage Cluster Analysis programmes were kindly supplied by Dr. G. Ross. Algol 60 versions of these programmes had previously appeared (Ross, 1969 a, b, c) but were unsuitable for use on the available IBM 370 computer.

The Chi-Square programmes to test, the overall homogeneity of the binary characters, the independence of the binary characters in each group, the χ^2_{10} value for each band and the homogeneity of binary characters between each pair of groups, were written by the author. Other programmes written by the author include those to: (1) generate all possible phenotypes for binary data, (2) calculate the posterior probabilities for each phenotype given the discriminant function coefficients and constant for each group, (3) sort and count the phenotypes in each population, (4) assign each phenotype to its "best" group, (5) compute and print an inter-population distance matrix given a set of coordinates for each population, (6) convert the distance matrix to a similarity matrix and find the maximum, total and average intersample distances.

(3) Major Analyses

The following analyses were made on the data using the programme BMDO7M. The code names of the various analyses

are given with a description of the variables involved and the purposes for which the analyses were done.

ALLALL contained all 33 of the meristic and morphometric variables for all of the 13 populations. The analysis was allowed to proceed until all the variables had been included in the discriminant functions. This analysis was done to determine the total variation between all populations using all variables in their raw (untransformed) form.

ALLNO1 contained all variables except the one which had the highest F value in the ALLALL analysis. This variable was omitted to determine the effect of the most significant variable in ALLALL. ALLNO2 contained all variables present in ALLNO1 except the one with the highest F values. These three analyses were done in an attempt to determine the extent to which each excluded variable contributed to the correct classification of the individual fish.

From the "ALL" series of analyses it becomes apparent that a major division of the populations was between the freshwater populations, excluding the two "stunted" populations, and the salt water ones. FWTSWT contained all variables and compared the freshwater populations to the salt water ones.

FWTALL was done on all the freshwater populations excluding the "stunted" populations and contained all

variables.

SWTALL included all the populations from the salt water sources and all the variables were considered.

Morphometric measurements are often combined to form indices or expressed as a function of some other measurement. This is done in an attempt to overcome the effect of size between populations and to minimize the standard deviations. The morphometric data was transformed to form the various indices and functions seen in Table 2 (Results). TGNALL was run containing all the populations and only the transformed variables. TGNALL was done to determine the discriminatory effect of the transformed variables.

TGNMER contained all populations and all transformed and meristic variables. This was done as an analagous analysis to ALLALL and thus allows comparison between the use of raw morphometrics plus meristics and transformed morphometrics plus meristics.

From an investigation of ALLALL and TGNMER, it was apparent that the major distinction was between the two "stunted" populations and the remainder. Therefore, variables were included that enhanced this major distinction. In ALLXHM, the two "stunted" populations were excluded from consideration and all variables were used to discriminate between the eleven remaining populations.

TGNXHM was similar to ALLXHM in that the two "stunted"

populations were not considered. In this analysis, all transformed variables and meristics were available for inclusion in the discriminant function.

MERIST contained only the eleven meristic counts. All thirteen of the populations were included in this analysis. The purpose of this analysis was to determine the value of meristics in the discrimination of the various groups.

MORPHPO contained the twenty morphometric measurements for all populations. This analysis is similar to TGNALL except in this case the data were in their raw form. The purpose of this analysis was to determine the value of the morphometric data in the discrimination of the various populations.

ESTALL was run on the eleven populations and included the binary data from the eight esterase bands. This analysis was done, not only to determine the classification value of the various esterase bands, but also to determine the spacial distribution of the populations on the basis of their esterase components. It was hoped that this would aid in the interpretation of the esterase results from the Chi-Square tests.

SZOTMO contained all 13 samples and all the transformed variables except those that might be expected to contain a size factor (Wt./Age and $\log_{10}$ St.L.).

SZOTME was similar to SZOTMO, except that the

meristic characters were also included. These two analyses were run to evaluate the effect of size on the outcome of the results.

3RAW01 contained the best meristic character and the two best raw morphometric characters (Table 5, Results).

3RAW02 contained the best meristic character and the next two best raw morphometric characters.

3TGN01 contained the best meristic character and the two best transformed characters that were not related to size. These three analyses were done to test the extent to which the results of the previous tests could be approached by the information in just a few variables (redundancy of the data).

RESULTS

(A) Data Analysis

(1) Raw Data

The first step in the analysis was to obtain an idea of the size ranges for the various morphometric and meristic characters. This was done by calculating the overall mean and the standard deviation of the groups' means for each character. These results are seen in Table 1. The third column of this table contains the character code which will be used in other parts of the text. Thus, from Table 1, an overall estimate of the values of the characters and their respective degree of variability was obtained.

Two of the samples were composed of individuals that were unmistakably shorter than even the shortest members of the other normal samples. Subsequently, they were referred to as stunted (Delisle, 1969). Their weights averaged about four grams as compared to an average of more than 35 grams for the remainder of the samples. As a result, the weights of the stunted fish were recorded in decigrams to one decimal place, in order to maintain the same format and number of significant digits as had been used for the other normal populations. This resulted in a more conservative effect from the weights, as the difference between the stunted and

TABLE 1

CHARACTER, CODE, OVERALL MEAN AND STANDARD DEVIATION OF THE GROUP MEANS, FOR THE RAW DATA FROM THIRTEEN SMELT POPULATIONS

No.	Character	Code	Overall Mean	Standard Deviation of Group Means
1	Weight	Wt.	38.81 g (33.25)	12.72 (18.30)
2	Age	Age	3.06 year	0.67
3	Standard length	St.L.	141.11 mm	32.88
4	Anal fin to maxillary length (*)	A.F.-Mx.	106.83 mm	24.82
5	Dorsal fin to maxillary length	D.F.-Mx.	73.60 mm	16.99
6	Pelvic fin to maxillary length (*)	Pl.F.-Mx.	76.29 mm	14.74
7	Pectoral fin to maxillary length (*)	Pc.F.-Mx.	35.27 mm	8.06
8	Posterio-Ventral Region length (*)	P.V.R.	42.10 mm	10.01
9	Base of Dorsal fin length	B.D.	12.44 mm	3.13
10	Height of Dorsal fin	H.D.	21.61 mm	4.57
11	Length of Pectoral fin	L.Pc.F.	20.54 mm	4.32
12	Number of Pectoral fin Rays	#Pc.R.	11.65	0.47
13	Length of Pelvic fin	L.Pv.F.	17.29 mm	3.84
14	Number of Pelvic fin Rays	#Pv.R.	7.99	0.04

No.	Character	Code	Overall Mean	Standard Deviation of Group Means
15	Base of Anal fin length	B.A.	20.10 mm	4.81
16	Height of Anal fin	H.A.	12.67 mm	2.67
17	Length from adipose fin to base of the tail (*)	Ad.-T.	21.61 mm	5.21
18	Depth of Caudal Peduncle	D.Cd.P.	8.05 mm	1.83
19	Length of Head	L.Hd.	35.39 mm	8.24
20	Length of Orbit	L.Or.	8.10 mm	1.50
21	Length of Snout	L.Sn.	10.14 mm	2.71
22	Inter-Orbital Width	I.O.W.	8.17 mm	2.32
23	Length of Upper Jaw	L.Mx.	16.23 mm	4.22
24	Number of branchiostigeal Rays	#Brgl.R.	4.99	0.08
25	Number of Pored Lateral Line scales	#Pd.L.L.	16.94	1.38
26	Number of Gill Rakers - upper arm	Sp.G.R.	9.90	.90
27	Number of Gill Rakers - lower arm	I.G.R.	21.55	1.82

No.	Character	Code	Overall Mean	Standard Deviation of Group Means
28	Total number of gill rakers	Σ G.R.	31.46	2.70
29	Number of hemal vertebrae	He.V.	39.92	0.69
30	Number of caudal vertebrae	Ca.V.	21.70	0.41
31	Total number of vertebrae	Σ V.	61.62	1.08
32	Number of Dorsal fin Rays	#D.R.	10.08	0.29
33	Number of Anal fin Rays	#A.R.	14.98	0.28

(*) = measurements described by Berg (1948)

other samples was not as great as it otherwise would have been. The true values of the overall mean and the standard deviation of the weight measurement are seen in brackets.

(2) Transformed Data

The raw data was also transformed to produce ratios of the various measurements to others. The St.L. measurement was converted to its $\log_{10}$ form. The transformed values for each fish were individually calculated and their overall mean values are seen in Table 2. Some of the transformed variables were multiplied by one of two constants (0.1 and 10). This was done so that the resulting values would all be of the same order of magnitude. It can be seen from Table 2 that not only the size of the means but the standard deviations of the groups' means have been reduced.

(3) Age Distribution

The results of the age determinations for each population are seen in Table 3. The frequency of each age class was converted to a percentage to facilitate the comparison of the samples. A number of points emerge from this table. The N.A. (69) sample has a much higher percentage of four and five-year-olds than does any other sample.[1]

---

[1]The N.A. (69) sample, or simply, N.A. (69), refers to the sample of *O. e. mordax* taken from the Norris Arm sampling area (Materials and Methods [A]). The abbreviations used for the other smelt samples are similarly defined and referenced in Materials and Methods (A).

TABLE 2

CHARACTER CODE, OVERALL MEAN AND STANDARD DEVIATION OF THE GROUP MEANS, FOR THE TRANSFORMED DATA FROM THIRTEEN SMELT POPULATIONS

No.	Transformed Character Code	Overall Mean	Standard Deviation of Group Means
1	Wt./Age x 0.1	1.311 dag/year	.424
2	St.L./A.F.-Mx.	1.321	.012
3	$\log_{10}$ St.L.	2.134	.226
4	St.L./D.F.-Mx.	1.917	.030
5	St.L./Pl.F.-Mx.	1.843	.036
6	St.L./Pc.F.-Mx.	3.998	.110
7	St.L./P.V.R.	3.370	.076
8	H.D./B.D.	1.772	.128
9	St.L./L.Pc.F.	6.838	.380
10	St.L./L.Pv.F.	8.136	.374
11	B.A./H.A.	1.654	.104
12	St.L./Ad.-T.	6.561	.205
13	D.Cd.P./L.Hd. x 10	2.283	.101
14	L.Hd./St.L. x 10	2.510	.072
15	L.Or./L.Hd. x 10	2.336	.222
16	L.Sn./L.Hd. x 10	2.834	.143
17	L.Mx./L.Or. + L.Sn. x 10	8.810	.454
18	L.Hd./I.O.W.	4.433	.370
19	L.Mx./L.Hd. x 10	4.551	.191
20	L.Or. + L.Sn./L.Hd. x 10	4.171	.117
21	L.Or./I.O.W. x 10	10.415	1.821

TABLE 3

PERCENTAGE AGE DISTRIBUTION FOR THE VARIOUS SAMPLES

Sample \ Age	2	3	4	5	6
N.A.(69)	2	9	59	25	5
C.B.	30	60	5	5	0
L.R.	5	70	25	0	0
S.B.	50	35	15	0	0
I.R.	0	65	25	5	5
N.A.(70)	5	86	3	6	0
G.B.	5	67	22	6	0
G.L.	100	0	0	0	0
L.S.	38	36	24	2	0
L.H.	13	80	6	1	0
L.E.	3	74	18	3	2
L.H.G.	0	25	65	10	0
L.H.S.	100	0	0	0	0
Average *	27	47	20	5	1

*See page 232.

Eighty-nine percent of the individuals from N.A. (69) are more than three-years-old. The L.R. and S.B. samples differ especially in the high number of two-year-olds in the S.B. sample. The two stunted samples (G.L. and L.H.S.) contained exclusively two-year-olds. On the other hand, the L.H.G. sample contained four-year-olds at a greater than average frequency. Of the Great Lakes samples, the L.H. and L.E. samples had a relatively high frequency of three-year-olds while the L.S. sample presented an age distribution that was near the average age distribution over all the samples. The average percentage age distribution shows that approximately fifty percent of the fish in the study were three years old while approximately twenty-five percent of the animals were two and four years old. Only a few individuals lived more than four years.

(B) Discriminant Function Analysis

This study is concerned with the variation between 13 populations of smelt, *O. e. mordax*. Rather than trying to obtain measures of distinctness for these populations by making a series of comparisons amongst the individual morphometric characters, the analysis attempts, by the use of discriminant functions (Rao, 1952), to use all the measurements simultaneously. Therefore, when a decision is required regarding the distinctness of two populations, the total of the information available is used in making

that decision. Discriminant functions are linear functions which characteristically space out the means of the groups to a maximum.

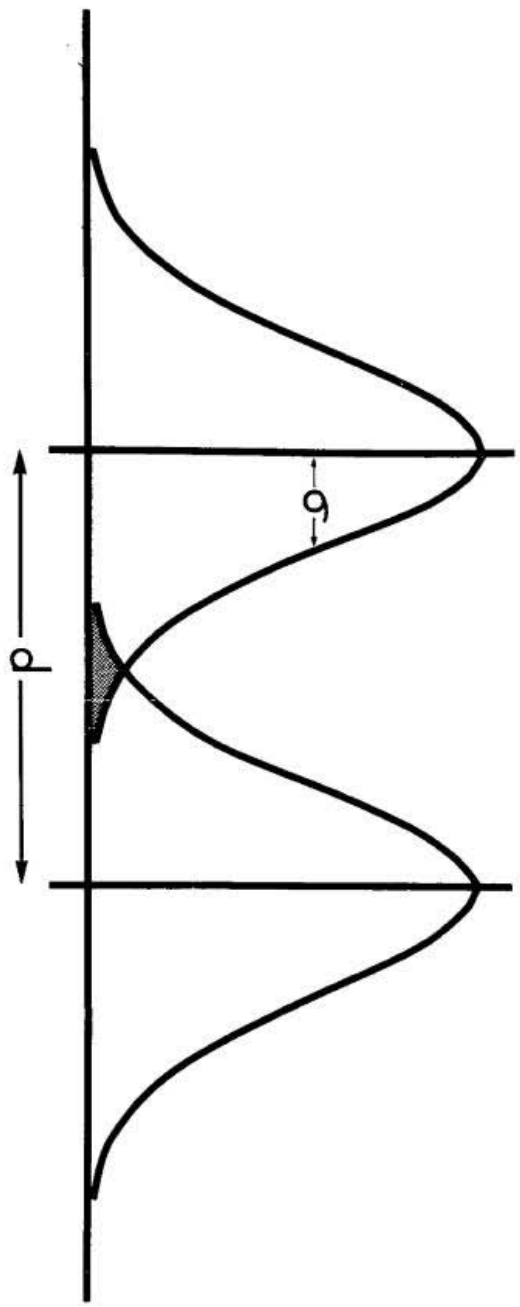
(1) Latent Roots and Vectors

The principle upon which this analysis is based is illustrated in Figure 1, which shows the frequency function of two Normal populations with a common standard deviation σ. If 'd' is the distance between the two means, the amount of overlap between the two populations decreases as d/σ increases. Any linear combination of the set of measurements, produces a single compound measurement from which a diagram such as Figure 1 can be constructed. To obtain the best linear combination, the coefficients should be chosen so that d/σ is maximum.

This approach generalises to the situation in which there are more than two groups of objects. Again, a compound measurement is sought which separates the group means as widely as possible. The ratio of the group means to the assumed common variance within the groups is taken as an overall measurement of the separation of the groups. The result is concisely expressed in matrix notation. For example, suppose that B is the matrix containing the variances and covariances of the group means, W is the variance-covariance matrix within groups and c is the vector of weights to be determined. c is chosen by maximizing the

FIGURE 1

DIAGRAM OF FREQUENCY FUNCTION OF TWO NORMAL POPULATIONS


q
p

variance ratio $c^T Bc / c^T Wc$ and this can be shown to be equivalent to solving the set of equations, $(B - \lambda W)c = 0$; from which the vector c can be determined (Gower, personal communication). The quantity λ, the latent root, is evaluated at the same time as c (Healy, 1965) and is the maximum value of the variance ratio. Therefore, λ indicates the extent to which the group means are dispersed along a single dimension. The dispersion is determined by the co-efficients, c, which have been calculated to make the separation maximal. An introduction to the mathematics involved is given by Sawyer (1966), with a more rigorous account of the calculations given by Anderson (1966) and Seal (1964).

If there are p measurements and m populations, the equation $(B - \lambda W)c = 0$ will have Min (p, m-1) solutions for c and non-zero λ's. The single optimum solution is the one associated with the largest value of λ. Next, a second compound measurement is sought which maximizes the remaining dispersion of the group means. The restriction is applied that the first compound measurements sought must be uncorrelated with the first set. This is given by the c vector, associated with the second largest value of λ, where λ measures the dispersion along dimension c. Further independent compound measurements could be obtained in the same way.

As λ is the extent to which the group means are

dispersed along a single dimension, the total extent of dispersal in p dimensions is represented by $\sum_{i=1}^{p} \lambda_i$. The extent to which a limited number of linear functions account for the total variability can be judged by comparing the sum of the λ's used to the sum of all p of the λ's. Bartlett (1947) has suggested a test for the hypothesis that all roots after the k^{th} root can be given values of zero. Calculations for this test are given in Seale (1968). In other words, although there are Min (p, m-1) non-zero roots, are they all significant, or are some of these not significant and can therefore be given zero value? When these tests of significance were conducted on the roots of some of the more critical analyses in this study, a general pattern emerged. It would found that Min (p, m-1)-2 of the roots were significant. That is, for ALLALL and SZOTME there were 10 significant roots, while in ESTALL, the first 6 roots were significant.*

(C) Summary of the Analyses in the Study, Latent Roots and Variables

For each of the major analyses, the latent roots and vectors were obtained. A summary of these results is seen in Table 4. From this table it can be seen that the sums of the latent roots are quite variable for the different analyses. This is because the various analyses contained both a different total number and a different set of

---

*See page 232.

TABLE 4

LATENT ROOTS AND ACCUMULATIVE PERCENTAGE O[illegible]
DISPERSION FOR EACH ANALYSIS PERFORMED
WITH DATA FROM SMELT POPULATIONS

Analysis	Sum of Latent Roots	Canonical Equation I		Canonical Equation II		Canonical Equation III		Canonical Equation IV		Canonical Equation V		Canonical Equation VI		Number of Variables Included	λ at which 95.0% of Variation Included
	($\Sigma\lambda$)	λ_I	Acc.%	λ_{II}	Acc.%	λ_{III}	Acc.%	λ_{IV}	Acc.%	λ_V	Acc.%	λ_{VI}	Acc.%		
ALLALL	64.17	49.35	76.9	5.25	85.1	2.47	89.0	2.12	92.3	1.49	94.6	1.06	96.2	33	6
ALLNO1	53.57	43.23	80.7	4.11	88.4	1.72	91.6	1.45	94.3	1.26	96.7	0.75	98.1	13	5
ALLNO2	50.68	40.24	79.4	4.10	87.5	1.94	91.3	1.64	94.5	1.28	97.1	0.67	98.4	13	5
ALLNO3	20.23	10.74	53.1	3.18	68.8	1.85	78.0	1.40	84.9	1.13	90.5	0.71	94.0	16	7
TGNMER	54.49	36.75	68.7	4.76	77.6	3.03	83.3	2.73	88.4	2.14	92.4	1.23	94.7	25	7
TGNALL	46.06	34.54	75.0	2.71	80.9	2.49	86.3	2.22	91.1	1.32	94.0	0.98	96.1	21	6
MORPHO	52.25	43.62	83.5	2.44	88.1	2.14	92.2	1.38	94.9	1.13	97.0	0.68	98.3	20	5
MERIST	9.17	7.10	77.4	0.83	86.5	0.57	92.7	0.24	95.3	0.20	97.4	0.11	98.6	10	4
FWTSWT	2.78	2.78	100.0											7	1
FWTALL	16.33	12.56	76.9	2.00	89.2	1.76	100.0							10	3
SWTALL	9.51	3.47	36.5	2.65	64.4	1.88	84.2	1.15	96.3	0.24	98.8	0.12	100.0	14	4
ALLXHM	14.85	5.88	39.6	2.77	58.2	1.95	71.4	1.50	81.5	1.06	88.6	0.77	93.8	20	7
TGNXHM	16.22	5.45	33.6	3.06	52.4	2.21	66.1	1.62	76.1	1.51	85.5	0.84	90.7	20	8
ESTALL	2.66	1.47	55.2	0.61	78.0	0.27	88.2	0.15	94.0	0.09	97.2	0.07	99.7	8	5

Analysis	Sum of Latent Roots	Canonical Equation I		Canonical Equation II		Canonical Equation III		Canonical Equation IV		Canonical Equation V		Canonical Equation VI		Number of Variables Included	λ at which 95.0% of Variation Included
	($\Sigma\lambda$)	λ_I	Acc.%	λ_{II}	Acc.%	λ_{III}	Acc.%	λ_{IV}	Acc.%	λ_V	Acc.%	λ_{VI}	Acc.%		
SZOTM0	21.54	13.55	62.9	2.31	73.7	1.89	82.4	1.28	88.4	1.03	93.2	0.70	96.4	19	6
SZOTME	29.06	17.23	59.3	3.93	72.8	2.23	80.5	1.97	87.2	1.26	91.6	0.93	94.8	19	7
3RAW01	22.81	19.63	86.1	2.21	95.8	0.96	100							3	3
3RAW02	11.69	8.38	71.7	2.26	91.0	1.05	100							3	3
3TGN01	13.89	11.19	80.6	1.70	92.8	1.00	100							3	3

variables. It was noted that a large sum of latent roots could be expected with the combination of all populations and morphometric data in an analysis. ALLALL through MORPHO (Table 4) are of this type, while MERIST through ESTALL do not contain the previously mentioned combination. The stunted samples were not considered in ALLXHM. It was noted that the sum of the roots for ALLXHM (14.85) was approximately equal to the sum of the roots from ALLALL minus the first root (64.17 - 49.35 = 14.82). This suggests that the dispersion associated with the first canonical equation in ALLALL is primarily involved with the separation of the stunted samples from the normal samples.

The accumulated percent columns (Table 4) indicate the level to which the variation in the data has been accounted for, by the inclusion of that particular root. With the data used in these analyses, it was common to have more than 85% of the variation accounted for with the addition of the third root. The root at which 95% of the variation has been accounted for is also seen (Table 4). It was noted that in all cases, the particular root was most probably significant on the basis of Bartlett's test (Section B, Results).

For example, it can be seen from Table 4 that in the analysis ALLALL, thirty-three measurements and counts were included when the canonical equation was solved. At

least 95.0% of the dispersion of the data was explained with the inclusion of the sixth latent root and its associated vector. On the other hand, 76.9% of the dispersion was accounted for by the first solution of the equation $(B - \lambda W)c = 0$. The first and second solutions of the equation, together, account for 85.1% of the dispersion while with the inclusion of the first six latent roots, 96.2% of the dispersion had been accounted for. All of these roots were significant.

Other interesting points emerge from Table 4. In the ALLALL through ALLNO3 series of analyses, the following was seen: the removal of the variable with the highest F-value resulted in decreases in the total dispersion of only 16.4% in the first case (ALLALL-ALLNO1) and of 5.9% in the second case (ALLNO1-ALLNO2). However, ALLNO3 showed a marked decrease in the dispersion; only 31% of the dispersion seen with the ALLALL analysis apparent in ALLNO3. Therefore, the majority of the dispersion in ALLALL, resulted from those three variables that were not included in ALLNO3 but were considered in ALLALL. These three variables were: the total number of the gill rakers, the depth of caudal peduncle and the weight. Of these, the weight, appears to make the greatest contribution to the total dispersion, since it was the measurement that was omitted between ALLNO2 and ALLNO3 (Table 5).

TABLE 5

FIRST FIVE VARIABLES INCORPORATED INTO THE DISCRIMINANT FUNCTIONS FOR EACH ANALYSIS PERFORMED WITH DATA FROM SMELT POPULATIONS AND THE LEVEL OF SIGNIFICANCE FOR THE HYPOTHESIS THAT THE GROUPS' MEANS ARE THE SAME WITH THE INCLUSION OF THE FIFTH VARIABLE

Analysis	First Five Variables Incorporated in the Discriminant Functions					
	First	Second	Third	Fourth	Fifth	Significance
ALLALL	Σ G.R.	D.Cd.P.	Wt.	St.L.	L.Or.	< .001
ALLNO1	D.Cd.P.	Wt.	St.L.	L.Or.	I.G.R.	< .001
ALLNO2	L.Pl.F.	Wt.	St.L.	I.G.R.	L.Or.	< .001
ALLNO3	L.Pl.F.	I.G.R.	L.Or.	St.L.	L.Sn.	< .001
FWTSWT	Σ G.R.	Σ Vert.	#D.R.	L.Pl.F.	L.Sn.	< .001
FWTALL	L.Mx.	St.L.	I.O.W.	L.Or.	Age	< .001
SWTALL	I.O.W.	L.Or.	Σ G.R.	L.Sn.	St.L.	< .001
TGNALL	$\log_{10}$ St.L.	Wt./Age x .1	Hd./St.L. x 10	Or./Hd. x 10	Or./I.O.W. x 10	< .001
TGNMER	$\log_{10}$ St.L.	Wt./Age x .1	Σ G.R.	Hd./St.L. x 10	Or./Hd. x 10	< .001
ALLXHM	Σ G.R.	L.Mx.	L.Or.	Wt.	L.Sn.	< .001
TGNXHM	Σ G.R.	Hd./St.L. x 10	Or./Hd. x 10	Or./I.O.W. x 10	Hd./I.O.W.	< .001
MERIST	Σ G.R.	Σ Vert.	Pd.L.L.	#D.R.	#Pc.R.	< .001
MORPHO	D.Cd.P.	Wt.	St.L.	L.Or.	L.Sn.	< .001

Analysis	First Five Variables Incorporated in the Discriminant Functions					
	First	Second	Third	Fourth	Fifth	Significance
ESTALL	1	7	6	4	5	< .001
SZOTME	Or./I.O.W. x 10	Σ G.R.	Hd./St.L. x 10	Or./Hd. x 10	Hd./I.O.W.	< .001
SZOTMO	Or./I.O.W. x 10	Hd./St.L. x 10	Or./Hd. x 10	Hd./I.O.W.	St.L./D.F.-Mx.	< .001
3RAW01	Σ G.R.	D.Cd.P.	Wt.	-	-	< .001
3RAW02	Σ G.R.	L.Or.	St.L.	-	-	< .001
3TGN01	Or./I.O.W. x 10	Σ G.R.	Hd./St.L. x 10	-	-	< .001

It is interesting to note the number of latent roots required to account for 100% of the dispersion. As a mathematical consequence of the analytical procedure, the number of non zero λ's is Min (p, m-1), where p is the number of variables and m is the number of samples included in the analysis. In the majority of the analyses, the number of variables included was greater than m-1 and therefore, 100% of the dispersion was accounted for with the inclusion of the $m-1^{th}$ root. In five of the analyses (MERIST, ESTALL, 3RAW01, 3RAW02 and 3TGN01), the number of variables was less than m-1 and 100% of the dispersion in these analyses was accounted for with the inclusion of the p^{th} root.

In Table 5, the first five variables which were incorporated into the discriminant functions are given. For each analysis, the significance level is given for the hypothesis that the means of all the samples are equal, after the fifth variable had been included. It was noted that this level of significance was often achieved even when fewer variables were contained in the discriminant function. This table indicates that on the basis of any of a host of variables, the populations from which the samples were drawn are not uniform with respect to each other.

(D) Classification Matrices[1]

(1) ALLALL Analysis

After each step in each analysis, discriminant functions were generated for each group. By the use of these, it was possible to produce a classification matrix.* Each individual was assigned to the group whose coefficients and constant resulted in the maximum value when the discriminant functions were solved. Each cell of the matrix contains the number of animals from each sample (indicated in the column) that were classified into each group (top row). Table 6 shows the classification matrix resulting from the inclusion of the first two variables (Table 5) in the discriminant functions of ALLALL. For example, there were fourteen fish from the N.A. (69) sample classified into the N.A. (69) group. No fish from the N.A. (69) sample were classified into the C.B., L.R. or S.B. groups. However, five were classified into the I.R. group and one into the N.A. (70) group. Since there were twenty fish in each sample, the sum of all rows is twenty.

---

[1]The use of the word "classification" in "classification matrix" and "correct classification" follows the convention used at the Health Sciences Computing Facility, U.C.L.A., California. However, in this context, the use of the word "identification" would be more precise, though less common (Kendall and Stuart [1966]). Therefore, "to correctly classify" is properly defined as "to correctly identify and assign."

*See page 232.

TABLE 6

CLASSIFICATION MATRIX OF SMELT SAMPLES
FROM THE ANALYSIS ALLALL, TWO
VARIABLES CONSIDERED

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.	L.H.G.	G.L.	L.H.S.
N.A.(69)	14	0	0	0	5	1	0	0	0	0	0	0	0
C.B.	0	14	1	3	0	2	0	0	0	0	0	0	0
L.R.	1	6	3	3	2	3	0	2	0	0	0	0	0
S.B.	0	5	3	4	3	4	1	0	0	0	0	0	0
I.R.	6	0	2	3	3	2	1	2	0	0	1	0	0
N.A.(70)	1	1	6	1	2	9	0	0	0	0	0	0	0
G.B.	0	1	0	1	2	0	6	2	0	2	6	0	0
L.S.	0	2	1	0	1	0	1	5	2	5	3	0	0
L.H.	0	0	0	1	0	0	0	2	17	0	0	0	0
L.E.	0	0	1	1	0	0	2	4	0	12	0	0	0
L.H.G.	0	0	0	1	0	0	3	0	0	2	14	0	0
G.L.	0	0	0	0	0	0	0	0	0	0	0	15	5
L.H.S.	0	0	0	0	0	0	0	0	0	0	0	1	19

To determine the number of animals correctly classified in each matrix, the diagonal elements were summed. Thus for Table 6, 135 animals were placed in their correct groups. This represents 51.9% of the animals correctly classified. The number of animals correctly classified on the basis of chance alone would have been 20, (260 x 1/13) which represents 9.1% of the sample. Therefore, there is more than a five fold improvement over chance, with the classification based on only two characters. Other interesting points emerge from this matrix. Two groups, each containing two samples, were geographically and temporally quite close together (nearest neighbours). These were the I.R. and N.A. (70) group and the L.R. and C.B. group. If these two groups are combined, the percentage of correct classification on the basis of two characters, rises to 55.7. Secondly, the matrix can be partitioned into regions of correct classification for the salt water samples and the freshwater samples.[1] That is the upper left 7 x 7 matrix represents those animals from salt water samples that were classified into salt water groups. The lower right 6 x 6 matrix represents those fish from freshwater samples that were classified into a freshwater group. Therefore, the sum of all elements of these two submatrices represents

---

[1]Throughout this thesis, "salt water samples" and "freshwater samples" will be used to refer to the samples of *O. e. mordax* which lived primarily in the marine-estuarian environment or exclusively in freshwater, respectively.

the number of salt water and freshwater fish which were correctly identified as such. There were 233 of the 260 animals in these two submatrices (Table 6) which represents 89.6% of the individuals correctly identified. A further subdivision of the freshwater submatrix is possible. The extreme lower right 2 x 2 matrix contains the number of fish from the two stunted populations which were so classified. With one variable included, this value was 92.5% while with the inclusion of the second and subsequent variables, this value became 100%.

When thirty-two variables were included in the discriminant functions, the classification matrix seen in Table 7 resulted. From this table, it can be seen that the number of fish correctly classified had risen to 243 which represented 93.4% correct classification. The L.R. and S.B. samples and the I.R. and N.A. (70) samples were very similar with respect to their geographic area and time of capture. The combination of these nearest neighbours increased the correct classification to 96.5%. Salt water fish were correctly separated from the freshwater ones, at a 99.2% level. The two stunted samples were again completely distinct from the rest of the populations under consideration.

The percentage of fish correctly classified for each successive classification matrix, in the ALLALL analysis is seen in Figure 2. The percentages for the separation of the

TABLE 7

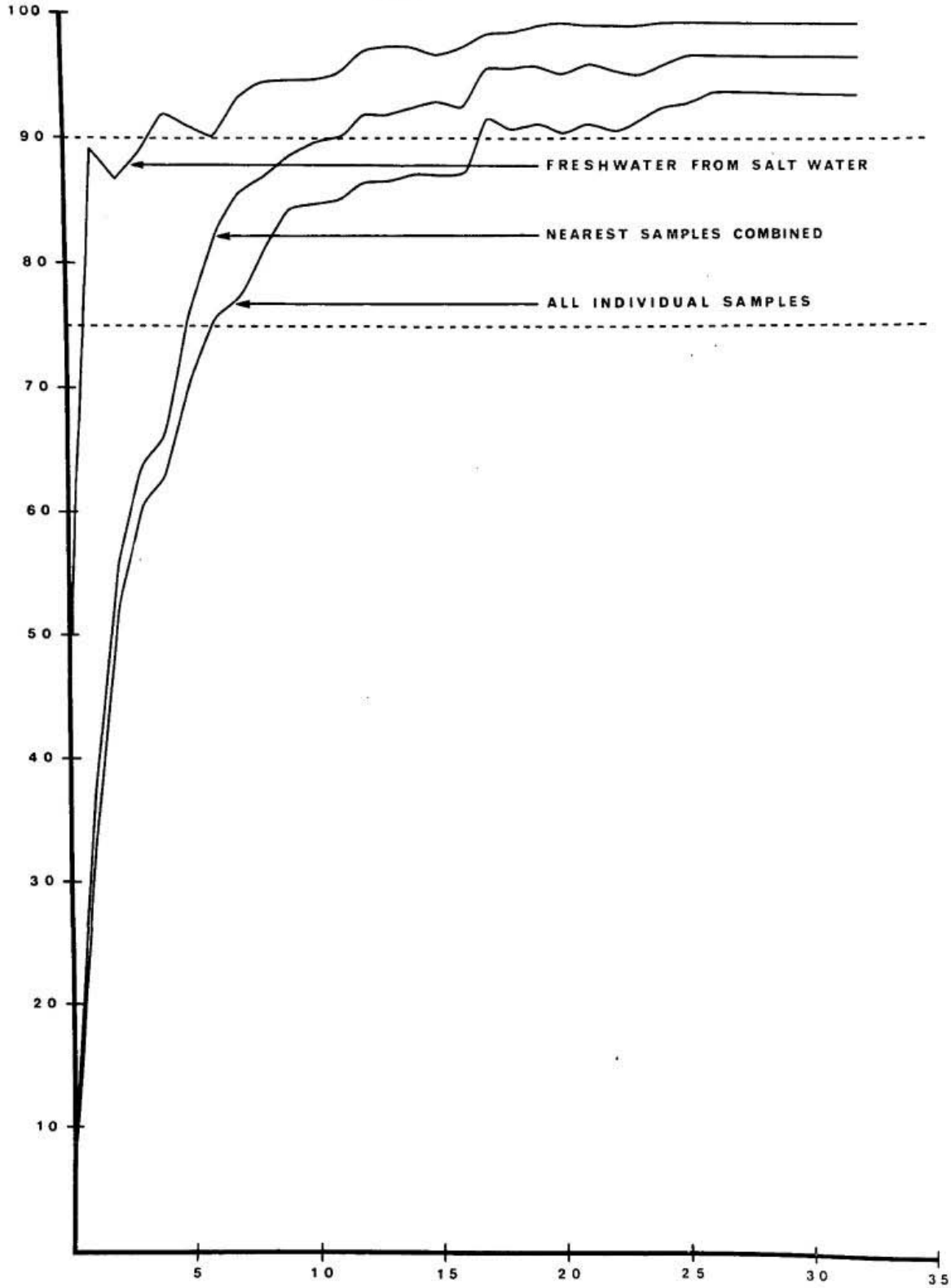
CLASSIFICATION MATRIX OF SMELT SAMPLES FROM

ALLALL, ALL (32) VARIABLES CONSIDERED

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.	L.H.G.	G.L.	L.H.S.
N.A.(69)	18	0	1	0	1	0	0	0	0	0	0	0	0
C.B.	0	20	0	0	0	0	0	0	0	0	0	0	0
L.R.	0	0	17	3	0	0	0	0	0	0	0	0	0
S.B.	0	0	3	17	0	0	0	0	0	0	0	0	0
I.R.	0	0	0	1	16	1	2	0	0	0	0	0	0
N.A.(70)	0	0	0	1	1	18	0	0	0	0	0	0	0
G.B.	0	0	0	0	0	0	20	0	0	0	0	0	0
L.S.	0	0	0	1	0	0	0	19	0	0	0	0	0
L.H.	0	0	0	0	0	0	0	0	19	1	0	0	0
L.E.	0	0	0	1	0	0	0	0	0	19	0	0	0
L.H.G.	0	0	0	0	0	0	0	0	0	0	20	0	0
G.L.	0	0	0	0	0	0	0	0	0	0	0	20	0
L.H.S.	0	0	0	0	0	0	0	0	0	0	0	0	20

FIGURE 2

GRAPH SHOWING PERCENTAGE OF INDIVIDUAL SMELT CORRECTLY CLASSIFIED VERSUS THE NUMBER OF VARIABLES INCLUDED IN THE DISCRIMINANT FUNCTIONS FOR THE ANALYSIS ALLALL


100
90
80
70
60
50
40
30
20
10
5
10
15
20
25
30
35
FRESHWATER FROM SALT WATER
NEAREST SAMPLES COMBINED
ALL INDIVIDUAL SAMPLES

stunted populations were not plotted, as for the second and subsequent matrices, this value was 100%. The plot for individual populations and nearest neighbours combined rises quite sharply with the inclusion of the first seven to nine variables. The increase in the percent of correct classification continues to increase slowly until a plateau is reached at about the seventeenth step. The inclusion of the next 15 variables only increases the correctness by one or two percent. The plot of the freshwater versus salt water classifications exceeds the 90% correct level with the inclusion of the fourth variable. A plateau is also achieved in this case, at about the seventeenth step.

The 75 and 90 percent lines, are indicated on the graph. Thus, it can be seen that for all populations classified individually, six variables were required for 75% correct classification while seventeen were required for 90% correct classification. With the combination of nearest neighbours, only five and eleven variables are required to produce the same results. Almost 90% separation of salt water and freshwater forms results on the basis of the first variable, which has the highest F-value.

It should also be noted that although there is a general increase in the percentage of correct classification with an increase in number of variables, there are occasional dips in these curves. A closer investigation of Figure 2 reveals that where dips occur in one line, there are usually

rises in one or both of the other lines. This provides an indication as to the cause of the dips. For example, the second variable entered decreases the percentage of correct classification in the salt water from freshwater distinction while the percent correct classification of the other two comparisons rises sharply. Therefore, the size of the F-value for the second variable is not so much the result of the differences between the salt water and freshwater samples but rather the result of differences between all the individual samples and the overall mean. It can also be said that with respect to variable 2 (ALLALL, Table 5) some freshwater individuals resemble some salt water samples more than the other freshwater samples. They are, therefore, incorrectly classified into a salt water group on the basis of variable 2. A similar statement can be made for salt water individuals incorrectly classified into freshwater groups.

The relationship between the total dispersion and correct classification will be dealt with later (Section O).

(2) Classification Summary for All Analyses

For each analysis, a graph, similar to Figure 2 was obtained. The shapes of the plots from the various analyses were similar to those obtained from ALLALL. However, the rates of increase for the various analyses differed as did the maximum percentages that were obtained. A summary of the results of these graphs is seen in Table 8. For each analysis, four comparisons of the data are given. The

TABLE 8

SUMMARY OF THE CLASSIFICATION VALUE OF THE
VARIABLES IN THE VARIOUS ANALYSES
CONDUCTED WITH DATA FROM
SMELT SAMPLES

Analysis	All Populations Individually			Nearest Neighbours Combined			Freshwater from Salt Water			Stunted from Others		
	75%	90%	Max.%	75%	90%	Max.%	75%	90%	Max.%	75%	90%	Max.%
ALLALL	6	17	93.4	5	11	96.5	1	4	99.2	1	1	100
ALLNO1	7	NR	86.8	5	10	92.0	3	5	96.9	1	1	100
ALLNO2	6	NR	87.3	5	10	92.6	3	4	97.7	1	1	100
ALLNO3	6	NR	89.6	4	10	95.4	2	2	98.6	1	1	100
TGNMER	5	21	92.7	5	11	96.2	3	4	98.8	1	1	100
TGNALL	9	NR	86.9	7	14	93.1	3	10	95.4	1	1	100
MORPHO	9	NR	85.0	6	NR	89.2	5	6	94.1	1	1	100
MERIST	NR	NR	61.1	NR	NR	66.2	1	2	92.7	1	1	97.5
FWTSWT	-	-	-	-	-	-	1	2	95.8	-	-	-
FWTALL	2	4	98.8	-	-	-	-	-	-	-	-	-
SWTALL	5	NR	84.3	4	8	93.6	-	-	-	-	-	-
ALLXHM	7	20	90.0	5	9	95.0	1	4	98.6	-	-	-

Analysis	All Populations Individually			Nearest Neighbours Combined			Freshwater from Salt Water			Stunted from Others		
	75%	90%	Max.%	75%	90%	Max.%	75%	90%	Max.%	75%	90%	Max.%
TGNXHM	7	NR	88.2	6	11	95.0	1	6	98.6	-	-	-
ESTALL	NR	NR	37.4	NR	NR	41.1	1	NR	89.3	-	-	-
SZOTMO	10	NR	82.7	8	NR	87.3	1	10	93.5	1	1	100
SZOTME	7	NR	89.2	4	14	92.7	1	2	97.7	1	1	100
3RAW01	NR	NR	60.4	NR	NR	63.8	1	1	90.8	1	1	100
3RAW02	NR	NR	68.8	NR	NR	73.4	1	1	90.8	1	1	100
3TGN01	NR	NR	63.5	NR	Nr	70.3	2	2	90.8	1	1	100

NR = the particular level of correct classification was not reached

- = the particular comparison was not applicable to the analysis

first was obtained by considering each sample individually. The second resulted by combining N.A. (70) with I.R. and L.R. with S.B. The third and fourth comparisons represent the results of the two partitions of the matrix that were done for each step in each analysis (salt from freshwater, stunted from normal). Each comparison contains the number of variables that were included in the discriminant functions, when the percentage of correct classification was equal to or greather than the levels indicated. The maximum percent of correct classification is also indicated for each comparison. This often occurred when all variables were included in the discriminant function, however, it occasionally occurred before this point, in instances where the graph included an upper plateau.

Three N.R.'s appear in the 90% column for ALLNO1 through ALLNO3. This may be a bit misleading, with regard to the classification value of the variables involved, if direct comparisons with ALLALL are made. Table 4 has shown that these three analyses contained 13, 13 and 16 variables respectively when the analyses were terminated. When there were 13 and 16 variables included in the discriminant functions of ALLALL, the levels of correct classification were 86.5% and 87.2% respectively. Thus, the variables in ALLNO1 through ALLNO3 in fact compare quite favourably with those of ALLALL with regards to their classification value. Therefore, it appears that the 90% level would have been reached in these three analyses with the inclusion of

at the most, 20 variables. In other cases where N.R. occurs, (TGNALL, MORPHO, MERIST, SWTALL, TGNXHM and ESTALL), direct inference seems justified. In SWTALL and TGNXHM, all possible variables were not included at the termination of the analyses. However, in both cases, the number was approximately equal to twice the number of samples considered in these analyses. This should have been adequate to indicate the upper limit of correct classification available from all the data in these analyses. The data in ESTALL exhibited the poorest classification value. These particular results will be dealt with more extensively later in the text.

(E) Plotting Samples Along Canonical Axes

The Stepwise Discriminant Analysis programme produced as part of its results a set of canonical equations in the general form

$$y_{ij} = c_{1i}x_{1j} + c_{2i}x_{2j} + c_{3i}x_{3j} + \ldots c_{pi}x_{pj} \qquad (1)$$

where y_{ij} = the j^{th} value of the i^{th} canonical variate

$i = 1 \ldots p$ $\quad$ p = the number of variables considered

$j = 1 \ldots m$ $\quad$ m = the number of units considered

From the equation $(B - \lambda W)c = 0$, there are p different values for λ and the vector c for which the equation can be solved. If $m \leq p$, then $p - m+1$ of the λ's will be zero. In the analysis ALLALL, 33 variables were included in the

canonical equations for 13 populations. Therefore, there were twelve (13-1) different sets of coefficients actually generated for the 33 possible canonical equations. As has been seen in Table 4, 89% of the total dispersion of the populations in ALLALL is accounted for by the first three canonical variate axes and the remaining 11% is accounted for by the nine remaining axes. Thus, the first three axes hold the majority of the information available concerning the dispersion of the populations. From Table 4, it can also be seen that for ALLALL, 76.9% of the dispersion is accounted for by the first canonical axis, 8.2% is accounted for by the second and 3.9% by the third. These values then, give the relative importance of each of these three axes in explaining the variation between the populations.

In this study, as many as 33 measurements were used. The set of measurements on each single animal can be represented by a point in p-dimensional space, where p equals the number of measurements considered. A p-dimensional representation accounts for all the variation in the samples, but is rather difficult to visualize. Therefore, it is natural to investigate whether a diagram with fewer dimensions, preferably three or less, could depict the situation without the sacrifice of too much of the essential information.

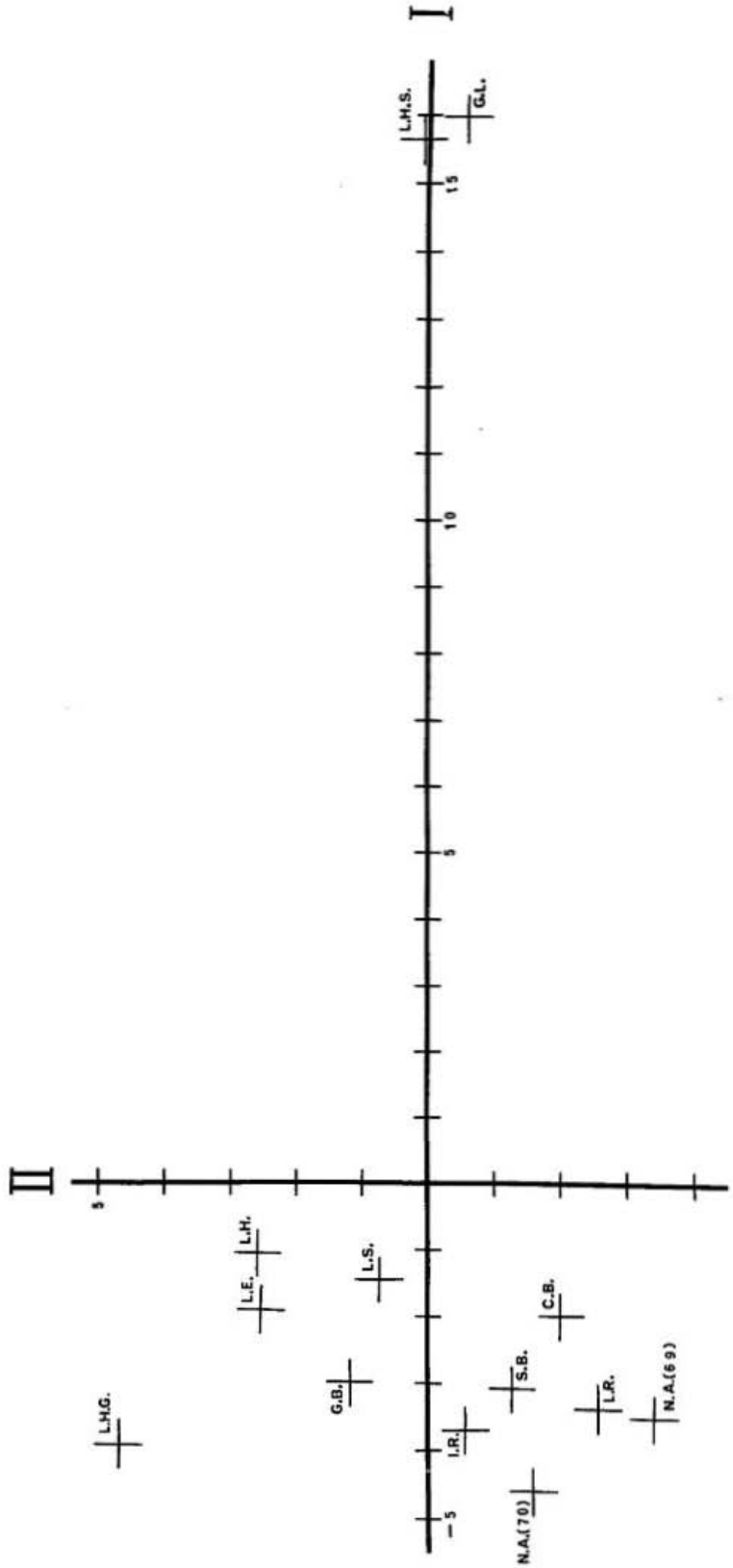
(1) Two Dimensional Representation

Figure 3 is a two dimensional representation of the

FIGURE 3

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II

Arms of Crosses Extend for One Standard Error of the Mean


II
I
5
-5
10
15
L.H.G.
L.E.
L.H.
L.S.
G.B.
I.R.
N.A.(70)
S.B.
C.B.
L.R.
N.A.(69)
L.H.S.
G.L.

dispersion of the samples resulting from the ALLALL analysis. The coordinates for each population along the first and second canonical axes were found in the following manner. The means of each sample's measurements were substituted in Equation 1. Thus, the value y_{ij} represents the i^{th} coordinate for the mean of the j^{th} sample. Similarly, by substituting the mean values with the third and subsequent sets of canonical coefficients, it is possible to find the coordinates for the sample means in the third through p^{th} dimension. The arms of the crosses which are centered on the mean of each sample extend for one standard error of the mean. They were computed as in Seal (1964) and after Gower (personal communication).

From Figure 3, some interesting relationships become apparent. Firstly, the two stunted samples (L.H.S. and G.L.) are more than 17 units removed from the nearest sample (L.H.) along the first canonical variate axis. Secondly, the salt water samples are found, rather closely grouped, in the bottom left quadrant. The G.B. sample is the exception to this. Thirdly, the freshwater samples are found in the upper left-hand quadrant. With the exception of L.H.G., they seem to be slightly more displaced toward the L.H.S.-G.L. group than are samples from salt water. In the freshwater quadrant, the L.H.G. sample appears somewhat remote.

(2) Three Dimensional Model

The two dimensional representation, such as is seen

in Figure 3, represents clearly the dispersion of the means of the samples. However, it does not suggest how tightly the individuals in a sample are distributed, with respect to their mean. To represent this distribution, it is customary to draw circles about the mean which enclose 90% of the individuals in each sample. The radii of these circles are found in the following manner. For a k-dimensional representation, the squared radius of the confidence region is χ^2_k at the percentage level of interest (Gower, personal communication). For a two dimensional representation, at the 90% level, χ^2 with two degrees of freedom is 4.61 and the radius of the circles would be 2.14. If one was interested in enclosing only 70% of the individuals, the radii of the circles would be 1.56. In either case, the result of drawing the boundaries would be an interlocking mesh of circles which does little to clarify relationships, especially when several samples are involved in the study. This is partially due to the diameter of the circles and partially to the fact that the multidimensional dispersion has been compacted and stacked into only two dimensions.

In an attempt to more clearly represent the multi-dimensional dispersion, the following was done. First, the samples' means were used with the third set of canonical coefficients, to determine each samples' coordinate along the third canonical variate axis. Therefore, it was possible to position each sample's mean in three dimensional space.

Secondly, by using the measurements of the individuals with the sets of canonical coefficients, it was possible to determine the position of each individual in each sample. This was done for their positions along the first and second axis. Therefore, it was possible to calculate for each sample the actual dispersion (standard deviation) which occurred for the individuals about their mean. These results are seen in Table 9. From columns, two and three it is apparent, that some of the samples are more diffuse than others. For example the L.H.S. sample is much more compact than the L.H.G. sample. Secondly, it appears that there is a trend for both standard deviations to be of approximately the same size for each sample. The L.E. sample is the greatest exception to this. However, when the averages of the samples' standard deviations are taken, they appear to be very close (.970 and .972). This was expected because of the condition for no correlation between the first and second compound measurements (Section B, Results). It also seems reasonable to assume that the standard deviation along the third axis would be of the same size as along the first two. This in effect means that in two dimensions the sample's dispersion can be represented by a circle while in three dimensions it would be represented by a sphere.

A three dimensional model was constructed, using for each sample, the coordinates which had previously been calculated. The scale used in the construction was one unit of standard deviation equals one inch. Two inch diameter styrofoam balls were used to represent the sample. For a

TABLE 9

STANDARD DEVIATION OF THE INDIVIDUALS OF EACH SMELT SAMPLE ABOUT THEIR MEANS, FOR CANONICAL AXES I AND II

(ALLALL)

Sample	σ_I	σ_{II}	$\bar{\sigma} = \frac{\sigma_I + \sigma_{II}}{2}$
N.A.(69)	1.135	1.220	1.177
C.B.	.974	.768	.871
L.R.	1.364	.934	1.149
S.B.	.986	.896	.941
I.R.	1.116	1.010	1.063
N.A.(70)	.723	.852	.788
G.B.	.802	.922	.862
G.L.	.820	1.096	.949
L.S.	1.049	1.186	1.118
L.H.	1.083	.720	.902
L.E.	.840	1.468	1.154
L.H.G.	1.301	1.037	1.169
L.H.S.	.410	.528	.469
Average	.970	.972	.971

σ_I = sample's standard deviation along Canonical axis I

σ_{II} = sample's standard deviation along Canonical axis II

$\bar{\sigma}$ = average of σ_I and σ_{II}

three dimensional representation a sphere with a radius of one unit would contain more than twenty percent of the individuals of the sample. A three inch diameter sphere would have contained fifty percent of the individuals while a five inch diameter sphere would have been required to enclose ninety percent of the individuals in each sample. Thus, the two inch balls represent an inner shell of the sample density sphere.

Photographs of the three dimensional model are seen in Figure 4 (A), (B) and (C). All three photographs were taken from the positive side of axis III. In Figure 4 (A), the majority of the samples appear to form a rather compact crescent-shaped cluster. The two stunted samples are far removed while the N.A. (69) and to a lesser extent the L.H.G. samples are apart from the main body of the cluster. The balls in Figure 4 (B) approximate the results of the two dimensional plot (Figure 3). In this case, the apparent circles enclose approximately fifty percent of the individuals in each sample. The inferences that can be drawn from this figure have already been made from Figure 3. From Figure 4 (C), it can be seen that the crescent-shaped cluster of Figure 4 (A) is actually composed of two clusters. Again, the two stunted populations, N.A. (69) and L.H.G. are remote from the two central clusters. The difference between the two clusters seems to be partially explained on the basis of size. The samples on the left side of the space separating

FIGURE 4

THREE DIMENSIONAL MODEL OF THE DISPERSION OF THE SMELT SAMPLES IN THE ANALYSIS ALLALL

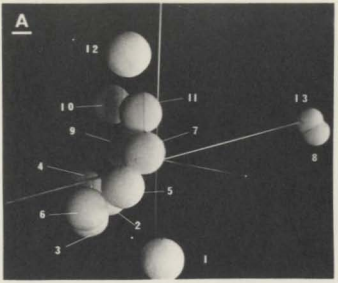
A Upper left hand view

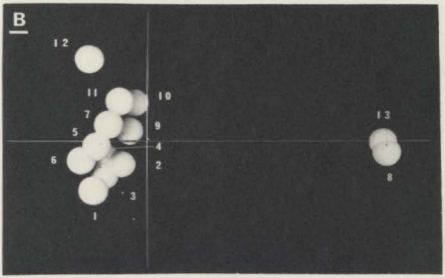
B View along canonical axes III

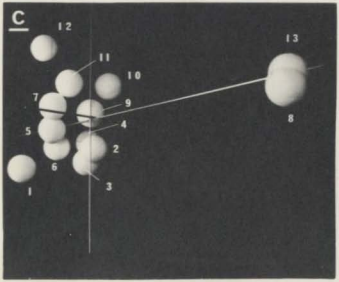
C Bottom right hand view

Legend

1 = N.A. (69)	6 = N.A.(70)	10 = L.H.
2 = C.B.	7 = G.B.	11 = L.E.
3 = L.R.	8 = G.L.	12 = L.H.G.
4 = S.B.	9 = L.S.	13 = L.H.S.
5 = I.R.		


A
12
11
13
10
9
7
8
4
5
6
2
3
1


B
12
11
10
7
9
13
5
4
6
2
8
3
1


C
12
13
11
10
7
9
8
5
4
2
6
3
1

the two clusters (Figure 4 C) are larger than those on the right. The average lengths of the two groups were 163 and 139 mm respectively. This suggests that overall size may be responsible for the separation. This grouping of the samples is similar to that obtained from the analysis TGNALL, in which $\log_{10}$ St.L. and Wt./Age were the two most important variables. Both of these variables are concerned with size.

(F) Confidence Region for Sample Means

Although only about 20% of the individuals of each sample are enclosed by the spheres of the model, the spheres have a greater significance than might be expected on that basis. The confidence region of the sample mean is also related to square root of Chi Square with k degrees of freedom. However, in the case of the sample means, this value is divided by the square root of the number of individuals in each sample (Seal, 1968). Therefore,

$$r = \frac{\sqrt{\chi^2_k}}{\sqrt{N_l}} \qquad (2)$$

where, r = radius of the confidence region of the mean

N_l = the number of individuals in sample l

χ^2_k = Chi-Square with k degrees of freedom at the significance level required.

In this study, there were twenty indiviudals per sample. Thus, with a radius of 1, and a sample size of 20 χ^2 will be 20. For the three dimensional representation

(k=3), (Figure 4 A, B, C), the two inch (unit) diameter balls represent the >99.99% confidence region for the position of the sample means. The 90% confidence region for the samples' means could have been represented by balls with a radius of .56 units, or a little more than one-half the size of the spheres used in the model. Thus for significant separation of the means, the 90% spheres would have to be 1.12 units apart (2 x .56). This is the case for all the samples in the three dimensional model, with the exception of the two stunted populations which are only .087 units apart (Table 10). When twelve dimensions are used in the representation (the point at which the accounted for dispersion >99.9% of the total), the 90% confidence spheres would have radii of 1.02, with twenty individuals per sample. Thus, the means would be 90% separated at a distance of 2.04 or greater. All the intersample distances of Table 11 are greater than this value and the 90% confidence regions of the sample means do not overlap.

(G) Calculation of the Intersample Distances

When the canonical analysis was done, a set of canonical equations was generated which had the general form seen in Equation 1. The solution of the set of equations produced a set of coordinates for the unit under consideration. Thus, if one was interested in obtaining the set of coordinates for an individual, the values of the measurements

made on that individual would be used for the corresponding x values in the equation. On the other hand, if the position of the population mean was sought, the value of the group's mean for each measurement would be used as the appropriate values for the different values of x. Therefore, to find the complete set of coordinates for the unit of interest, the appropriate values of x would be substituted with all sets of canonical coefficients to yield a set of k co-ordinates for that unit.

In this study, the complete set of coordinates was generated for the population means of all populations, in each analysis. Once a set of coordinates was available for each unit, it was possible to determine the distance between any two units. To do this, the following equation was solved:

$$d_{ij} = \left(\sum_{l=1}^{n} (y_{li} - y_{lj})^2 \right)^{\frac{1}{2}} \qquad (3)$$

where d_{ij} = the distance between the i^{th} and j^{th} units

n = the number of coordinates to be considered

y_{li} = the value of the l^{th} coordinate for unit i

If m units are considered, the solution of this equation yields a triangular matrix containing, $(m \times \frac{m-1}{2})$ elements. The nature of this relationship between the number of units and the size of the inter-unit distance matrix, restricted the calculations and subsequent analysis, to the consider-ation of only the distances between population means. For example, if 13 units (population means) are used the matrix

contains (13 x 12/2) = 78 elements, while if there were 20 individuals in each of the 13 populations there would be the inter-unit distances of the 260 units to calculate. This would have resulted in a matrix containing 33,670 elements. This matrix is 431.7 times larger than the one obtained from the use of coordinates of the population means and is too larger for use in further analysis.

The distance between population means can be calculated using any number of coordinates in the coordinate sets. Thus, if the distances between the means of the populations was sought for a three dimensional representation, the first three coordinates would be used from the set of k coordinates. Similarly, inter-population distances can be calculated using 1 through k coordinates for each population. If all coordinates are used, (n = p), the distance between the i^{th} and j^{th} population is D_{ij} (Mahalanobis' D^2_{ij}) using fewer dimensions (coordinates), the distances give approximations to D_{ij} that are the best possible in the least-squares sense, that is, the sum of the squares of the residuals is minimum (Gower, personal communication). Table 10 is the distance matrix obtained by using the coordinates from the first three canonical axes for each population from the analysis ALLALL. When each population's coordinates from all canonical axes of the ALLALL analysis were used in the distance computations, the inter-population distance matrix seen in Table 11 resulted. In this matrix, each element is D_{ij}, or the square root of Mahalanobis' distance.

TABLE 10

DISTANCES BETWEEN SMELT SAMPLES' MEANS COMPUTED
FROM THE FIRST THREE CANONICAL VARIATE
AXES OF THE ANALYSIS ALLALL

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.	L.H.G.	G.L.
C.B.	5.28											
L.R.	4.86	1.52										
S.B.	5.58	1.38	1.41									
I.R.	3.50	3.62	3.45	3.23								
N.A.(70)	4.05	3.03	2.09	2.29	1.97							
G.B.	5.03	4.43	4.75	4.00	1.90	3.56						
L.S.	6.26	2.89	3.92	2.77	3.40	4.03	2.74					
L.H.	8.53	4.78	5.77	4.41	5.41	5.90	4.27	2.32				
L.E.	6.81	5.01	5.72	4.61	3.70	5.01	1.94	2.31	2.79			
L.H.G.	8.33	7.57	7.87	6.81	5.26	6.52	3.61	5.17	5.01	3.04		
G.L.	19.96	18.22	19.65	19.33	19.79	20.72	19.14	17.61	17.64	18.25	20.65	
L.H.S.	19.83	17.91	19.36	18.98	19.51	20.43	18.81	17.22	17.15	17.85	20.23	0.87

TABLE 11

GENERALISED DISTANCES (D) BETWEEN SMELT SAMPLES' MEANS
COMPUTED FROM TWELVE CANONICAL VARIATE AXES
OF THE ANALYSIS ALLALL

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.	L.H.G.	G.L.
C.B.	7.23											
L.R.	5.59	5.26										
S.B.	5.98	5.13	2.78									
I.R.	5.09	5.20	5.09	4.71								
N.A.'(70)	6.05	4.91	4.89	4.50	3.19							
G.B.	6.61	6.37	6.25	5.27	3.78	4.63						
L.S.	6.83	6.33	5.90	4.33	5.91	6.35	5.45					
L.H.	8.87	7.01	6.82	5.50	6.06	7.16	5.68	4.85				
L.E.	7.54	7.03	6.65	5.55	5.02	6.09	4.34	4.46	4.60			
L.H.G.	9.07	8.24	8.40	7.58	6.55	7.83	6.00	7.11	6.84	5.60		
G.L.	20.16	18.64	19.92	19.57	19.99	21.07	19.45	18.02	17.84	18.50	21.00	
L.H.S.	20.14	18.42	19.63	19.17	19.77	20.63	19.14	17.78	17.64	18.23	20.47	4.50

As has been seen in Table 4, 89% of the total dispersion of the populations was accounted for by the first 3 canonical axes. The sum of all distances in Table 10 is 731.6. The sum of all the inter-population distances from Table 11 is 842.5 and this represents the sum of the distances between all populations using all axes. Therefore, the distances produced from the coordinates of the first 3 axes represents 87% of the total inter-population distances. This is in general agreement with the previous results (Table 4).

(H) Cluster Analysis (ALLALL)

(1) Cluster Analyses

When the interpopulation distance matrices had been computed, they were the starting points for further analyses. The use of Dr. Ross's programmes, MST and SLCA, required the conversion of a distance matrix to one of similarity coefficients. This was done using the formula,

$$S_{ij} = 1 - \frac{d_{ij}^2}{d_{max}^2} \qquad (4)$$

where S_{ij} = is the similarity coefficient between the i^{th} and j^{th} sample

d_{ij}^2 = the square of the distance between the i^{th} and j^{th} sample

d_{max}^2 = the square of some distance equal to or greater than the largest intersample distance.

This is the suggested method of conversion (Ross, personal communication).

Two types of cluster analyses were performed: Single Linkage Cluster Analysis, (S.L.C.A.), Sneath (1957), and Unweighted Pair Group Method Analysis, (U.P.G.M.A.), Sokal and Michener (1958). The S.L.C.A. operates by joining together into disjoint sets those individuals whose distance apart is less than some threshold distance, d_i. The value of d_i can be set at zero and then incremented in either a continuous or a discrete fashion. When the former is done, two individual samples will form clusters, or individual samples will join preexisting clusters, one at a time. If the increments are discrete, many samples may form clusters or join preexisting clusters at one step. This method of cluster analysis, tends to form elongated clusters. This is because the distance between an individual sample and any member of an existing cluster must be less than or equal to the threshold distance. For two clusters to join, it is only necessary for one of the inter-sample distances, between the two clusters, to be less than or equal to the threshold distance.

On the other hand, U.P.G.M.A. tends to form compact clusters. An incremented distance threshold is also used as the criterion by which it is decided whether an individual may join a preexisting cluster. However, in this case, it is the average distance of the individual with all the members

of the preexisting clusters which is judged against the threshold value. If two clusters are to join, the average intersample distance of all the members of both clusters must be equal to or less than the threshold value.

A Minimum Spanning Tree (M.S.T.) can be constructed whilst computing a S.L.C.A. The M.S.T. is a system of interconnected lines containing no closed loops.[1] The M.S.T. line system represents the minimum distance that is required to connect all the points into a tree. This tree, in conjunction with U.P.G.M.A. is of great help in understanding the relationships that exist between the samples in the analyses.

(2) Three Canonical Axes (ALLALL)

The results of the M.S.T. and U.P.G.M.A. from the three canonical variate distance matrix are seen in Figure 5 (A) and Figure 5 (B) respectively. Figure 5 (A) shows the M.S.T. linking the sample means which are plotted in the first two canonical axis of the analysis ALLALL. From the three dimensional model, it was seen that L.R., S.B. and C.B. form a rather tight group (Figure 4 [A], [B], [C]). When only three dimensions are considered, it can be seen, from the M.S.T., that L.R. is closer to N.A. (70) than are either of the other two samples to any other sample in the analysis. Secondly, the connection between the salt water and freshwater samples is shortest between G.B. and L.E. The L.H. sample was the nearest to the two stunted samples, L.H.S. and G.L.

---

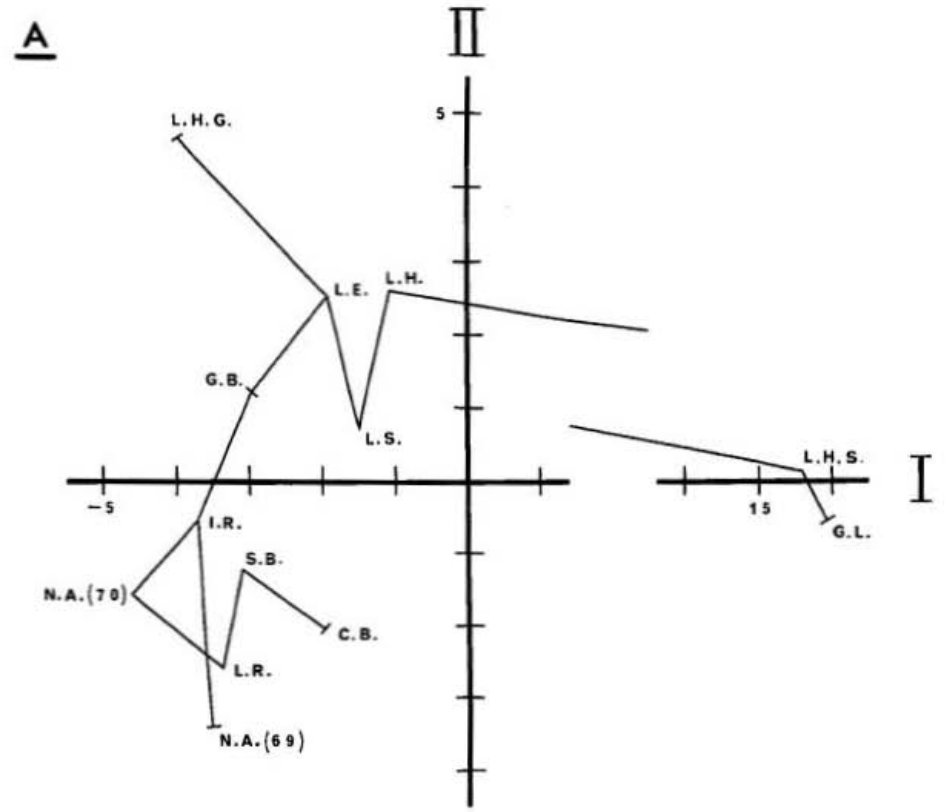
[1]Gower and Ross, 1969.

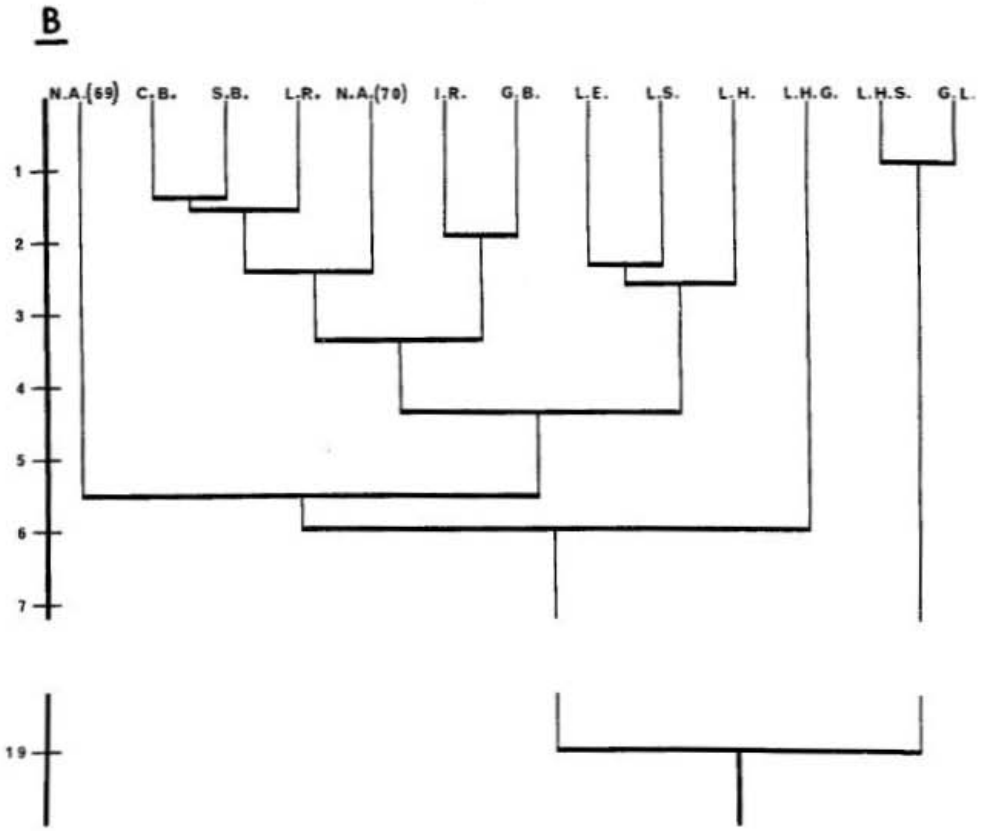
FIGURE 5(A)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II AND M.S.T. COMPUTED FROM INTERSAMPLE DISTANCES DERIVED FROM THE FIRST THREE CANONICAL AXES IN THE ANALYSIS ALLALL

FIGURE 5(B)

DENDROGRAM OF SMELT SAMPLES FROM U.P.G.M. CLUSTER ANALYSIS ON INTERSAMPLE DISTANCE MATRIX DERIVED FROM THE FIRST THREE CANONICAL AXES IN THE ANALYSIS ALLALL


A
II
I
5
−5
15
L.H.G.
L.E.
L.H.
G.B.
L.S.
L.H.S.
G.L.
I.R.
S.B.
N.A.(70)
C.B.
L.R.
N.A.(69)


B
N.A.(69)
C.B.
S.B.
L.R.
N.A.(70)
I.R.
G.B.
L.E.
L.S.
L.H.
L.H.G.
L.H.S.
G.L.
1
2
3
4
5
6
7
19

Although in two dimensions it may appear that L.E. and L.H. are close together, each of their distances from L.S. is shorter than is the distance between themselves.

These relationships are of value when setting up the dendrogram from the U.P.G.M. analysis. The U.P.G.M.A. produced a series of clusters and the level at which these clusters form. However, there is no information available regarding the arrangement of the samples at the ends of the branches. As the first cluster to form, L.H.S.-G.L., (Figure 5 [B]) is far removed from all others, it was placed on the extreme right-hand side. From the M.S.T. it was seen that the connection of this cluster with the others occurs between L.H. and L.H.S. Therefore, L.H.S. was oriented towards the body of the samples in the dendrogram. Similarly, among the Great Lakes samples, L.E. was oriented towards G.B. (and thus the body of the salt water samples) and L.H. is found on the side closest to the two stunted samples. For the remainder of the salt water samples, the clusters were arranged in the following manner. From the M.S.T., the connections indicated that the linkage is N.A. (70) - L.R. - S.B. - C.B. The first cluster from the U.P.G.M.A. was S.B. - C.B. To this cluster, L.R. was added. As L.R. was closer to S.B. than to C.B., it was placed on the S.B. side of the first salt water cluster. N.A. (70) was the next sample to be added. It was placed on the L.R. side rather than on the C.B. side, because of the information obtained from the M.S.T. The next

question was whether this four sample cluster should be rotated 180° so that N.A. (70) would have been beside N.A. (69) and C.B. would be beside I.R. The C.B. - N.A. (69) distance is quite large, 5.28, (Table 10) and might suggest that this is not the best arrangement. However, the dendogram as shown, has the shortest sum of distances; i.e. ((N.A. (70) to I.R.) + (C.B. to N.A. (69)) < ((N.A. (69) to N.A. (70) + C.B. to I.R.). This was also the arrangement suggested by the M.S.T. At this stage in the clustering procedure, there were three clusters and two unclustered populations. The clusters were, the salt water group, the freshwater group and the stunted group. The N.A. (69) and L.H.G. samples were atypical in so far as they did not join their salt water or freshwater groups respectively, before these two major groups had joined. However, as they lie on the extreme opposite ends of axis II, along which the freshwater and salt water fish are divided, they are considered the exaggerated expression of their respective groups. Therefore, they were placed on the extreme opposite sides of the central salt water - freshwater cluster.

(3) All Canonical Axes (ALLALL)

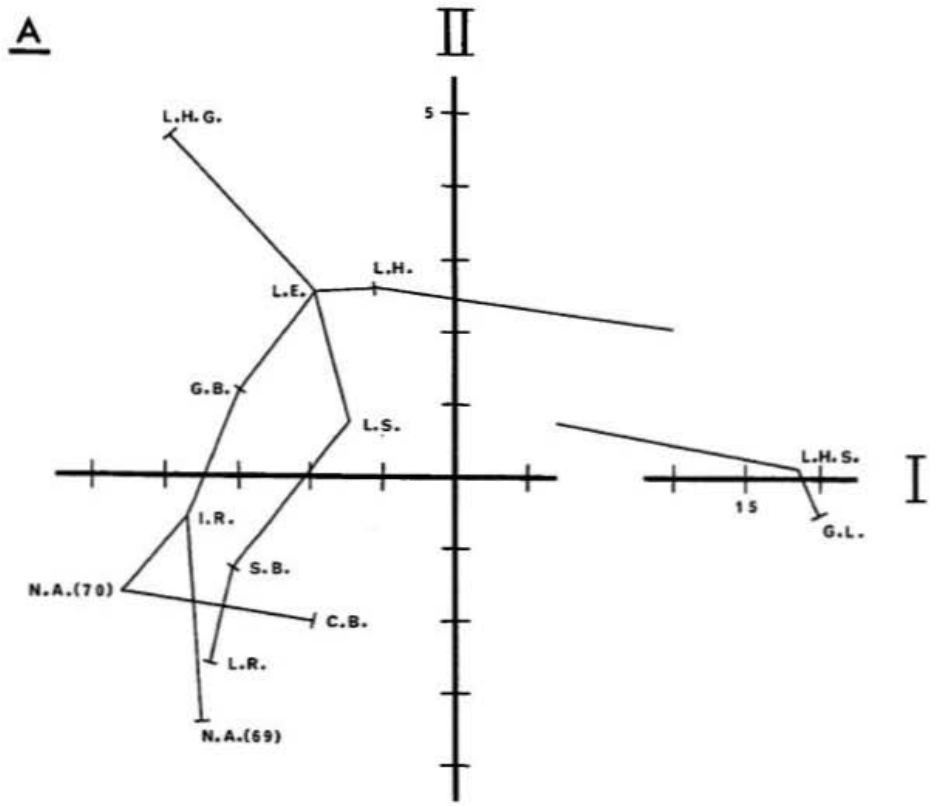
When all canonical axes from ALLALL were considered the M.S.T. of Figure 6(A) and the U.P.G.M. dendrogram of figure 6(B) resulted. With the availability of the additional information, some interesting changes occurred in the M.S.T. Firstly, the connection between the L.H. sample and the

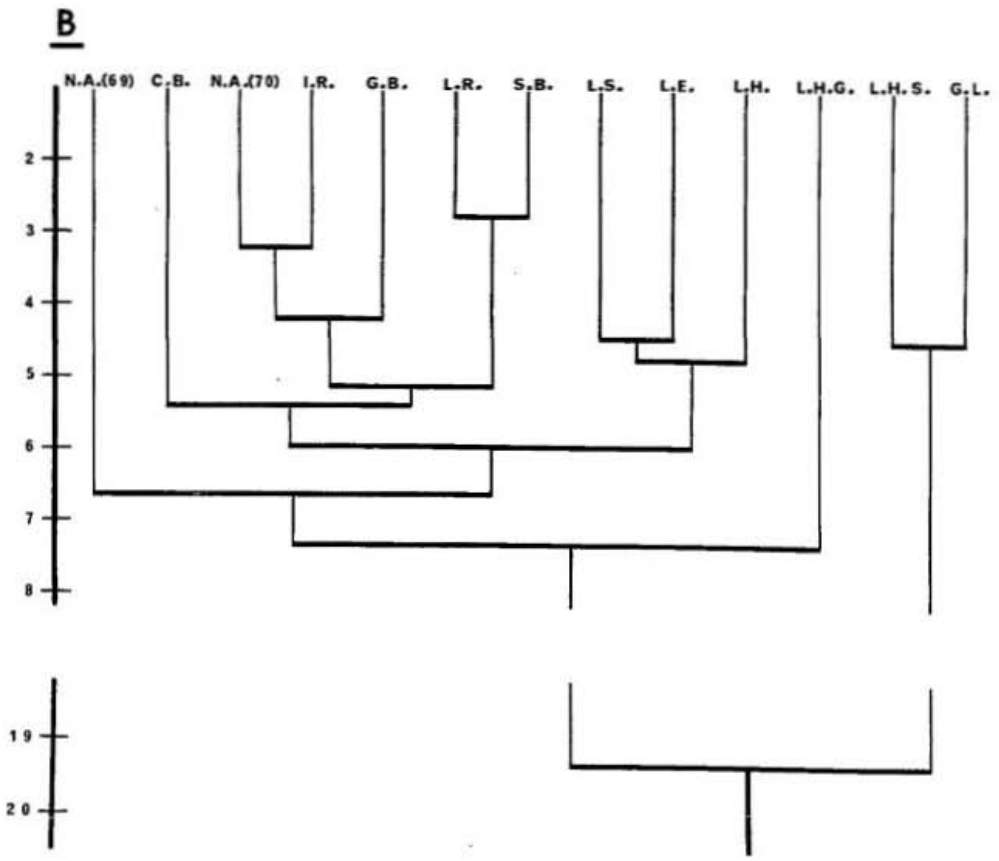
FIGURE 6(A)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II AND M.S.T. COMPUTED FROM THE INTERSAMPLE DISTANCES DERIVED FROM ALL CANONICAL AXES IN THE ANALYSIS ALLALL

FIGURE 6(B)

DENDROGRAM OF SMELT SAMPLES FROM U.P.G.M. CLUSTER ANALYSIS ON INTERSAMPLE DISTANCE MATRIX DERIVED FROM ALL CANONICAL AXES IN THE ANALYSIS ALLALL


A
II
I
5
15
L.H.G.
L.H.
L.E.
G.B.
L.S.
L.H.S.
G.L.
I.R.
S.B.
N.A.(70)
C.B.
L.R.
N.A.(69)


B
N.A.(69)
C.B.
N.A.(70)
I.R.
G.B.
L.R.
S.B.
L.S.
L.E.
L.H.
L.H.G.
L.H.S.
G.L.
2
3
4
5
6
7
8
19
20

remainder of the non-stunted samples is with L.E. instead of L.S. Secondly, the tight S.B. - I.R. - C.B. cluster seen in the three dimensional model has become somewhat dispersed. Thus, C.B. is now closer to N.A. (70) than to the other two members of its former cluster. Also, the connection of the L.R. - S.B. cluster is now through S.B. to L.S. This suggests that the two elongated clusters seen in Figure 4C, have become even more widely separated with the exception of C.B. and L.E. which are probably intermediate between the two chains. The M.S.T. shows that the G.B. - L.E. link is shorter than either of the distances between C.B. and L.R. or S.B.

These changes in the M.S.T. are reflected in the arrangement of the branches of the dendrogram seen in Figure 6(B). The reversal of the positions of L.E. and L.S. result because of the L.S. - S.B. and the L.E. - L.H. links. This situation is best satisfied by the arrangement shown. Similarly for the N.A. (70) - I.R. cluster, N.A. (70) is linked with C.B. while I.R. is linked with the G.B. sample. The N.A. (70) - I.R. - G.B. cluster is oriented as shown because of the link of G.B. with L.E., a freshwater sample. C.B. therefore, lies between N.A. (69) and N.A. (70). Once again, the clustering shows that N.A. (69) and L.H.G. remain as the two exaggerated expressions of their respective types and are thus again placed on the appropriate edges of the salt water - freshwater cluster.

When all the meristic and morphometric information is considered, (all canonical axes), the three main clusters that were seen in Figure 5(B), (three canonical axes) are again apparent. There are two major differences between the two dendrograms. Firstly, there is a certain amount of rearrangement of the subclusters in the main salt water cluster. This is the result of the separation of C.B. from L.R. - S.B. and C.B.'s subsequent association with N.A. (70). The G.B. sample also becomes separated from I.R. which then forms a primary cluster with N.A. (70). Secondly, with the consideration of the fourth and subsequent axes, the lengths of most of the stems increase by between one and three units. This is especially noticeable with the L.H.S. - G.L. cluster which forms at 0.87 (Table 10) in the first case, but not until 4.50 (Table 11) in the second. On the other hand, there is little change in the distance at which the stunted cluster joins with the remainder of the samples. Thus, the majority of the information separating these two main groups is contained in the first three axes. The relative level at which these two groups join should also be noted. Therefore, the difference between the stunted cluster and the main body of the samples is almost three times larger than any of the difference between any samples in the main body.

(I) Esterases*

(1) The General Pattern and Sample Comparisons

The gels stained for esterase activity were examined and the following pattern was observed. There appeared to be a possibility of a maximum of eight bands being present in any individual. Figure 7(B) shows an electropherogram which resulted from running liver extracts from five individuals from the L.H. sample. From the electropherogram, the eight band composite pattern can be seen. This figure also shows the type of variation which was seen between individuals of a sample. Electropherograms of this type were the results from which the presence or absence of bands were recorded.

The question as to whether or not the bands from one sample were the same as those from another was answered in the following way. A number of individuals from the various samples were run side by side on the same gel. These experiments showed that the bands from one sample migrated at the same rate as did those of any other sample. Examples of the results from these types of experiments are

---

*In Figure 7(A) the "constant" pattern of the muscle myogens between samples is seen. Experimental variation in bands 3, 4 and 5 is seen when one compares individuals 1-3 with 4 and 5. This type of variation was not seen if individuals from different samples were compared in the same run.

FIGURE 7

PHOTOGRAPHS OF ELECTROPHEROGRAMS OF SOME

SMELT BIOCHEMICAL SYSTEMS

A

Myogen Patterns Between Samples

1 Three individuals
2 From the Lake
3 Erie (L.E.) Sample

4 Two individuals
5 From the Grand Bank (G.B.) Sample

B

Esterase Patterns Within a Sample (L.H.)

1 Four individuals
2 From the Lake
3 Huron (L.H.)
4 Sample

C

Esterae Patterns Between Freshwater Samples

1 A Green Lake (G.L.) individual
2 A Lake Superior (L.S.) individual
3 A Lake Erie (L.E.) individual
4 A Lake Huron (L.H.) individual
5 A Green Lake (G.L.) individual

D

Esterase Patterns Between a Freshwater and Salt Water Sample

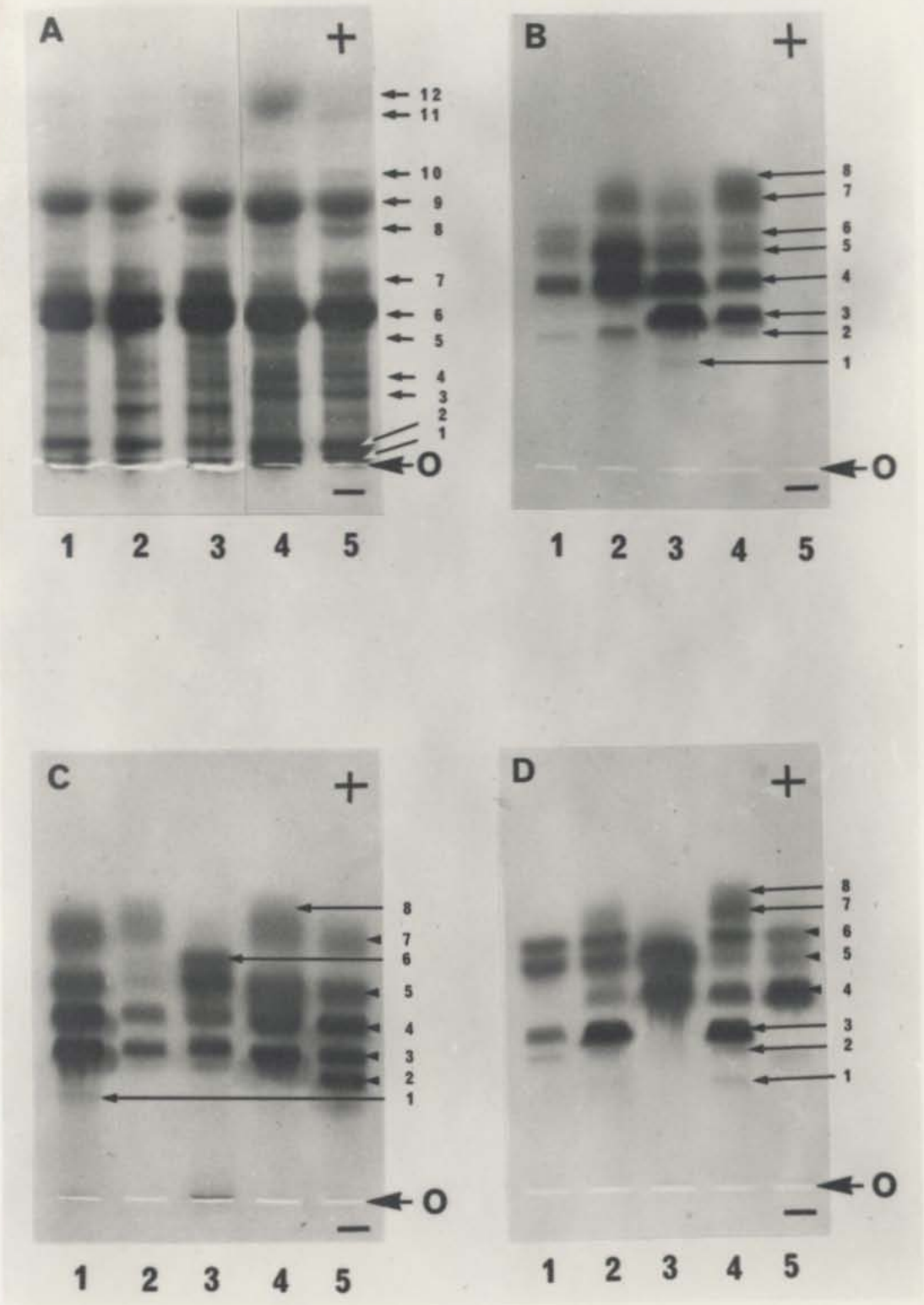
1 Three individuals
3 From the Lake
5 Erie (L.E.) Sample

2 Two individuals
4 From the Norris Arm 1970 (N.A. (70)) Sample

+ = the anodic end of the gel

- = the cathodic end of the gel

0 = the origin


A
+
12
11
10
9
8
7
6
5
4
3
2
1
O
−
1 2 3 4 5
B
+
8
7
6
5
4
3
2
1
O
−
1 2 3 4 5
C
+
8
7
6
5
4
3
2
1
O
−
1 2 3 4 5
D
+
8
7
6
5
4
3
2
1
O
−
1 2 3 4 5

seen in Figure 7(C) and 7(D). Representatives of G.L. and the Great Lakes are seen in Figure 7(C). This is a small example of the variation which exists between samples. Figure 7(D) is a comparison of individuals from N.A. (70) and the L.E. sample. This figure not only shows variation within the two samples, but also offers a comparison between a salt water and freshwater sample. If a band from one sample migrated at the same rate as that of any other sample, it was recorded as the same band for the purpose of the cross comparison of the samples. An investigation of the patterns seen in Figures 7(B), 7(C) and 7(D) indicates that the bands obtained from individuals from the various samples, in fact, appear to be the same.

(2) The Frequency of Esterase Bands

The frequencies of each band in each sample were computed. These results are seen in Table 12. From this table, some interesting points emerge. The frequency of occurrence for band 1 in all the salt water samples ranges between .801 and .979. The Great Lake samples have a very low frequency for this band, between .083 and .281. The Green Lake sample shows a more median frequency of .526. Therefore, there appears to be a distinction between freshwater and salt water smelts with respect to band 1. Band 3 exhibits a similar situation. It was almost always present in the salt water samples and could be used as a marker

TABLE 12

FREQUENCY OF EACH ESTERASE BAND IN EACH GROUP OF SMELT AND THE CHI-SQUARE VALUE FOR EACH BAND

Group	Esterase Bands								N
	1	2	3	4	5	6	7	8	
N.A.(69)	.857	.995	1.000	.862	.872	.601	.606	.281	203
C.B.	.979	1.000	.990	.948	.958	.866	.866	.412	97
L.R.	.917	.969	.979	.937	.812	.885	.771	.333	96
S.B.	.821	1.000	.981	.981	.632	.972	.981	.594	106
I.R.	.801	1.000	.988	.562	1.000	.530	.645	.135	96
N.A.(70)	.926	.926	1.000	.672	1.000	.728	.884	.114	96
G.B.	.926	1.000	1.000	.957	1.000	.676	.926	.458	96
G.L.	.526	.957	.655	1.000	1.000	.177	.998	.655	96
L.S.	.125	.946	.374	1.000	1.000	.499	.936	.426	96
L.H.	.281	.780	.478	.988	.988	.260	.936	.343	96
L.E.	.083	.988	.551	.915	1.000	.718	.144	.021	96
χ^2 *	172.8	4.3	75.1	22.4	15.8	97.7	87.3	113.1	
Significance Level	<.001	N.S.	<.001	<.025	N.S.	<.001	<.001	<.001	

N = the number of individuals in each group

*degrees of freedom = 10

N.S. = Not Significant

point with which to compare gels of various runs; however, it was present at considerably lower frequencies among the freshwater samples. Band 4 was generally present at a high frequency in both salt and freshwater samples. However, N.A. (69), N.A. (70) and I.R. (the Notre Dame Bay samples) exhibited a somewhat lower frequency of occurrence. Band 6 occurred much more frequently in the Lake Erie sample than it did in the other freshwater samples. Green Lake had the lowest frequency of occurrence for Band 6 followed by Lake Huron. Band 7 occurred very infrequently in the Lake Erie sample which was quite distinct from the remaining samples with regard to this band. Band 8 again showed the frequency of the Lake Erie sample to be considerably lower than the rest of the freshwater groups. Again, the Notre Dame Bay samples appeared to have Band 8 present at a lower frequency than did the other salt water samples.

To test the significance in the observed differences between samples, in the frequency of occurrence of each band, Chi-Square tests were performed by solving

$$\chi^2_m = \sum_{i=1}^{n} \frac{(\text{obs}(i) - \text{exp}(i))^2}{\text{exp}(i)} \qquad (5)$$

where m = the number of the band

n = the number of groups

obs (i) = the number of times band m was observed in group i

$$\text{exp}(i) = N_i \times \left(\sum_{i=1}^{n} \text{obs}(i) \Big/ \sum_{i=1}^{n} N_i \right)$$

The results of these tests are also seen in Table 12. These results show that only Bands 2 and 5 are not significant (N.S.), whereas the other bands are all rather highly significant. Therefore, it was apparent that the esterase system in the smelt was not uniform and that significant differences did exist between the various samples.

(3) Joint Occurrence Matrices

Although variation did occur between samples with respect to the frequency of occurrence of the esterase bands, it was felt that this method of analysis did not specifically capture and show the variation of the pattern types and band association between samples.

An exhaustive treatment would have entailed classifying the patterns obtained into the 256 (2^8) possible combinations, of the eight bands being either present or absent. This approach would have been slow and costly and would not necessarily have provided more informative results. The approach taken was to classify the bands in pairs, for each sample. This results in 28, $\binom{8}{2}$, pairs of bands. These twenty eight pairs give some measure as to how the individual bands are associated with the other bands and thus give a measure of the variability of the patterns within a sample. Therefore, for each sample, twenty eight frequencies were recorded and arranged in a matrix whose elements are the frequency with which band i and band j occur in the same

individual in that sample. Thus by this method, it should be possible to determine whether any two bands are associated at a higher level than would have been expected on the basis of their joint probability.

(4) Chi-Square Tests

In the analysis of the data from the esterase patterns, the following questions were considered. Is it possible to distinguish various samples or groups of smelts on the basis of the esterase patterns, or are all the groups in effect homogeneous? If the individual samples are taken in pairs and their band frequencies compared, are there any significant differences between samples and if so, what are the levels of the significance? To answer these questions, it is necessary to postulate that the occurrence of each band is independent of all other bands present in that organism. Thus it was necessary to test to see if this assumption could legitimately be made.

The Chi-Square test for the hypothesis that all groups are homogeneous and the test for the homogeneity of all pairs of groups will be based on the assumption that the bands occur independently of each other in each fish. Therefore, it was necessary to test this assumption to see if it was valid and therefore to test the validity of the two subsequent tests. To do this, a Chi-Square test was set up in the following manner. Previously, matrices of joint

frequency of occurrence had been generated for each population. The elements of a matrix were the frequencies with which band i and band j occurred together in that population. Therefore, the question was do the bands occur together in an independent fashion or do the bands occur in pairs in a non independent fashion? The latter possibility would result in some cells of the matrix having higher or lower values than might be expected on the basis of the joint probability, calculated from the occurrence of each band alone. When the null hypothesis is that the bands occur independently in population A, where A can represent any population,

$$\chi^2_A = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} N_A \left[\frac{p(i \cap j) - (p_i \times p_j)}{p_i \times p_j} \right] \qquad (6)$$

where N_A = the number of individuals in population A

$(p_i \times p_j)$ = the frequency with which i occurs × the frequency with which band j occurs in population A.

$p(i \cap j)$ = the frequency with which i and j occur together in population A

In this test, the number of degrees of freedom is $[m \times (m-1)/2]-m$, 20 in the case of an eight band pattern. The results of this test are seen in Table 13. From these results it can be seen that none of the Chi-Square values are significant. Thus, the null hypotheses, that the esterase bands occur in an independent fashion, cannot be rejected. Therefore, it appears that the overall Chi-Square

TABLE 13

INTRAGROUP CHI-SQUARE VALUES FROM ESTERASE DATA AND LEVELS OF SIGNIFICANCE, FOR EACH SAMPLE OF SMELT

Group	Intragroup Chi-Square	Significance Level*
N.A.(69)	26.93	> .100
C.B.	1.75	> .995
L.R.	4.32	> .995
S.B.	2.40	> .995
I.R.	15.33	> .750
N.A.(70)	6.15	> .995
G.B.	0.51	> .995
G.L.	15.43	> .750
L.S.	13.44	> .750
L.H.	30.52	> .050
L.E.	19.75	> .250

*degrees of freedom = 20

to test the homogeneity of all groups and the Chi-Square test for the homogeneity of each pair of populations does not rest on a faulty assumption--the independence of the occurrence of the various esterase bands.

The second test evaluated the hypothesis that all groups were homogeneous with respect to their total esterase. Thus the null hypothesis was that all samples were homogeneous with respect to the occurrence of the bands. In this case Chi-Square was defined as:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(\text{obs}(i,j) - \text{exp}(i,j))^2}{\text{exp}(i,j)} \qquad (7)$$

where obs (i,j) = the number of observations of band j in sample i

exp (i,j) = the expected number of observations of band j in sample i

= (Number of individuals in sample i) × (Grand total of the number of occurrences of band j) / (total number of individuals in the study.)

n = the number of groups

m = the number of the band

For this test the degrees of freedom are $[(n \times m)-m]$. The results of this test on all eleven samples for which the esterase electrophoresis was done resulted in a Chi-Square of 588.53 with 80 degrees of freedom. When the degrees of freedom for a Chi-Square test are greater than thirty, a normal approximation is quite accurate (C.R.C. Handbook of Tables for Probability and Statistics). The expression

$(2\chi^2)^{\frac{1}{2}} - (2n-1)^{\frac{1}{2}}$ is approximately normally distributed as the standard normal distribution. Thus, χ^2_α the α point of the χ^2 distribution, may be computed by the formula

$$\chi^2_\alpha = \tfrac{1}{2}[x_\alpha + (2n-1)^{\frac{1}{2}}]^2 \qquad (8)$$

where x_α is the α point of the cumulative normal distribution. Thus for an α point of .9999, and 80 degrees of freedom,

$$\chi^2_{.0001} = \tfrac{1}{2}[3.62 + (2 \times 80 - 1)^{\frac{1}{2}}]^2$$
$$= 131.5$$

Since the Chi-Square obtained through the analysis was 588.53 which is greater than 131.5, the results are highly significant ($p << .0001$) and the null hypothesis must be rejected. Thus the samples are not homogeneous with respect to their esterase patterns.

The final question was concerned with the homogeniety between all sample pairs. It has just been seen that all the samples are not homogeneous with respect to their esterase patterns. Therefore, is the lack of homogeneity exhibited, the result of an overall situation, or are certain of the samples especially distinct from others? To test this question, a third Chi-Square test was done. Here, the null hypothesis was that each pair of samples was homogeneous with respect to the occurrence of the esterases. In this case the Chi-Square value for each pair of samples was found by evaluating

$$\chi^2_{ab} = \left[\sum_{k=1}^{m} \frac{[\text{obs}\ (a,k) - \text{exp}\ (a,k)]^2}{\text{exp}\ (a,k)}\right] + \left[\sum_{k=1}^{m} \frac{[\text{obs}\ (b,k) - \text{exp}\ (b,k)]^2}{\text{exp}\ (b,k)}\right] \quad (9)$$

where obs (a,k) = the number of observations of band k in sample a

exp (a,k) = the expected number of observations of band k in sample a

$$= (\text{number of individuals in sample a}) \times \left[\frac{(\text{number of occurrences of k in a}) + (\text{number of occurrences of k in b})}{(\text{number of individuals in a}) + (\text{number of individuals in b})}\right]$$

obs (b,k) and exp (b,k) were similarly defined. The results of this test are seen in Table 14. In this test, the degrees of freedom equal m, eight in this case. The levels of significance of the between populations Chi-Square values are seen in Table 15. From this table, it can be seen that the salt water populations as a group are all highly significantly different from the freshwater populations. On the other hand, within the salt water populations, there appears to be a lower significance level in the cross comparisons. Similarly, Lake Huron shows less heterogeneity with Green Lake or Lake Superior than do any of the salt water - freshwater comparisons. Lake Erie on the other hand shows a high level of heterogeneity with all populations, both from salt and freshwater.

If one concentrates on the magnitude of the Chi-Square values rather than their significance level a further

TABLE 14

CHI-SQUARE MATRIX TO TEST THE HOMOGENEITY
BETWEEN SMELT SAMPLES, BASED ON
THE ESTERASE DATA

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.
C.B.	18.86									
L.R.	11.93	2.78								
S.B.	50.42	12.98	13.19							
I.R.	15.65	35.58	29.22	67.67						
N.A.(70)	21.52	22.20	18.47	50.57	8.90					
G.B.	18.37	2.89	7.90	16.32	34.14	24.98				
L.S.	106.18	100.28	98.67	102.90	105.70	113.51	88.86			
L.H.	85.06	89.21	86.85	100.77	77.68	90.02	72.73	16.65		
L.E.	139.62	175.34	154.83	200.57	122.47	146.98	179.44	105.41	123.28	
G.L.	82.37	70.24	76.70	75.52	80.89	93.19	49.62	50.23	22.29	186.92

TABLE 15

LEVELS OF SIGNIFICANCE FROM BETWEEN SMELT SAMPLES

CHI-SQUARE MATRIX (TABLE 14)

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.
C.B.	+									
L.R.	N S	N S								
S.B.	+ + +	N S	N S							
I.R.	+	+ + +	+ + +	+ + +						
N.A.(70)	+ +	+ +	+	+ + +	N S					
G.B.	+	N S	N S	+	+ + +	+ +				
L.S.	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +			
L.H.	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+		
L.E.	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	
G.L.	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +	+ +	+ + +

N S = p > .100

\+ = .100 > p > .010

\+ + = .010 > p > .001

\+ + + = p < .001

quantification of the relationships is possible. Table 16 presents the within group and between groups average values for the Chi-Squares of all salt water populations, all freshwater populations and freshwater populations excluding Lake Erie. From this table it can be seen that the average Chi-Square between salt water and freshwater populations excluding Lake Erie is approximately three times the average Chi-Square within each group. Also, the inclusion of the Lake Erie population has the result of considerably boosting both the average within freshwater Chi-Square and the average between groups Chi-Square. Although Lake Erie shows a high Chi-Square with both salt and freshwater populations, (Table 14) it shows a higher average Chi-Square with the salt water populations (159) than it does with the freshwater ones (138).

(J) Classification of Samples on Esterase Data

When the stepwise discriminant analysis was done on the esterase data, the following results were obtained. These results were generally similar to those obtained from the investigations of the frequency of occurrence tables and Chi-Square tests previously discussed.

(1) One Band Considered

The different variables are included in the discriminant functions on the basis of the decreasing F-value of the between groups over within group variance ratio. On the basis of this criterion, the first band to be considered was Band 1. As the esterase data is binary, it is

TABLE 16

AVERAGE WITHIN AND BETWEEN GROUPS CHI-SQUARE FOR SELECTED GROUPS OF SMELT (COMPILED FROM TABLE 14)

Group	Average Within Group Chi Square	Average Between Group Chi Square
All Salt Water Populations	23.1	105.9
All Freshwater Populations	84.1	105.9
All Freshwater Populations Excluding Lake Erie	29.7	88.0

TABLE 17

CLASSIFICATION MATRICES OF SMELT SAMPLES

BASED ON ESTERASE DATA

(A) One Band Considered

(B) Two Bands Considered

Band	1									1
Present/Absent	P									A
Sample | Group	N.A. (69)	C.B.	L.R.	S.B.	I.R.	N.A. (70)	G.B.	G.L.	L.S.	L.H.	L.E.
N.A.(69)		86									14
C.B.		98									2
L.R.		92									8
S.B.		82									18
I.R.		80									20
N.A.(70)		93									7
G.B.		93									7
G.L.		52									48
L.S.		12									88
L.H.		28									72
L.E.		8									92

Band	1,7						1,7		1,7		1,7
Present/Absent	P A						P P		A P		A A
Sample | Group	N.A. (69)	C.B.	L.R.	S.B.	I.R.	N.A. (70)	G.B.	G.L.	L.S.	L.H.	L.E.
N.A.(69)	34						52		9		5
C.B.	12						86		1		1
L.R.	19						73		4		4
S.B.	0						82		16		2
I.R.	22						58		6		14
N.A.(70)	10						83		6		1
G.B.	7						86		7		0
G.L.	0						52		47		1
L.S.	1						11		83		5
L.H.	0						28		66		6
L.E.	7						1		10		82

only possible to divide the samples into two groups on the basis of one character, that is, those with Band 1 and those without. Furthermore, as there are only two groups possible, these groups will be the two populations with Band 1 present at the highest and lowest frequencies, respectively. These are, in fact, the results that were obtained, as seen in Table 17(A). All individuals with Band 1 were classified into the Chuff Brook group, (C.B.); whereas, all those individuals who lacked Band 1 were classified into the Lake Erie group, (L.E.). On the basis of this single character, 89% of the salt water fish were correctly identified as such and 84% of the Great Lakes fish were correctly identified as coming from a Great Lakes population. The sample from Green Lake was intermediate, as 52% were classified as salt water and 48% were classified as coming from Lake Erie.

(2) Two Bands Considered

The second band to be entered into the discriminant function was Band 7, which had the second highest F-value after Band 1 had been removed. These results are seen in Table 17(B). On the basis of the two bands, four groups are possible. These are: both present, both absent, Band 1 present with Band 7 absent and Band 1 absent with Band 7 present. From an investigation of Table 12, it may have been possible to deduce the identity of the four groups, but it would have been much more difficult. Thus, N.A. (69) is the group with Band 1 present and Band 7 absent,

G.B. becomes the group for both Bands 1 and 7 present, L.S. for Band 1 absent and Band 7 present and L.E. for both bands absent. The percentages of correct classification into salt water and Great Lakes remains unchanged as the two major groups have already been split, on the basis of Band 1.

(3) Three Bands Considered

The third band to be included in the discriminant function is Band 6. The primary distinction between the salt water and Great Lakes populations remains the same with regard to the percentage of each group correctly classified. However, two new groups were added. From an examination of Tables 17(B) and 18(A), it appears that the numbers classified into N.A. (69) and L.E. remain the same. The G.B. group of Table 17(B) has apparently given rise to two new groups. C.B. and G.L. in Table 18(A). An investigation into the values of Table 12 and the values of Table 18(A), suggest that the C.B. group contains those individuals with Bands 1, 7 and 6 present; whereas, the G.L. group contains those individuals with Bands 1 and 7 present and Band 6 absent. The L.S. group of Table 17(B) has been split to form the groups L.S. and L.H. of Table 18(A). Once again, from an investigation into the frequencies of Tables 12 and 18(A), indicates that the L.S. group contains those individuals with Bands 1 and 6 absent, and Band 7 present. The L.H. group appears to contain those members with Band 1 absent and Bands 7 and 6 present.

TABLE 18

CLASSIFICATION MATRICES OF SMELT SAMPLES

BASED ON ESTERASE DATA

(A) Three Bands Considered

(B) All Bands Considered

Band	1,7,6	1,7,6						1,7,6	1,7,6	1,7,6	1,7,6
Present/Absent	P A -	P P P						P P A	A P A	A P P	A A -
Sample | Group	N.A. (69)	C.B.	L.R.	S.B.	I.R.	N.A. (70)	G.B.	G.L.	L.S.	L.H.	L.E.
N.A.(69)	34	35						17	7	2	5
C.B.	12	75						11	1	0	1
L.R.	19	68						5	4	0	4
S.B.	0	81						1	15	1	2
I.R.	22	32						26	5	1	14
N.A.(70)	10	63						20	6	0	1
G.B.	7	61						25	3	4	0
G.L.	0	8						45	9	37	1
L.S.	1	4						7	42	41	5
L.H.	0	8						20	17	49	6
L.E.	7	1						0	8	2	82

Sample | Group	N.A. (69)	C.B.	L.R.	S.B.	I.R.	N.A. (70)	G.B.	G.L.	L.S.	L.H.	L.E.
N.A.(69)	27	26		13	14	0	8	6	3	0	3
C.B.	9	72		4	4	1	8	1	0	0	1
L.R.	14	51		19	5	1	2	1	1	3	3
S.B.	0	51		37	1	1	1	0	7	0	2
I.R.	7	28		0	41	4	7	5	5	0	3
N.A.(70)	1	52		0	18	14	7	0	4	3	1
G.B.	6	60		0	3	1	8	19	3	0	0
G.L.	0	7		0	0	0	3	48	37	4	1
L.S.	1	3		0	0	0	1	10	75	5	5
L.H.	0	6		0	1	0	6	16	47	22	2
L.E.	7	1		0	7	0	0	2	8	1	74

(4) All Bands Considered

With the addition of the fourth and subsequent bands to the discriminant functions, interpretation of the classification matrix became much more difficult. This was undoubtedly because of the increased complexity of the interactions of the weights assigned to each band by means of the coefficients of the discriminant functions for each population. The classification matrix resulting after the inclusion of all 8 of the esterase bands is seen in Table 18(B). The number of animals correctly classified from each group is found in the diagonal elements of the matrix, (Table 18[B]) and represents 37% for all samples and 89% for the salt-freshwater partition.

(5) Classification of Each Phenotype

The interpretation of the phenotype of the members of each cell of the matrix was possible by inspection and deduction for the first two or three bands entered (Tables 17[A], [B] and Table 18[A]). However, the increasing interaction of the coefficients prevented the use of this method for the inclusion of more than three bands. Table 18(B) is the result of the classification of a possible 256, 8 band phenotypes, into one of the 11 available groups. To determine into which group each phenotype would be classified, the following procedure was adopted.

The discriminant analysis yields at each step, for

each population, a set of coefficients and a constant. The number of coefficients is equal to the number of characters being considered at that time. Thus, when all 8 bands are considered, each population has a function in the form

$$y_i = c_{1i}x_1 + c_{2i}x_2 + \ldots c_{8i}x_8 + k_i \qquad (10)$$

where y_i = the value of the discriminant for the phenotype when used with the coefficients and constant of population i

i = the number of the population

$x_1 \ldots x_8$ = the value of bands 1 through 8

Therefore, any particular phenotype would have as many values for y_i as there were groups into which the phenotype could be classified. A particular phenotype is classified into the group whose coefficients and constants, when used with the values of the phenotype produce the maximum value for y_i.

Therefore, each of the eleven discriminant functions was evaluated for each of the 256 possible phenotypes yielding 2,816 values for y. From these, the group whose coefficients and constant produced the maximum value of y_i, for each phenotype, was the group into which a fish with that phenotype would be classified. A cross comparison table was constructed in this manner. That is, for each possible phenotype, it was established into which group the phenotype would be classified by reference to the particular discriminant function which had yielded the maximum value

when evaluated with that phenotype. By the use of this reference table, it was possible to classify any of the 256 phenotypes into its' appropriate group.

In order to use this table, it was necessary to sort the individuals of each population with respect to their phenotypes and to record the number of occurrences of each phenotype in each population. When this was done, it was possible to determine for any population how many animals in each cell of the classification matrix were the result of each phenotype present in that population. For example, from Table 18B, it can be seen that for the C.B. sample, 9 individuals from this population were classified into the N.A. (69) group. From an examination of the number of each phenotype present in the C.B. population, and the group into which each phenotype was classified, it can be seen that there were nine fish with the 11111122 phenotype, which resulted in these nine fish being classified into the N.A. (69) group. Table 19 shows the remainder of the results for the C.B. sample. Although there were 72 animals from the C.B. sample, classified into the C.B. group, it can be seen that only two phenotypes, which occurred frequently in this sample, accounted for the 72 animals being correctly classified. These two phenotypes were very common in all the salt water samples. They were also the only two phenotypes that actually occurred, in any sample, which caused an animal to be classified into the C.B. group.

TABLE 19

THE NUMBER OF INDIVIDUAL SMELT FROM THE CHUFF BROOK SAMPLE (C.B.) CLASSIFIED INTO EACH GROUP (AREA) ON THE BASIS OF THE PHENOTYPES

Phenotype 1 2 3 4 5 6 7 8	Group	Area	Number
1 1 1 1 1 1 2 2	1	N.A.(69)	9
1 1 1 1 1 1 1 1	2	C.B.	39
1 1 1 1 1 1 1 2	2	C.B.	33
1 1 1 1 2 1 1 1	4	S.B.	2
1 1 1 1 2 1 2 2	4	S.B.	1
2 1 1 1 2 1 1 1	4	S.B.	1
1 1 1 2 1 2 1 2	5	I.R.	2
1 1 1 2 1 2 2 2	5	I.R.	2
1 1 1 2 1 1 1 2	6	N.A.(69)	1
1 1 1 1 1 2 1 2	7	G.B.	8
1 1 1 1 1 2 1 1	8	G.L.	1
2 1 2 1 1 1 2 2	11	L.E.	1

1 = band present

2 = band absent

No phenotypes were present in the C.B. sample that would have caused animals to be classified into the L.S. or L.H. groups. As with the rest of the populations, no animals from C.B. were classified into the L.R. group.

In the entire study, only 55 out of the possible 256 phenotypes were recorded as actually occurring. Of these, three would cause an animal to be classified into the N.A. (69) group, 2 caused classification into N.A. (70) and only 1 caused classification into the G.B. group. On the other hand, L.H. was the most diffuse group, as 13 phenotypes caused classification into this group. I.R., L.E. and L.S. were progressively less diffuse, with 10, 8 and 7 phenotypes causing classification into these groups, respectively. Five phenotypes caused inclusion in the S.B. group while 4 caused inclusion in G.L. None of the 8 phenotypes that would have caused an animal to be classified into the L.R. group actually occurred in this study, and as a result all of the L.R. group's cells are empty.

(K) Intersample Distances from Esterase Data

From the analysis ESTALL, a set of eight coordinates was obtained for each sample. By substituting these coordinates into Equation 3 the generalized distance matrix based on the esterase data was produced (Table 20). As there were no esterase data available from L.H.G. and L.H.S.

TABLE 20

DISTANCES ($\sqrt{D^2}$) BETWEEN SMELT SAMPLES' MEANS

DERIVED FROM ESTERASE DATA

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.
C.B.	1.11									
L.R.	0.85	0.78								
S.B.	1.56	1.43	1.02							
I.R.	1.25	1.56	1.61	2.20						
N.A.(70)	1.57	1.27	1.40	2.00	1.03					
G.B.	1.18	0.55	1.15	1.64	1.56	1.38				
L.S.	3.13	3.10	3.20	3.23	3.23	3.29	2.86			
L.H.	3.09	3.16	3.19	3.33	3.24	3.19	2.87	1.17		
L.E.	2.81	3.33	3.13	3.56	3.14	3.60	3.36	2.72	3.09	
G.L.	2.56	2.54	2.77	2.83	2.78	2.89	2.09	1.58	1.53	3.29

the resulting matrix contained fifty-five elements as compared to seventy-eight as seen when thirteen populations were considered. Also, it was noted that the average distance in this matrix was 2.31 while for ALLALL, it was 9.68. This reflects the relatively greater overall dispersion in ALLALL as compared with ESTALL.

(1) Between and Within Groups Distances

Table 21 shows the average distance between and within selected groups. The values in this table reflect the situation that was previously seen (Table 16). Again, the between group distances average more than twice the within group distances if the Lake Erie sample is disregarded. Therefore, it appears from an investigation of the esterase data, that there are two compact groups; the freshwater populations and the salt water populations. The Lake Erie sample is remote from the rest, but appears to be somewhat closer to the freshwater groups than to the salt water ones.

(2) Canonical Analysis

When the canonical analysis was run on the esterase data, a set of canonical coefficients was generated in the manner previously described (Section (B), Results). These were used in conjunction with the mean values for each esterase band to determine the coordinates for each sample in each of the eight dimensions. The positions of the sample means are plotted with respect to the first two

TABLE 21

AVERAGE WITHIN AND BETWEEN GROUPS DISTANCES BASED ON ESTERASE DATA FOR SELECTED GROUPS OF SMELT SAMPLES

Group	Average Within Group Distance	Average Between Group Distance
All Salt Water Populations	1.338	3.053
All Freshwater Populations	2.228	3.053
All Freshwater Populations Excluding Lake Erie	1.424	2.979

canonical axes (Figure 8[A]). From this figure, two things become apparent. Firstly, the relative dispersion of the samples about the first two axes is not nearly as great as has been seen with the meristic and morphometric data. (Figure 3), where the Lake Heney populations and a different number of variables had been included. Thus, it was necessary to double the scale of the plot in order to clarify the positions of the sample means. This situation was indicated by the sum of the latent roots for this analysis (Table 4) and the relatively poor results with the classification of the samples for this analysis (Table 8). These two factors are connected and indicate a considerable amount of overlap, especially among the salt water samples. This second point is seen in Figure 8(A). The seven salt water sample means are very tightly clustered on the left hand side of axis I. The freshwater samples are twice as far removed from the salt water samples as the salt water samples are dispersed amongst themselves. The G.L. sample is intermediate between the freshwater and salt water samples but is tending toward the freshwater group. The L.E. group is removed from all other samples. These results reflect the results obtained from the interpopulation Chi-Square tests of Table 14.

(3) Cluster Analysis

The M.S.T. derived from the esterase distance matrix (Table 20) is also seen in Figure 8(A). From this tree,

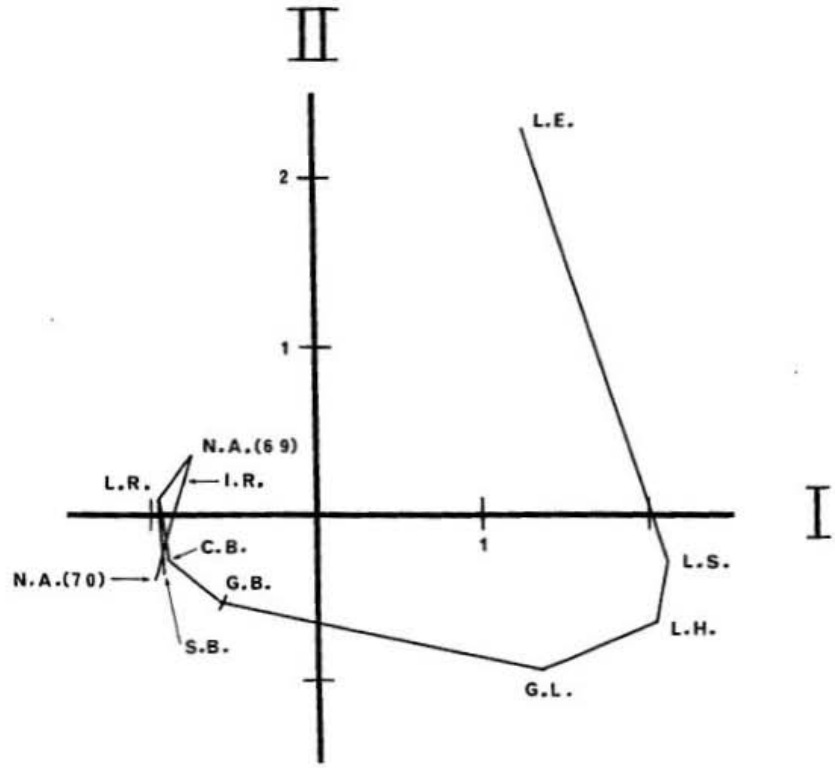
FIGURE 8(A)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II AND M.S.T. COMPUTED FROM THE INTERSAMPLE DISTANCES DERIVED FROM ALL CANONICAL AXES IN THE ANALYSIS ESTALL

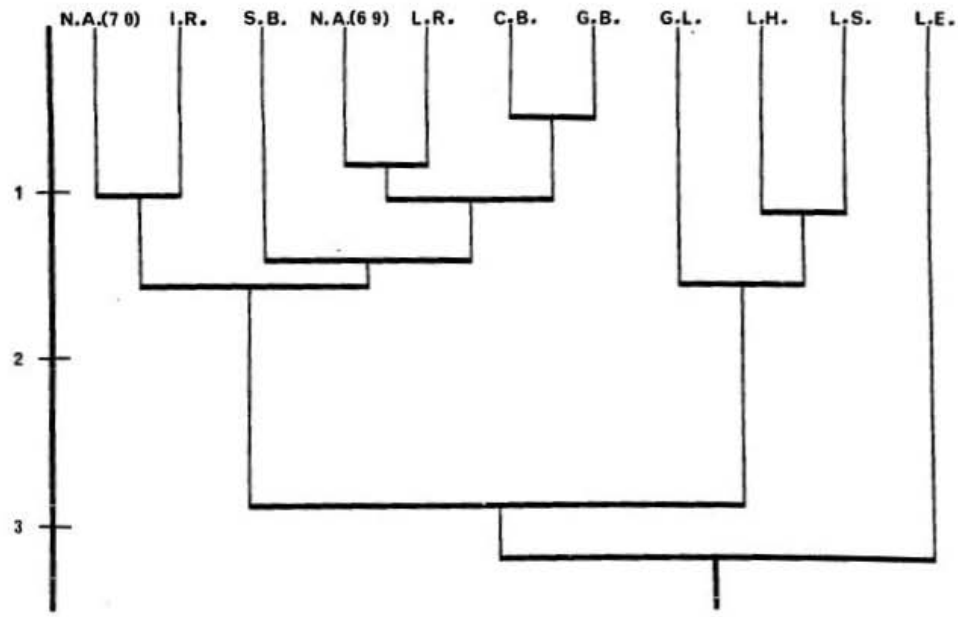
FIGURE 8(B)

DENDROGRAM OF SMELT SAMPLES FROM U.P.G.M. CLUSTER ANALYSIS ON INTERSAMPLE DISTANCE MATRIX DERIVED FROM ALL CANONICAL AXES IN THE ANALYSIS ESTALL

A


II
I
2
1
1
L.E.
N.A.(69)
L.R.
I.R.
C.B.
N.A.(70)
G.B.
S.B.
L.S.
L.H.
G.L.

B


N.A.(70)
I.R.
S.B.
N.A.(69)
L.R.
C.B.
G.B.
G.L.
L.H.
L.S.
L.E.
1
2
3

some of the relationships between the samples can be seen. The atypical L.E. sample is connected to the Great Lakes group through the L.S. sample. G.L. is intermediate to the salt water and freshwater groups and through it, G.B. is connected to L.H. The sample L.R. seems to be a central point for the salt water samples. The samples C.B., S.B. and N.A. (69) all are connected to the L.R. sample. N.A. (69) is one end of a chain that connects I.R. and N.A. (70). Although on the two dimensional representation, N.A. (70), S.B. and C.B. appear very close together, the M.S.T. shows that they are in fact each closer to L.R. than they are to each other. As there was no esterase data available for L.H.S. and L.H.G., they do not appear on these figures.

The relationships from the M.S.T. were again used with the results of the U.P.G.M. analysis to prepare the dendrogram seen in Figure 8(B). As L.E. was the atypical group, it was placed on the extreme right side. The G.L., L.H. and L.S. samples were then positioned as shown because of the L.S. - L.E. link and because G.L. forms the connection with the salt water samples through the G.B. sample. The M.S.T. indicates that L.R. should be positioned next to C.B. and that I.R. should be close to N.A. (69). From an examination of the minimum distances, calculated by the method previously used, S.B. should have been placed between L.R. and the C.B. sample. However, this was impossible because of the clusters that had formed. However, the

position that it takes in the dendrogram indicates the minimum distance of the two positions available.

In the esterase dendrogram, there are two main clusters and the L.E. sample. These main clusters contain the salt water and freshwater samples respectively. The increase of the distance threshold required to combine these two groups, was almost twice that required to form each separate group. This indicates that the differences between the salt water and freshwater groups were about twice those found in each group individually. The U.P.G.M. analysis also shows that although N.A. (70), S.B. and C.B. are very close in two dimensions, they do not all become part of the single cluster until the last step in the clustering of the salt water samples. Therefore, the differences between them are the result of their relative positions along the third and subsequent axes.

It is also interesting to note some of the first clusters to form. For example, the I.R. - N.A. (70) cluster is composed of the samples which were captured at nearly the same time and at locations approximately fifty miles apart by sea. With the exception of the G.B. sample, the remainder of that salt water cluster were captured in the same year. It is also interesting to note the order of clustering in the freshwater samples. The two samples from the upper Great Lakes, L.S. and L.H. cluster first, indicating their greater degree of similarity. The G.L. sample is then added

while the L.E. sample is then quite remote, as evidenced by the fact that it is the last to form any cluster.

(L) Combination of Intersample Distance Matrices from ALLALL and ESTALL

Interpopulation distance matrices had been generated from the coordinates of the various analyses. The distances from ALLALL (Table 11) resulted from the consideration of the meristic and morphometric data while those from ESTALL (Table 20) were the result of the dispersion in the esterase data. A joint analysis was not run, as there were different numbers of individuals (20 versus $\geq$ 96) for whom the different types of data were available. However, a comparison of the dispersion of the samples based on both types of data was sought.

The interpopulation distances for all analyses were computed using Equation #3. From this equation, it can be seen that any number of coordinates can be used in the calculation of d_{ij}. Thus, to calculate the interpopulation distances based on the coordinates of ALLALL and ESTALL, both sets of coordinates could be used in the equation. In this case, the total number of coordinates (n) would be equal to forty (32 + 8).[1] However, as the individual distance matrices had already been computed, they were combined using

---

[1]Although 40 sets of coordinates are possible, only (12 + 8) = 20 sets have non-zero value in this work.

the formula"

$$D_{ij}\ (\text{Total}) = (D^2_{ij}\ (\text{ALLALL}) + D^2_{ij}\ (\text{ESTALL}))^{\frac{1}{2}} \qquad (11)$$

where D_{ij} (Total) = Mahalanobis's distance between the i^{th} and j^{th} population.

The results obtained are seen in Table 22. These values are the same as those that would have been obtained by solving Equation 3, with all forty coordinates, ignoring any possible correlations between the meristic, morphometric variables and esterase variables.

Although, there was no information available concerning the esterase patterns for the two Lake Heney samples, it was desirable to include these two populations in the final analysis. To do this, it was necessary to estimate a distance component for their unavailable esterase components. This was done by summing all the elements in Table 20 and taking the average. This value (2.311) was then used in Equation 11 to find the estimated D_{ij} (Total) values for L.H.G. and L.H.S. These values are seen as the bottom two rows of Table 22.*

(1) Cluster Analysis

The results of the M.S.T. and U.P.G.M. analyses done on the combined distance matrix are seen in Figure 9(A) and Figure 9(B). As the ALLALL analysis contributed by far the greatest amount of the dispersion, the connections of the M.S.T. are shown on the two dimensional plot from that

---

*See page 233.

TABLE 22

DISTANCE MATRIX ($\sqrt{D^2}$) DERIVED FROM COMBINED MERISTIC MORPHOMETRIC AND ESTERASE DATA FROM THIRTEEN SMELT SAMPLES

	N.A.(69)	C.B.	L.R.	S.B.	I.R.	N.A.(70)	G.B.	L.S.	L.H.	L.E.	L.H.G.	L.H.S.
C.B.	7.31											
L.R.	5.65	5.32										
S.B.	6.18	5.33	2.96									
I.R.	5.24	5.43	5.34	5.20								
N.A.(70)	6.25	5.07	5.09	4.92	3.35							
G.B.	6.71	6.39	6.35	5.52	4.09	4.83						
L.S.	7.51	7.05	6.71	5.40	6.74	7.15	6.15					
L.H.	9.39	7.69	7.53	6.43	6.87	7.83	6.36	4.99				
L.E.	7.96	7.78	7.35	6.59	5.92	7.07	5.49	5.22	5.54			
L.G.G.*	9.36	8.56	8.71	7.92	6.95	8.16	6.43	7.48	7.72	6.06		
L.H.S.*	20.30	18.59	19.80	19.34	19.94	20.80	19.31	17.95	17.81	18.40	20.64	
G.L.	20.32	18.81	20.11	19.77	20.18	21.27	19.56	18.09	17.90	18.79	21.17	5.06

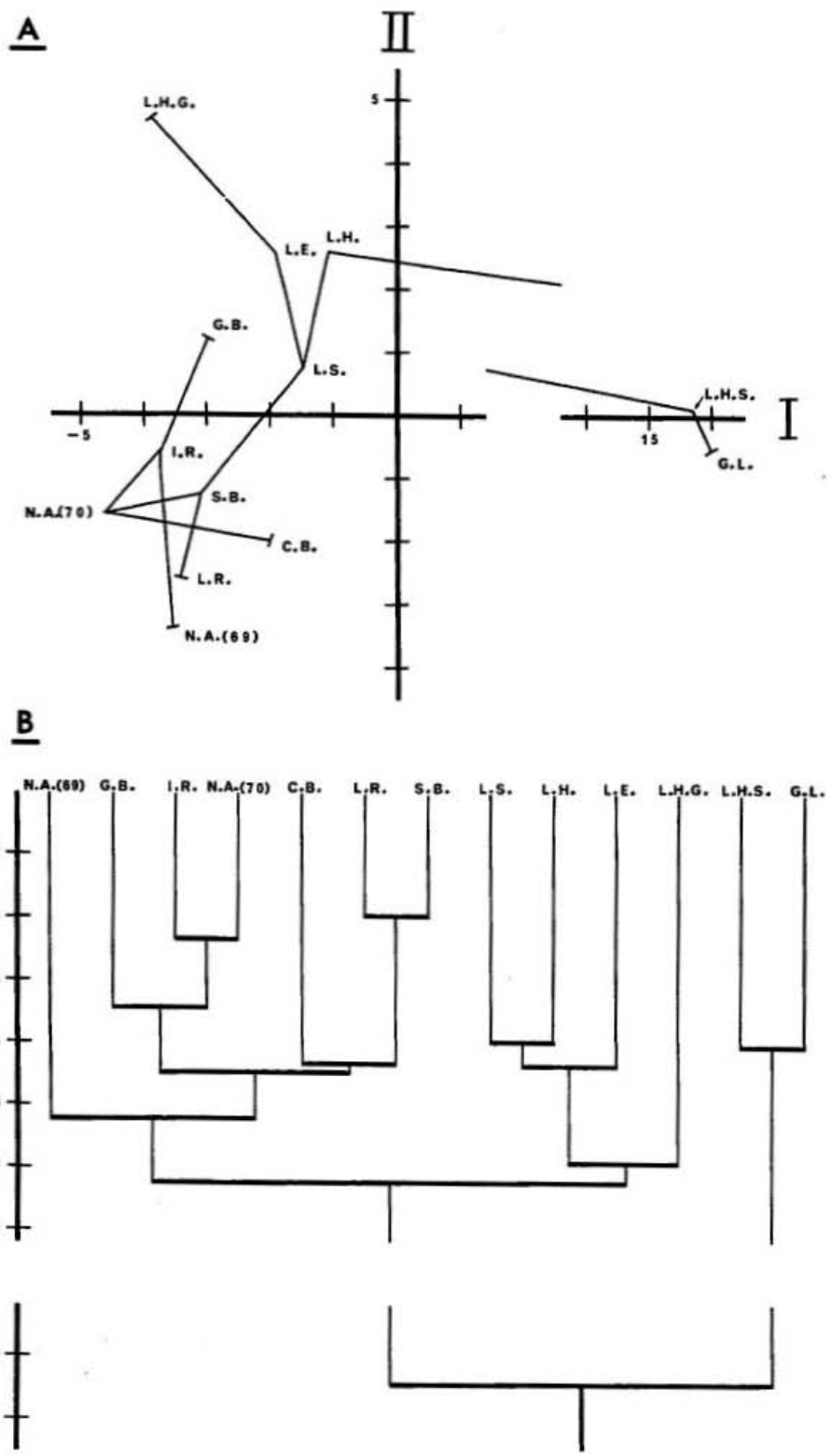
*estimated esterase component

FIGURE 9(A)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II FROM ALLALL AND M.S.T. COMPUTED FROM INTERSAMPLE DISTANCE MATRIX DERIVED FROM THE COMBINATION OF ALL CANONICAL AXES OF ALLALL AND ESTALL

FIGURE 9(B)

DENDROGRAM OF SMELT SAMPLES FROM U.P.G.M. CLUSTER ANALYSIS ON INTERSAMPLE DISTANCE MATRIX DERIVED FROM THE COMBINATION OF ALL CANONICAL AXES OF ALLALL AND ESTALL


A
II
I
5
−5
15
L.H.G.
L.E.
L.H.
G.B.
L.S.
L.H.S.
G.L.
I.R.
S.B.
N.A.(70)
C.B.
L.R.
N.A.(69)
B
N.A.(69)
G.B.
I.R.
N.A.(70)
C.B.
L.R.
S.B.
L.S.
L.H.
L.E.
L.H.G.
L.H.S.
G.L.
2
3
4
5
6
7
8
19
20

analysis. For the same reason, the M.S.T. from this analysis will be compared against that of ALLALL. The changes in the tree, can generally be accounted for by the displacement of the L.E. sample from the rest, when the esterase data were also considered. The first difference in this tree compared with the one of Figure 6(A) is that L.H. now joins the rest of the samples through a connection with L.S. sample. From the esterase tree (Figure 8[A]) it was seen that while the L.S. - L.H. link was short, the L.E. sample was remote from both of them. Thus, the L.H. - L.S. link is re-established. Secondly, the L.E. - G.B. lir' has been broken in favour of the S.B. - N.A. (70) link which has formed. Again, the distance between L.E. and G.B. from the esterase component was great enough to cause the L.E. - G.B. link not to be the shortest joining, the N.A. (69) - C.B. - N.A. (70) - I.R. - G.B. cluster, to the rest. This shortest link, of all the possible ones, was now S.B. - N.A. (70). The rest of the linkages remain the same as in Figure 6(A). The result of these two linkage changes has been to form a network amongst the salt water samples which is combined with the freshwater network by a single connection, S.B. - L.S. This is a desirable result, from the biological point of view, as the two groups come from different environments. This compares with the M.S.T.'s from the esterase data (Figure 8[A]) and three canonical axes of ALLALL, (Figure 5[A]) in which there were only single linkages between the two groups.

However, it contrasts with the M.S.T. from all axes of ALLALL which had two connections between the two groups.

The results of the M.S.T. (Figure 9[A]) were again used to position the ends of the branches of the dendrogram. The L.H.G., L.H.S. and G.L. samples were positioned for the same reasons as before. For the L.S., L.H., L.E. group, L.S. was placed next to S.B. and this caused L.H. to take the position indicated. L.E. was placed to the right of L.H. because it did form a link with L.H.G. and no longer had any connection with the salt water samples which lie to the left. The remainder of the positions are pretty straight forward. C.B. formed a link with N.A. (70) while G.B. formed a link with the I.R. sample. The presence of the I.R. - N.A. (70) cluster thus caused the remainder of the samples to be positioned as seen. The N.A. (69) sample was again on the extreme end of salt water group.

The U.P.G.M. cluster analysis resulted in a very "acceptable" dendrogram in that it reflects many of the temporal, environmental and geographical factors. From the bottom up, the two clusters seen at the 19 level are the stunted samples and the remainder of the freshwater and salt water samples. The latter group then divides to form two clusters, all the freshwater samples and all the salt water samples. The L.H.G. sample joins the freshwater cluster after the three Great Lakes samples have joined. The Great Lakes cluster is arranged, coincidentally, in a

fashion which reflects the north to south location of the sample sites. In the salt water cluster, N.A. (69) is the last sample to join the cluster. At the next step up, the two clusters that join are each composed of samples taken during the same year. That is, the G.B., I.R. and N.A. (70) samples were all captured in 1970, while C.B., L.R. and S.B. were captured in 1969. In each of these two "year class clusters," the samples that were temporally and geographically close together cluster first. That is, the I.R. - N.A. (70) and L.R. - S.B. are the first two salt water clusters to form.

(M) Transformed Data

When all the transformed variables and meristics were used in the analysis TGNMER, the results obtained were very similar to those of the analysis ALLALL. However, in these data, (TGNMER) there were still two variables which contained information which was directly related to size. These were $\log_{10}$ St.L. and Wt./Age. Although transforming length to its $\log_{10}$ form reduces the magnitude of the variables, the standard deviation is also reduced. Therefore, the ratio, $\bar{d}/\sigma$, remains large and the samples can still be differentiated on the basis of their size. Similarly, the calculation of a rough growth rate (Wt./Age) is also size dependent. This is because the variation of ages between the populations is not as great as that of weight. This

results in a larger fish having a greater value for Wt./Age and therefore, this also results in the separation of the samples due to a size factor. However, it was desirable to have an analysis of the data, in which no variable containing an obvious size factor was present. This was done by rerunning TGNMER with the two size related variables omitted.

(1) Classification from SZOTME

In the analysis SZOTME, all apparently size-related variables were omitted and meristics were also included. The classification matrices which resulted from this analysis provided some interesting results (Table 9). Although the 90% correct classification level was not reached for the individual samples when nineteen variables had been included, extrapolation from the curves suggested that this level would have been reached, if the number of variables in the discriminant functions were 22. The other key levels of classification and the number of variables included in the discriminant functions are seen in Table 8. The salt water fish can be correctly separated from the freshwater fish 76% and 91% of the time with the use of one and two variables respectively. The two stunted samples were separable from the rest of the fish at the 92% level with the use of a single variable, (Or./I.O.W. x 10). With the

inclusion of the second and subsequent variables, the stunted samples were completely distinguishable from the rest. These results show that discrimination between samples is possible, with the use of variables which are not obviously related to size. Also, the distinctions between the freshwater and salt water samples and the stunted and other samples are quite large using this type of information.

(2) Canonical Analysis

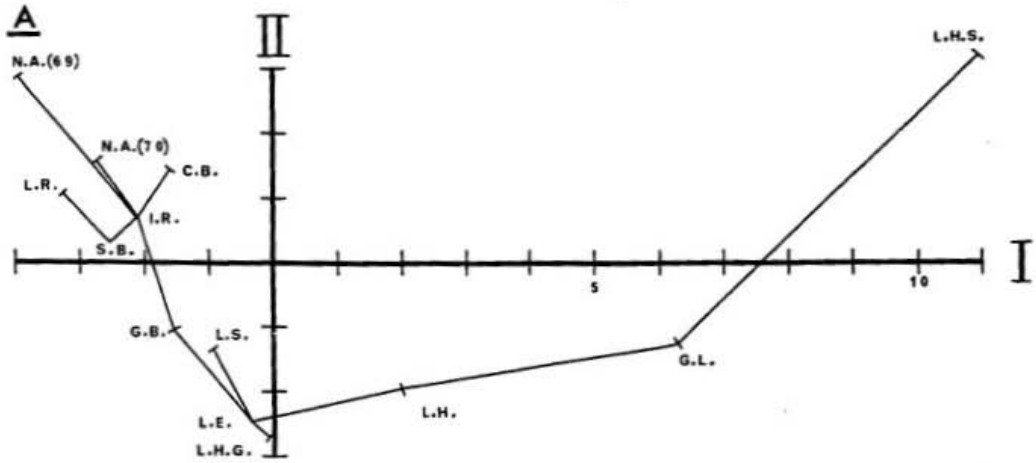
When the canonical analysis was run on the data, the coordinates of the sample means were found, as previously described. The plot of the samples' means along the first and second canonical axes is seen in Figure 10(A). From this figure, the salt water samples, with the exception of the G.B. sample, are located in the upper left corner of the plot. The I.R. sample appears to be the central sample of this rather tight cluster. The freshwater samples are seen in the lower middle portion of the plot. The L.E. sample appears to be central. From the two dimensional representation, the L.S. sample appears as the closest freshwater sample to the salt water group. The most striking point about this plot, when compared with that of ALLALL, is that the two stunted samples have been dispersed. They are still removed from the body of the samples, and the L.H. sample forms a bridge between these two and the bulk of the samples.

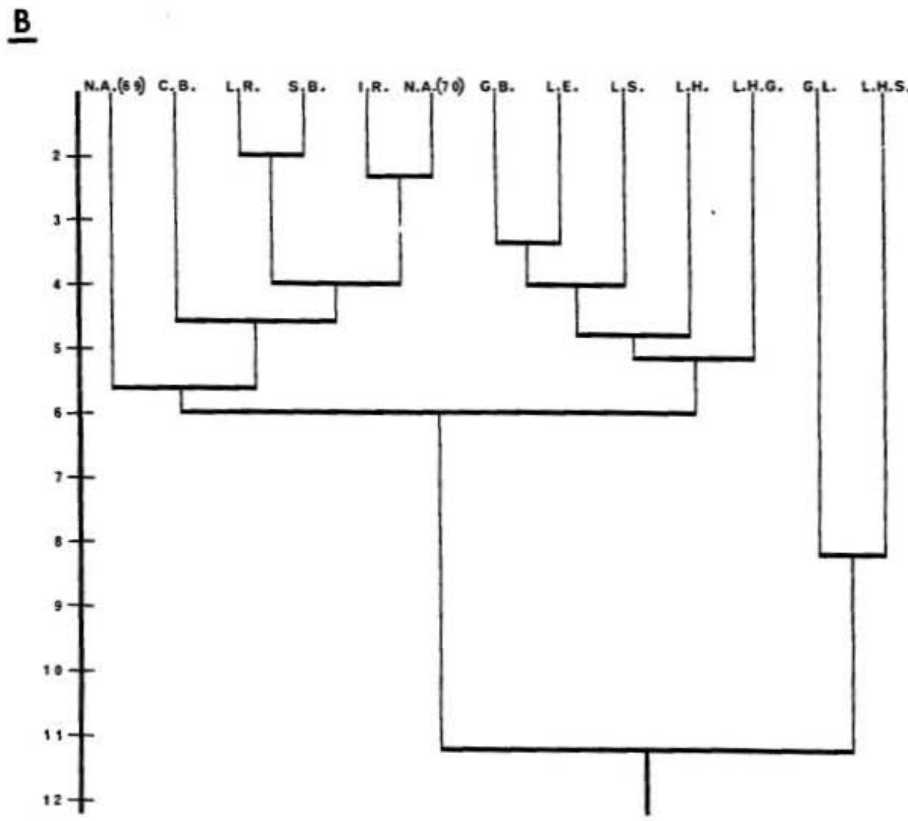
FIGURE 10(A)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II AND M.S.T. COMPUTED FROM INTERSAMPLE DISTANCE MATRIX DERIVED FROM ALL CANONICAL AXES IN THE ANALYSIS SZOTME

FIGURE 10(B)

DENDROGRAM OF SMELT SAMPLES FROM U.P.G.M. CLUSTER ANALYSIS ON INTERSAMPLE DISTANCE MATRIX DERIVED FROM ALL CANONICAL AXES IN THE ANALYSIS SZOTME


A
II
I
N.A.(69)
N.A.(70)
C.B.
L.R.
I.R.
S.B.
L.H.S.
5
10
G.B.
L.S.
G.L.
L.E.
L.H.
L.H.G.


B
N.A.(69)
C.B.
L.R.
S.B.
I.R.
N.A.(70)
G.B.
L.E.
L.S.
L.H.
L.H.G.
G.L.
L.H.S.
2
3
4
5
6
7
8
9
10
11
12

(3) Cluster Analyses

After the inter-sample distance matrix had been generated ($[D^2]^{\frac{1}{2}}$) as M.S.T. and U.P.G.M. cluster analysis were performed. Many of the links in this tree have been seen in the ALLALL analyses (Figure 6[A]). However, the major difference is that all the salt water samples connect with I.R., except for the L.R. sample which connects to I.R. through the S.B. sample. Also, the L.S. - S.B. link which was previously seen (Figures 6[A] and 9[A]) does not exist in this tree. The connection between the salt water and freshwater groups is through the G.B. - L.E. link. The M.S.T. also shows that although L.S. and G.B. appear close together in two dimensions, the distance between G.B. and L.E. is shorter when many dimensions are considered.

There is no indication from the M.S.T. of Figure 10(A) that there are two elongated clusters with this data. Instead, there appear to be two rather compact clusters (freshwater and salt water) and the two stunted samples extend to one side.

The results of the U.P.G.M. cluster analysis are seen in Figure 10(B). The positions of the arms of the dendrogram were arranged from the information in the M.S.T. There are two primary clusters, the two stunted samples and the remainder of the samples. The latter cluster splits to form one group composed of the salt water samples and a second group containing all the freshwater samples and the

G.B. samples. The salt water group contains the two subclusters composed of those samples that were geographically and temporally close together. The freshwater group does not contain any subclusters but consists of individual additions to the existing core of the cluster.

Two interesting points emerge from this dendrogram. The two stunted samples are not as close together as had been seen in Figures 5(B), 6(B) and 9(B). At the stage in the clustering procedure just prior to the joining of the freshwater and salt water populations, the G.L. sample has an average distance of 7.95 units from the samples in the freshwater cluster. Had the salt water and freshwater samples not joined at 5.90 but at some distance greater than 7.95, the G.L. would have joined the freshwater cluster, rather than joining with the L.H.S. sample at 8.18. Therefore, on the basis of this data, the G.L. sample is intermediate between the freshwater group and the L.H.S. sample.

Secondly, G.B. has joined the freshwater cluster through its link with the L.E. sample. From the distance matrix, it was seen that G.B. was actually closer to I.R. than to L.E. However, the I.R. - N.A. (70) cluster formed first and as a result, the distance between G.B. and the cluster was greater than between G.B. and the L.E. sample. Therefore, G.B. was incorporated in the freshwater cluster.

(N) Combination of Intersample Distance Matrices From SZOTME and ESTALL

The intersample distance matrix for the combined effect of the SZOTME and ESTALL data was computed using the individual distance matrices and Equation 11. The average value of the ESTALL matrix, 2.31, was again used for the two Lake Heney populations.

(1) Cluster Analysis

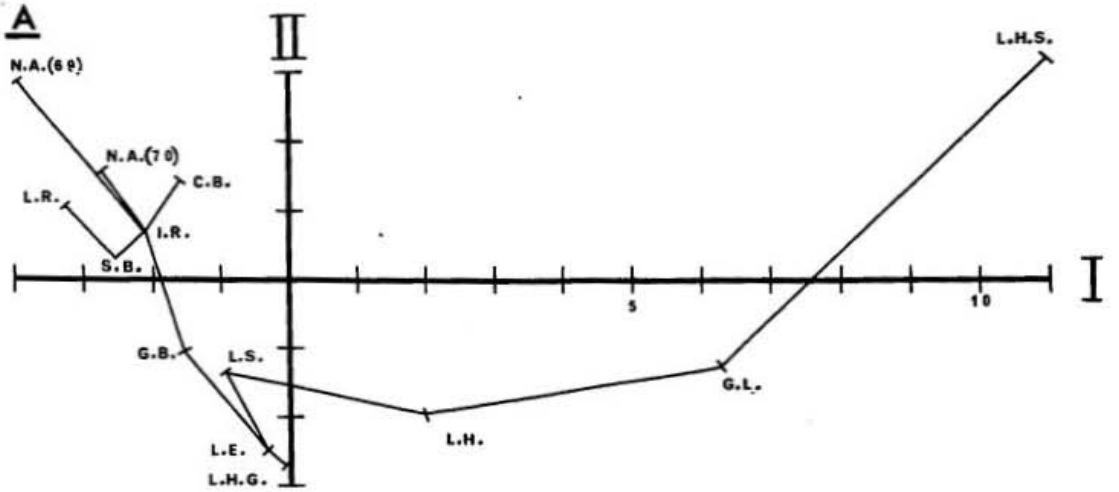
The results of the M.S.T. and U.P.G.M. analyses are seen in Figures 11(A) and (B). The M.S.T. is represented in the plot of the first and second canonical axes of SZOTME as this analysis contained a greater amount of sample dispersion than ESTALL (Table 4). This M.S.T. is very similar to that of SZOTME. The only difference is that L.H. now joins the body of the samples through L.S. rather than through the L.E. sample. From the M.S.T., there was no need to change any of the positions of the branches of the dendrogram that was seen with SZOTME (Figure 10[B]). The major differences in the two dendrograms is that the G.B. sample is now part of the salt water cluster. Secondly, the lengths of the stems of the salt water and freshwater clusters are now longer than with SZOTME. This indicates the increased differences between the freshwater and salt water groups. Again, there is a general lengthening of all stems with the combination of the two types of data.

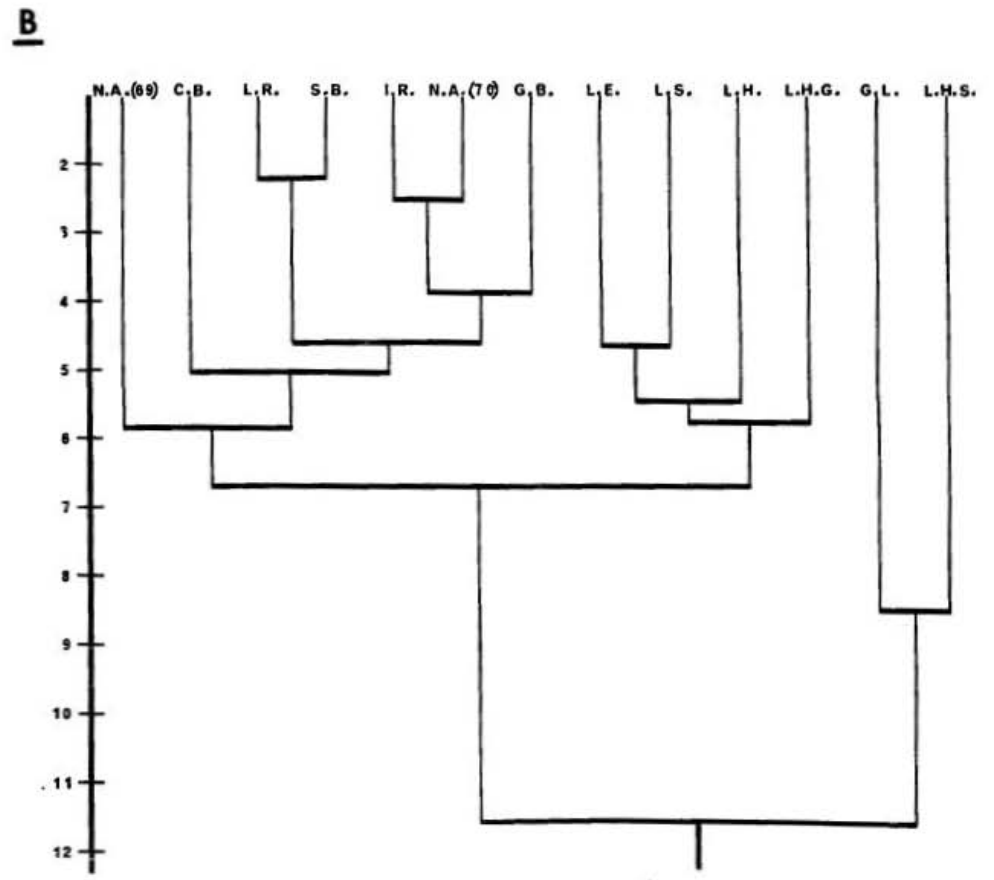
FIGURE 11(A)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II FROM SZOTME AND M.S.T. COMPUTED FROM INTERSAMPLE DISTANCE MATRIX DERIVED FROM THE COMBINATION OF ALL CANONICAL AXES OF SZOTME AND ESTALL

FIGURE 11(B)

DENDROGRAM OF SMELT SAMPLES FROM U.P.G.M. CLUSTER ANALYSIS ON INTERSAMPLE DISTANCE MATRIX DERIVED FROM THE COMBINATION OF ALL CANONICAL AXES OF SZOTME AND ESTALL


A
II
I
N.A.(69)
N.A.(70)
C.B.
L.R.
I.R.
S.B.
L.H.S.
5
10
G.B.
L.S.
G.L.
L.H.
L.E.
L.H.G.


B
N.A.(69)
C.B.
L.R.
S.B.
I.R.
N.A.(70)
G.B.
L.E.
L.S.
L.H.
L.H.G.
G.L.
L.H.S.
2
3
4
5
6
7
8
9
10
11
12

(O) Effect of Using a Reduced Data Set

In all the previous analyses which dealt with dispersion pattern of the thirteen populations, a large number of variables were included in the discriminant functions (Table 4).

The use of a large number of variables also means the solution of the canonical equation will require the use of a sophisticated computer and programme. The question, therefore arose, to what extent can a reduced data set mimic the results obtained by using a more complete set of variables.

It was decided that 3 variables would constitute a reduced data set. It was also decided to run three analyses on three different reduced data sets. The variables in each set were selected in the following manner. The analysis 3RAW01 contained the first three variables which were included in the discriminant functions of ALLALL, (Table 5). The second analysis 3RAW02 contained the first, fourth and fifth variables entered in ALLALL. The third analysis 3TGN01 contained the first three variables included in the analysis SZOTME and tested the effect of "size independent" variables (Table 5).

The results of plotting the samples along the first and second canonical axes of each analysis are seen in Figure 12(A), (B) and (C). Similarities are noted between

FIGURE 12(A)

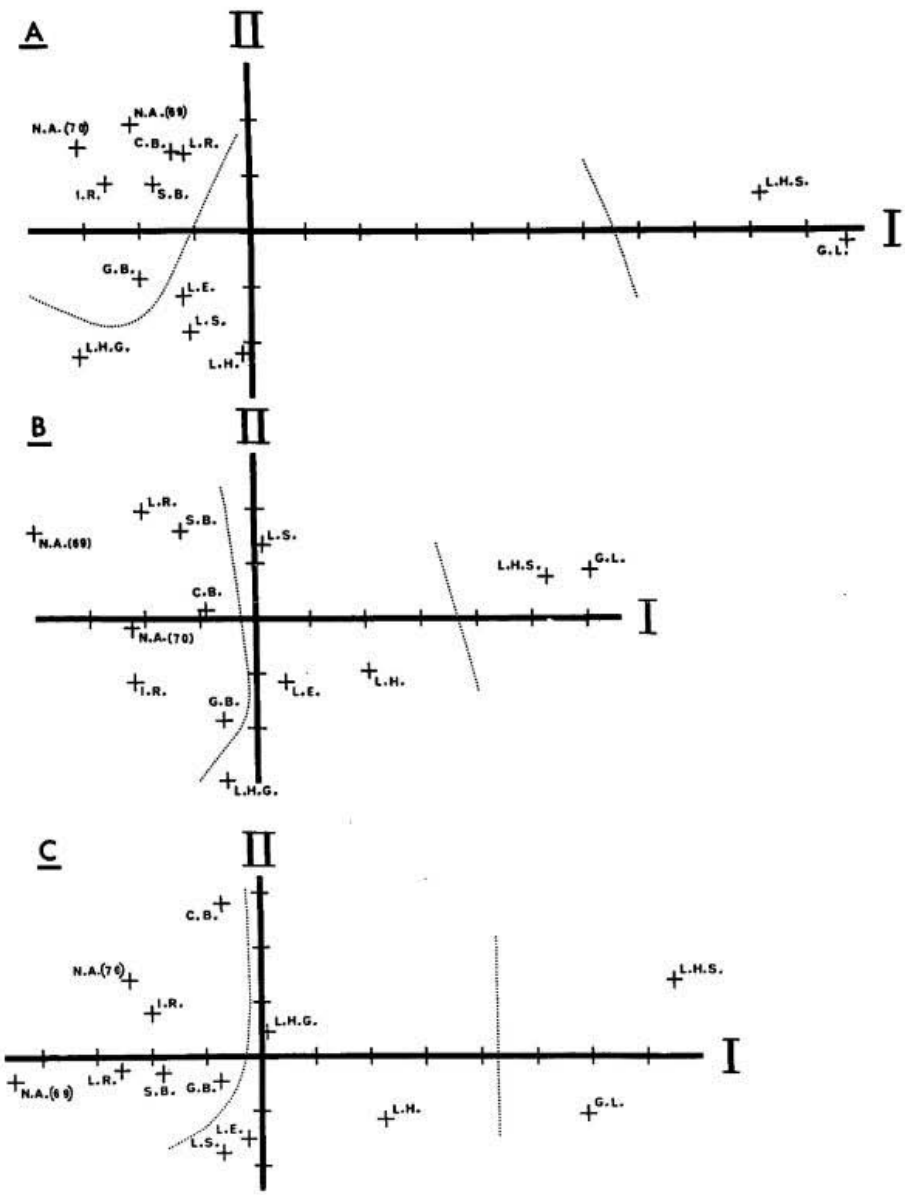
PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II FROM 3RAW01

FIGURE 12(B)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II FROM 3RAW02

FIGURE 12(C)

PLOT OF SMELT SAMPLES' MEANS ALONG CANONICAL VARIATE AXES I AND II FROM 3TGN01


A
II
I
N.A.(70)
N.A.(69)
C.B.
L.R.
I.R.
S.B.
L.H.S.
G.L.
G.B.
L.E.
L.S.
L.H.G.
L.H.
B
II
I
L.R.
S.B.
L.S.
N.A.(69)
L.H.S.
G.L.
C.B.
N.A.(70)
I.R.
L.E.
L.H.
G.B.
L.H.G.
C
II
I
C.B.
N.A.(70)
L.H.S.
I.R.
L.H.G.
L.R.
N.A.(69)
S.B.
G.B.
L.H.
G.L.
L.E.
L.S.

the plots from the three analyses. In each case, the stunted samples are separable from the remainder of the samples by a line, as indicated. The freshwater samples can be separated from the salt water forms by a straight line, however, to more clearly indicate the regions, a curved line was used in the figures. The Grand Bank sample, G.B., is again found in a position rather close to the freshwater samples in each of the three analyses. It was noted that similar results were seen when many variables were used in the canonical analysis (Figures 3 and 10[A].

There are also differences observed between the plots from the three analyses. It was noted that the stunted samples were further removed from the remainder of the samples in Figure 12(A) than in Figure 12(B). The results from Figure 12(C) are intermediate in this respect. On the other hand, the 'normal' samples appeared to be more dispersed in Figure 12(B) than in Figure 12(A). Again, Figure 12(C) appeared to be intermediate with respect to this point. In Table 4, it was noted that these analyses in order of decreasing total dispersion ($\Sigma\lambda_i$) were 3RAW01, 3TGN01 and 3RAW02. However, in Table 8 it was seen that in order of decreasing percent correct classification, the order was 3RAW02, 3TGN01 and 3RAW01.

To investigate this unexpected reversal in order of 3RAW01 and 3RAW02 on the basis of these different criteria, investigations were made into the intersample distances for

each analysis. These results are seen in Table 23. Two intersample distance matrices were calculated for each analysis. In the first case, the intersample distances were calculated between all thirteen of the samples. A second distance matrix was calculated between the 'normal' samples in each analysis. For every distance matrix, the sum of all the intersample distances and the average distance was calculated. These results as well as the percent correct classification amongst the samples under consideration are also seen in Table 23.

The sum of the distances for the first three distance matrices produce results which are in agreement with the order of the sum of the latent roots (Table 4). However, when only the eleven normal samples are considered this relationship does not hold. A new association emerges which provides an understanding of the apparent discrepancy between the total dispersion and the separability of the samples. It was noted that the order of the average distances for the second three analyses in this set is reflected by the order of the level of correct classification seen in these analyses, as well as that of the first three. Therefore, the level of correct classification appears to be linked more closely to the average intersample distance of the normal samples than to the total dispersion of the samples.

TABLE 23

EFFECTIVE DISTANCE BETWEEN SMELT SAMPLES, RELATED TO THE PERCENTAGE OF CORRECT CLASSIFICATION AND THE TOTAL VARIATION BETWEEN SAMPLES FOR THREE ANALYSES EMPLOYING REDUCED DATA SETS

Analysis	$\Sigma \lambda_i$	Number of Populations	Number of Distances	ΣD_{ij}	$\overline{D}_{ij}$	% Correct Classification
3RAW01	22.81	13	78	420	5.39	60.4
3RAW02	11.69	13	78	345	4.42	68.8
3TGN01	13.89	13	78	361	4.62	63.5
3RAW01	22.81	11	55	157	2.85	53.2
3RAW02	11.69	11	55	187	3.41	63.3
3TGN01	13.89	11	55	175	3.19	56.8

$\Sigma \lambda_i$ = Sum of the Latent Roots

Number of Populations = Number of Populations Considered in the Distance Matrix

Number of Distances = Number of Elements in the Distance Matrix

ΣD_{ij} = Sum of the Elements in the Distance Matrix

$\overline{D}_{ij}$ = Average Distance in the Distance Matrix

DISCUSSION

In the Introduction, the hypothesis was put forward that as *O. e. mordax* exists in rather isolated populations, they could be expected to exhibit a certain degree of divergence. Therefore, the initial problem became to determine if any divergence between the available samples could be demonstrated. When the initial analysis was performed using all variables and all samples, (ALLALL), the approximate F statistic to test the equality of group means was very highly significant ($F = 11.71$ with degrees of freedom = 384 and 2472). In fact, on the basis of the first or second variable there was no evidence that the group means were the same ($p << .001$). Therefore, from the initial analysis, it was determined that the groups were not homogeneous and further analysis could then proceed. It was also seen (Figure 3, Results) that there appeared to be a pattern in the dispersion of the samples. The stunted samples were far removed from the normal ones and a line could be drawn separating salt water from freshwater samples (Figure 3, Results).

The next step, was to study the effect of the deletion of certain key variables from the data set, to ascertain their value in classification and identification.

In the series of analyses ALLN01 - ALLN03, highly significant variables were successively removed. It was expected that the removal of key variables would result in considerable change on the classification matrices (Table 8). However, this did not appear to occur (Section D[2], Results). On the other hand, it was noted that there was a sizeable reduction in the total dispersion between ALLN02 and ALLN03. Further study of this situation indicated that this reduction in the sum of the latent roots had its primary effect along canonical axis I (separating the stunted and normal samples). The reduction in dispersion along the remaining axes was not nearly as dramatic (Table 4, Results). From this series of analyses, it became apparent that the source of the divergence among smelt populations that was initially seen (Figure 3), was not to be found in a few select variables, but was a more general phenomenon.

Once it had been established that the various populations showed diversity to a large degree and on many levels, it became desirable to perform further analyses. The initial analyses contained both meristic and morphometric characters. From Table 5 (Results), it was seen that a meristic character was found amongst the first five variables in ALLALL through ALLN03. Also, McAllister (1963) found that meristic characters were particularly useful in making decisions with regard to the existence of many supposed species of *Osmerus*. Therefore, it was decided to

examine the meristic and morphometric characters separately to compare their respective values in the identification and separation of samples. Although there were twice as many variables in MORPHO, there was five times more variation in MORPHO than in MERIST (Table 4, $\Sigma\lambda$). By examining the values in Table 8 (Results), it can be seen that the two types of characters appear to fulfil different functions in the classification of the samples. Neither the 75% nor 90% level of correct classification was reached (61.1% and 66.2%) for the first two subdivisions of the classification matrix with MERIST. On the other hand, about a 20% higher level of correct classification was achieved for MORPHO (79.4% and 85.5%) when ten variables (maximum number for MERIST) were included. That is, the morphological characters appear to be of greater importance when one is attempting to sort individuals into their own specific groups. However, if one considers the third subdivision of the classification matrix (Table 8, Freshwater from Salt Water), the importance of the meristic characters can be seen. Only one meristic variable was required to separate the two groups with a degree of success greater than 75% (89.2%), while in MORPHO considering only the first variable did not markedly improve the success achieved (56%) over chance. Similar results were seen for the inclusion of the second variable in each analysis (92% versus 68%). Therefore, the distinctions between the salt and freshwater forms were more easily

seen with meristics, while morphometric characters became valuable in assigning individuals to their correct group. With either type of data, the stunted samples were clearly distinct from the others.

To further investigate the apparent distinctions between and within the salt and freshwater samples, other analyses were performed with suitably arranged data sets. When the seven salt water samples were combined and run with the four combined normal freshwater samples, the importance of the meristic characters was reconfirmed. The first three variables to be included in the salt water - freshwater between groups analysis (SWTFWT) were meristic. Again, only two variables were required to correctly assign (separate), more than 90% of the individuals to one of these two groups. When the within normal freshwater analysis was performed (FWTALL), greater than 75% separation was achieved with the inclusion of two variables, while the 90% level was achieved with four. All of these four characters were morphometric which might have been expected on the basis of results from MORPHO and MERIST. Again, when the dispersion amongst the seven salt water samples was analysed (SWTALL), four of the first five variables included in the discriminant functions were morphometric (Table 5). It was also noted, that even when the nearest neighbours were combined, resulting in five groups, eight variables were required to achieve 90% separation. This contrasts with the results from FWTALL

in which only four variables were required to achieve 90% correct classification. This indicates that the four normal freshwater samples exhibited a greater amount of diversity than the salt water groups, in that half the amount of information was required to produce equally successful decisions.

In the Introduction, it was suggested that comparisons might be made between analyses in which raw data was used and those in which the data had been transformed as in Table 2 (Results). The earliest attempts at a quantitative assessment and comparison of shapes were almost all conducted with the aid of ratios of characters (Blackith and Reyment, 1971). This was done in an attempt to standardize variation and to eliminate the effects of size. Their longstanding use has resulted in the "acceptance" (Blackwelder, 1964) of this method as a primary operation in the comparison of samples. The main arguments against the use of ratios are that they do not in fact produce constants and that they provide only a poor appreciation of involved contrasts in form.

To test the effects of using transformed variables, four analyses were run. TGNMER and SZOTME are comparable with ALLALL in that characters of both meristic and morphometric origin were included. On the other hand, TGNALL and SZOTMO can be compared with MORPHO in that only morphometric characters and their transformations were included. By

selectively combining information in Tables 4 and 8 (Results), the similarities between the runs can be noted. When the size-related variables were removed from the data sets (SZOTME and SZOTMO) there was a considerable drop in the total dispersion ($\Sigma\lambda$, Table 4, Results), compared with the appropriate analyses (ALLALL or TGNMER and MORPHO or TGNALL, respectively). Information from Table 8 (Results) indicates that for the two groups of analyses (with and without meristics) those in which meristic characters are included reach a slightly higher level of correct classification than do the analyses in which they are omitted. Also SZOTME and SZOTMO produced lower levels than the two other analyses in each of their respective groups. However, the key point from these comparisons is that little difference between the matched runs (ALLALL and TGNMER or MORPHO and TGNALL) was observed. Not only were their sums of latent roots high (Table 4, Results) and their classification successes extremely similar (Table 8, Results) but their respective linkages in U.P.G.M. cluster analyses were in the same order (unpublished results). The dispersion patterns and cluster analysis results for the unmatched but comparable analyses, ALLALL and SZOTME are seen respectively in Figures 6(A) and (B) and 10(A) and (B), (Results). In both of these figures, the salt water, freshwater and stunted samples occupy different areas of the graphs resulting from plotting the sample means along canonical axes I and II. Similarly,

the respective U.P.G.M. cluster analyses indicate that these same major groupings hold. The differences seen in the dendrograms (Figure 6[B] and 10[B]) are basically the same except that the Grand Bank sample (G.B.) joins the freshwater cluster and the N.A. (69) and L.H.G. samples respectively join the salt water and freshwater clusters before their amalgamation. It should be noted however, that there are no essential differences between these two sets of results and the same conclusions that are reached about the structure of the dispersion based on raw data is neither contradicted, clarified nor made equivocal, by the use of transformed data. Therefore, there seems little point in continuing with investigations of this type, using transformed data.

It has been noted that in ALLALL and TGNMER, the primary dispersion pattern is between the normal and stunted samples. Therefore, it is reasonable to assume that the presence of the data from the two stunted samples would have an effect in the order in which the variables are included in the discriminant functions. To test this hypothesis, ALLXHM and TGNXHM were run with the two stunted samples omitted and using raw and transformed data respectively. In each of these analyses, the total dispersion was reduced (Table 4, Results) and approximately equal to the sum of the second through twelfth latent roots of ALLALL and TGNMER respectively. This again indicates that most of the

dispersion along canonical axis I is concerned with the separation of the stunted populations from the normals. However, in both analyses the freshwater and salt water forms can be clearly separated. There is also a difference in the order of inclusion of the variables in analyses with and without the stunted samples. In both comparisons, size-related variables were removed or demoted (Table 5, Results) when the stunted samples were omitted.

Biochemical characters provided yet another source of data with which the divergence between the samples could be investigated. As had been noted (Section I, Results), the pattern obtained for muscle myogens (Figure 7[A]) appeared to be 'constant' and was therefore of little value in an investigation of difference and diversity. The results obtained by investigating the esterase complements of the individuals provided considerable variation, within and between samples, and was therefore of use in evaluating intersample differences. When the discriminant function analysis ESTALL was performed, the total dispersion of the samples was small (2.66) when compared with the other analyses (Table 4, Results). However, the size of this value compares favourably with that reported by Calaprice (1971) for X-Ray spectrometric data from sockeye salmon. It was also noted that the data from the esterase investigations were not as efficient as the morphometric and meristic data for the correct identification of individuals (Table 8,

Results). However, 37.4% correct classification represents more than a four fold increase over the 7.7% expected on the basis of chance. Similarly 41.1% represents a 3.7 fold increase over chance. Therefore, the information from the esterase system is valuable in the identification of the group to which an individual belongs. It was also seen (Table 8, Results) that esterase information achieves maximum value in the distinction between the salt water and freshwater smelt. The level of correct identification compares well with that obtained from the other types of data (89.3% versus 95.8% for FWTSWT). When canonical analysis was performed (Figure 8[A], Results) the major distinction was between the salt and freshwater samples. Also, the sample from Lake Erie was rather far removed from the remainder of the freshwater samples. These relationships were also seen in Figure 8(B), (Results) in which the distance between the Lake Erie sample and the freshwater group was greater than that between the salt and freshwater groups. Again, the sample from Grand Bank provided the connection between the two main groups.

When one has m samples and p variables and wishes to perform a canonical analysis, it is mathematically desirable that $p \geq m-1$ (Seal, 1964). However, in almost any group of morphological characters, a considerable amount of redundant information may be expected to exist. Blackith and Reyment (1971) have reported on a number of studies in

which highly similar results were obtained with the use of either large or reduced data sets. The analyses 3RAW01, 3RAW02 and 3TGN01 were performed to investigate the extent to which the information obtained from ALLALL and TGNMER (33 and 25 variables) could be reproduced by employing reduced data sets (3 variables).

The selection of the three variables in each data set was not random but based on the order in which the variables had been included in the discriminant functions of ALLALL and SZOTME (Table 5, Results). The variables employed in the reduced data sets were also seen in Table 5 (Results). From Figure 12(A), (B) and (C), the major groupings of the samples can still be discerned using only three variables. Therefore, it appears that in studies of this type (samples with comparable quantitative characters), many important conclusions can be drawn by the use of a few selected characters. In the U.P.G.M. cluster analysis performed on the distance matrix produced from the coordinates of 3RAW01, (unpublished results) the linkages in the resulting dendrogram were similar to those of ALLALL (Figure 5[B], Results). From these results, the hypothesis that considerable redundancy existed in the data seemed to be substantiated. However, there is little that can be done from theoretical considerations of redundancy. This question remains one which must be individually decided by experiment (Hotelling, 1951).

As has been noted (Table 23, Results) the order of the average distances did not reflect the order of correct classification, when the thirteen populations were considered with reduced data sets. This suggests a model to explain the level of correct classification with respect to distance. There is some ideal intersample distance ($d_{optimum}$) at which there is the least chance of incorrectly classifying an individual into some neighbouring group. If this distance is exceeded the proportion of the dispersion that is involved in this extension of the distance is "wasted", with regard to achieving an optimal separation of the samples with a given amount of dispersion. However, should there be insufficient dispersion to allow each intersample distance to reach the $d_{optimum}$ level, then any distance which exceeds $\bar{d}$ can be considered as non-optimal. This fact is based on the Normal density distribution. Therefore, to achieve optimal discrimination, one requires a group of samples which are uniformly distributed. If this type of situation occurred in nature it would have unfortunate and unproductive repercussions for taxonomists as meaningful classification would be impossible.

Table 23 (Results) shows that although 3RAW01 had both the greatest total dispersion ($\Sigma\lambda$) and average intersample distance, it was the poorest for percent correct classification. In 3RAW02 the reverse is true. By examining the values in Table 23, with a view to the model

just developed, one can gain an understanding of the reversals. In 3RAW01, much of the dispersion is 'wasted' by the excessive distances between the two stunted samples and the normal samples. That is, the average of the twenty-two intersample distances involving a normal and stunted sample was 12.0. This greatly exceeds the average distance between normal samples (2.85), and the situation is far from ideal for discriminating the various individuals. When one compares the distance values from 3RAW02, a more optimal situation now exists. The average stunted - normal distance is reduced to 7.3 which is approximately twice the average intersample distance of the normals (3.41). The samples in 3RAW02 are more uniformly distributed with greater distances between normal samples permitting a higher level of correct identification.

There are in effect two average distances; the real average distance and the effective average distance. The effective average distance is the average distance between the members of some subgroup. The effective average distance for 3RAW01 would result from considering only the 'normal' samples, or more accurately the freshwater samples and the salt water samples. The subdivision of the normal samples into freshwater and salt water groups is less clear in Figure 12(B) than it is in Figure 12(A). This underlines the second point in the model. With the more uniform dispersion of 3RAW02, the ease with which the samples can

subdivided or classified has been diminished. Again, the results of the 3TGN01 analysis are interpreted as intermediate between the other two analyses.

The analysis FWTSWT was originally carried out to investigate the distinctions apparent along the second canonical axis of ALLALL in more detail. When seven characters had been included in the FWTSWT discriminant functions, it was possible to correctly classify more than 95% of the individuals into their appropriate groups. This indicates that the groups are rather distinct, with little overlap. In the canonical analysis, there was only one non-zero latent root as there were only two groups in the analysis. The group means were 3.45 units apart along the first canonical axes. It was of interest to determine the difference between the two sets of data and thereby clarify the importance of the difference obtained along the canonical axis.

The group means and standard deviations were computed for the thirty-two variables in each data set (pooled salt water and normal freshwater samples). Then, t-tests were performed on these values to determine the significance level of each variable. Only nine of the variables were not significant ($p > 0.10$). Three variables were significant between 0.10 and 0.05, eight between 0.05 and 0.01 and one variable between 0.01 and 0.001. There were twelve variable significant at a level less than

0.001. Of these, eight were meristic characters and four (L.Mx., Ad.-T., B.A. and B.D.) were morphometric. Therefore, it again appears that the freshwater and salt water forms can be better separated on the basis of their meristics. The t-tests do not take into account the correlations between the various measurements, but the results do allow an appreciation of the degree of difference between the two data sets and how this difference relates to the discriminant function and canonical analyses. The t-tests also show in a univariate way, the degree of difference between the salt water and freshwater samples.

If similar analytical procedures had been employed to evaluate the difference between each pair of samples, it would have involved performing 2,574 t-tests. This monumental task can be avoided by gaining an appreciation of the significance of the distances between samples as determined by their positions along the canonical axes. To accomplish this, a matrix of misclassification was first constructed. This was done by adding the number of animals from sample A that were classified into sample B and the number of animals from sample B that were classified into sample A. This number was expressed as a percentage of the sum of the individuals in the two samples. Therefore, for each pair of samples, a percentage of misclassification and the distance between the group means was available. When the percentages of misclassification were plotted

against the intersample distances, a second order curve resulted. The analysis 3RAW01 resulted in a lower level of correct classification, compared with the other analyses. Therefore, it would have many non zero points for the percent misclassification. A second order regression analysis was run on these data and the resulting of the line was

$$\% \text{ misclassification}_{(ij)} = 26.72 - 12.2d_{(ij)} + 1.4d^2_{(ij)}$$

where $d_{(ij)}$ = the distance between the i^{th} and j^{th} populations

when $d_{(ij)}$ is greater than 4.36, the level of % misclassification is taken to be zero

It is apparent that for a three variate situation, an intersample distance greater than 2.0 results in a high level of correct classification between pairs of samples (> 92%). That is, for 3RAW01 the expected percentage of misclassification with an intersample distance of 2.0 is 7.9. For two samples of twenty individuals each, one would expect only 3 of the 40 animals to be misclassified. From this example, it can be seen that the intersample distances computed from the sample means along the canonical variates give a real and succinct measure of the difference between the data sets for any two samples.

Curves, similar to that obtained from 3RAW01, were available from the other analyses. However, the greater intersample distances and the higher levels of correct classification resulted in fewer non zero points for the %

misclassifications. Therefore, the resulting equations were not as reliable.

Many of the hypotheses of Wright (1949), concerning the role of isolated populations in the evolutionary process, are supported by the results of this study. It has been proposed (Introduction) that the samples used in this study were obtained from populations which must have been at least partially, if not totally reproductively isolated. It has consistently been seen that there was a large amount of variation between the samples, as exhibited by the large amounts of dispersion repeatedly obtained between various combinations of samples, using different types and numbers of characters. This indicates that there was a very large store of variability within the samples and therefore within the species. Underscoring this is the fact that all of the intersample distances from ALLALL (Table 11, Results) indicated that the means of the various samples were significantly different from each other (Section F, Results). It is also interesting to note that the average distance among the freshwater samples was approximately the same as that exhibited by the samples from salt water. One might postulate that as all the Great Lakes samples came from a single transplant from the Green Lake stock (Creaser, 1925), they might have been expected to show less diversity. However, the results suggest that during fifty years (probably considerably less) the populations, resulting

from the single transplant, have achieved approximately the same degree of diversity as the salt water populations, which undoubtedly have been isolated for considerably longer periods. Therefore, the evidence suggests that when the transplanted smelt migrated to a new lake, they in effect spread into new niches with new complements of selective pressures. As a result, adaptive radiation proceeded in response to the pressures of the new environments or as a result of the removal of other pressures which had previously been operative.

A more striking example of the suggested adaptive radiation was seen in the results obtained from the esterase data. In Figure 8(A), the salt water samples formed a tight cluster which was distinct from the freshwater samples which were also more dispersed. This is especially true for the Lake Erie sample. According to Hankinson and Hubbs (1922), all the Great Lakes samples originated from a common stock. From Tables 20 and 21 (Results), both the Lake Superior and Lake Huron samples exhibited approximately the same degree of separation from the Green Lake sample. If there are selective pressures operating on the esterase system, it would indicate that the resultant forces in Lakes Superior and Huron are about the same size compared with those operating in Green Lake. On the other hand, the resultant between Lake Superior and Lake Huron is smaller, indicating that those selective factors which control the

esterase phenotype are more similar between these two lakes than they are to Green Lake.

When the Lake Erie sample was compared to the other freshwater samples, on the basis of the esterase data (Tables 20 and 21, Results), considerably larger differences were found. It is suggested that selective forces operating on the esterase system of Lake Erie smelt are appreciably different in direction and/or size, from those operating in the other freshwater lakes. The differences between the Lake Erie sample and the other freshwater samples were only slightly less than those between the salt water and freshwater samples (Table 20, Results). Therefore, the differences in the effective selective pressures between these two comparisons are of approximately equal size. From "the esterase system's" point of view, the Lake Erie environment is as different from that of the other freshwater lakes as it is from the sea.

A further conclusion can be drawn if one also considers the results from meristic and morphometric data (Table 11 and Figure 6[A] and [B], Results). In these results, the Lake Erie sample does not appear as an atypical freshwater sample, but is central to its cluster. At the same time, the Great Lakes samples appear to be rather uniformly dispersed. A major distinction exists between the salt water and freshwater samples. Selective factors appear to be operational with meristic and morphometric characters.

Between the salt water and freshwater environments, there are greater differences in the size of these factors than in those which are seen within each of the environments separately. This, however, is not necessarily the case with respect to the selection for the esterase pattern.

The required environmental data for each of the sampling areas were simply not available. This unfortunately restricts any attempt at uncovering any direct correlations between the patterns of dispersion and environmental factors. However, a few suggestions can be made in an attempt to rationalize the observed pattern. Amongst the normal samples, for which all types of data were available, a first approximation to the causes of the differences can be made by simply considering those elements which are different between the marine and freshwater environments. These would include both physical and biotic factors. The means of the total vertebrae counts for salt water samples were consistently higher than those for the freshwater groups. Barlow (1961) reported that variations of this type could be expected where marine eggs hatch in a cooler environment. On the other hand, the freshwater fish are not subjected to the effects of either tidal flow or the rich estuarine environment. Therefore, the increased number of gill rakers found in freshwater smelt may reflect a response to either a difference in available food organisms or the behaviour pattern by which they are encountered and captured. The

proposed difference in food supply may also be involved in the differences seen in the esterase system. However, it would be advisable to test these hypotheses before according much weight to them.

An apparent correlation exists between the esterase system of the Lake Erie sample and the high level of pollution (eutrophication) in the western basin of that lake. Again, it would be difficult to extend the results of this observation further. However, it does appear that factors present in the western basin of Lake Erie are producing a marked effect on the esterase system of the resident smelt population. Therefore, amongst the normal samples a response to undetermined selective pressures can be seen in the phenotype.

One of the primary questions in the biology of *O. e. mordax* is whether or not there are two types of smelt, large and small. Many authors have supported this theory including Kendall (1927), Greene (1930), Zilliox and Youngs (1958), Brooks and Deevey (1963) and Legault and Delisle (1968). According to Kendall there are two distinct sizes of smelts in Lake Champlain and some of the New England lakes. These fish also spawn at different times with perhaps a month between the heights of the spawning season of each type. He also presents some circumstantial evidence indicating that although large smelts have become small after being transplanted to a different lake, they

still maintain the large smelt habit of taking baited hooks. This behaviour pattern is uncommon for the small smelt who apparently subsist almost exclusively on plankton. In this regard, Kendall (1927) reports the results of a stomach contents study based on samples ranging over six years and containing smelts from 7/8 inch to 14 inches long. It appears that after passing a length of six inches, the smelt's diet changes from entomostraca to fish, usually small smelt.

Although many New England lakes are known to support populations of smelt, it is very difficult to determine for any lake whether its smelt population originated from "large" or "small" stocks. This difficulty is the result of two factors. It might be difficult to distinguish a well fed "small" smelt from an undernourished "large" smelt. This is especially true if samples from different years are compared, even if they are the same age. Productivity within a single lake may fluctuate from year to year resulting in an increase or decrease in the food supply available. This then would result in corresponding increase or decrease in the size of the smelt. Secondly, different lakes in the same year may differ in productivity, resulting in a variable food supply for the fish of different lakes. This problem could be investigated by sampling a number of lakes over a number of years and analysing the results by doing an analysis of variance

on the data. Ideally, as much environmental data as possible should also be gathered for the lakes. However, laborious work of this type is unknown on smelt.

The time of spawning has also been used as an indication of whether a smelt population consists of the large or small variety, as it has often been reported that the large fish spawn before the small. However, this is also a difficult criterion to use for the comparison of populations from different lakes or rivers. Rupp (1965) has suggested that a complex interaction and summation of various external and internal stimuli result in the release of the spawning drive in the smelt. If this is the case, it can be seen that the various components of this interaction must also vary from lake to lake making direct comparisons between any two lakes very difficult in one year and almost impossible between years. Rupp (1959) has also shown that both the duration and commencement date of the spawning run are quite variable not only between lakes, but also within a lake. The best situation in which to examine the differences between the large and small forms is when both forms live sympatrically. When this occurs, the variations in the influence of food on size and that of external/internal stimuli on spawning time are minimized. This contrasts with smelt population comparisons between lakes. Data of this type (sympatric) are presented in Kendall (1927). The United States

Bureau of Fisheries propagated smelts for a number of years at Green Lake, Maine. Green Lake was reported to contain the two forms and information is available concerning the spawning runs of the two smelt types. Over three consecutive years, (1921-23), on the average, fifty-three days elapsed between the start of the large smelts' run and the end of the small smelts' run. At the same time, there was an average interval of twenty days between the tailing off of the large smelts' run, and the height of the small smelts' run with an average time of twenty-eight days between the peaks. The duration of smelt runs is also quite variable, both between populations and between years, ranging from six and twenty days. Therefore, even if both the large and small forms in Green Lake exhibited spawning runs of the longest duration there would still have been a period of from eight to sixteen days between the two runs when no spawning would have occurred. Similar data are available for Lake Heney, Quebec. In 1967 the large smelt of Lake Heney spawned between March 18 and April 12 in water 20 to 40 feet deep. In the same year, the small form spawned between April 22 and May 10 on the sand and gravel shores of the lake (Delisle, 1969). There can be little doubt of the validity of these data as in the first case the fisheries officers who operated the smelt hatchery were on the site and propagated both large and small smelts, while in the second case, Dr. Delisle personally

made the observations over the two month period.

The main opponent of the "large smelt - small smelt" theory has been Rupp (1959, 1966), who presents a number of arguments which he thinks cast doubt on the existence of the two forms. Rupp (1959) cites the arguments of Greene (1930) which were:

(1) No anatomical basis for separating the two groups was found.

(2) Smelts which display intermediate growth and thus can be assigned to neither race are occasionally found. A further argument (Rupp, 1959) was:

(3) While Mooselookmeguntic Lake, Maine, was reported to contain smelts having two spawning runs, there was no evidence of this in 1955. Then, the run began with smelts averaging 7.0 (17.7 cm) inches total length which decreased to an average total length to 4.8 inches (12 cm) on the last night of the run.

From Table 8 it was seen that for each applicable analysis, the two stunted samples could be correctly separated from the rest, more than 90% of the time, on the basis of one variable. With the exception of one analysis, the two stunted samples were always totally separable from the rest. Therefore, it appears that the first objection to the existence of two forms of smelt, the lack of an anatomical basis for separation, is without grounds. The two forms are separable on the basis of (1) meristic

data (2) morphometric data and (3) ratios of the various body measurements to each other, formed in such a way that size is eliminated as a factor.

Rupp and Redmond (1966) also presented the results of a number of "controlled experimental introductions" of smelts to new lakes, for the purpose of testing "which characteristics of the parent populations might be altered by transfer to new environments and which might remain unaltered." They also suggest that the "existence of two kinds of smelts . . . some living allopatrically and others living sympatrically, depends ultimately on the assumption that smelt growth rates are controlled by genotypic factors which have sufficient penetrance to negate or virtually negate any environmental effects." A summary of Rupp and Redmond's results are seen in Table 24. Of the eight transplants, three were into lakes which had previously been reclaimed with rotenone, to remove the resident smelt populations (and presumably, all other fish as well), while the remaining five lakes did not previously contain smelt populations. The absolute change in growth was made by comparing the actual maximum length of the shorter lived population with the other population under consideration, at the same age. Following Rupp and Redmond, two exaggerated (overstated) hypotheses are possible:

(1) The size of a smelt (phenotype) is determined solely by its genotype.

TABLE 24

SUMMARY OF THE RESULTS OF TRANSPLANT STUDIES

ON *O. EPERLANUS MORDAX* IN MAINE

(RUPP AND REDMOND, 1966)

Donor Lake	Host Lake	Rotenone	Δ Growth Rate		
			Absolute (inch)	Expected if Environmental	If Genetic
	Rowe	-	δ + ve (0.5)	?	0
Branch	Squaw	-	0	?	0
	Basin	+	+ ve (1.0)	?	0
Cold Stream	Coleback	+	≈ 0	- ve (1.6)	0
Little Concord	Shagg	+	+ ve (2.1)	+ ve (2.1)	0
Kingsbury	Lower Tongue	-	≈ 0	?	0
Manson	Long Pond	-	≈ 0	?	0
Burnt Meadow	Colcord	-	≈ 0	?	0

\+ = host lake treated with rotenone

\- = host lake untreated

+ ve = positive

- ve = negative

? = data were not presented

(2) The phenotype of a smelt is determined by environmental factors.

If the first hypothesis was true, the expected difference in size between the donor and host lake would be zero, unless genetic change occurred. If the second hypothesis was true, the expected change would be towards the size of the population that originally inhabited the host lake. For four of the five lakes which were not treated with rotenone, the absolute change in growth was reported to be zero, while for the fifth lake, there was a slight increase in average length (0.5 inch), after three years of growth. Thus, these five transplant experiments are in agreement with the "all genetic" hypothesis. One is unable to make a decision on the degree to which these data fit the "all environmental" hypothesis as there were no original smelt populations in the host lakes with which to compare the transplants.

The three host lakes which were reclaimed with rotenone merit closer examination. At the outset, the smelt transplanted from Branch Lake to Basin Lake do not offer any assistance in deciding between the "all genetic all environmental" hypotheses, as no data are presented on the nature of the original Basin smelt population. Of the two other transplants, only the Little Concord-Shagg transplant give the type of results that would have been expected if the smelt's size is environmentally determined. On the

other hand, the Cold Stream-Coleback transfer did not support the "all environmental" hypothesis. The original Coleback smelt population averaged approximately 4.4 inches while at the same age, Cold Stream smelts averaged about 6.0 inches. If the environmental factors had been all important, the fish transplanted to Coleback Lake should have been about 1.6 inches smaller than were the ones actually observed, after two years growth. Thus, there is at best conflicting evidence for the "all environmental" hypothesis.

A more plausable hypothesis is that the phenotype is the result of the interaction of the genotype with its environment. The results of Rupp's experiments can more satisfactorily be explained in this context.

Thus, the five host lakes which did not previously contain smelts probably presented approximately the same magnitude of environmental pressure on the introduced smelts as that to which they had been subjected in their original lakes. There was some indication of this in the data presented by Rupp (1959) regarding the physical and biological components of the lakes. The results of the Cold Stream-Coleback transplant can be explained if one postulates that the environment of Coleback Lake was naturally deficient in some factors which are required for normal growth. This condition was offset by the removal of predators and competitors by rotenone treatment, resulting in a new environment in Coleback which exerted approximately

the same pressure on the smelt as they experienced in their original Cold Stream Lake.

The smelts of Little Concord Pond were probably the only stunted fish to be involved in Rupp and Redmond's study (10 cm). Thus, the results of the Little Concord-Shagg transplant can be explained if one considers that Shagg had a naturally more favourable environment for smelt which was enhanced by the removal of predators and competitors by rotenone. Apparently, this combination more than offsets the tendency of any "small" genes to produce small fish. The result of these three factors is an increase in the normal size of the fish from the Little Concord gene pool after two years growth in the much more favourable environment. An evaluation of the effects of rearing fish in a laboratory situation where the natural predators and competitors are absent has been reviewed by Barlow (1961).

Therefore, Rupp has failed to show that the existing differences between smelt populations are almost entirely the result of environmental factors and that genetic differences between populations are of minimal significance. What has been shown is that if smelt are placed in a lake which has been reclaimed, they will probably grow to a larger size than that to which they otherwise would have grown. However, if the lake is not reclaimed, evidence for the genetic nature of the size difference is obtained: that is "small" populations transplanted into new (natural

environments remain "small" and "large" forms remain "large" (Kendall, 1927; Brooke and Deevey, 1963).

The second line of "genetic" evidence is that "large" and "small" populations coexist in at least 7 lakes in Maine (Kendall, 1927) and 3 lakes in Quebec (Delisle and Veilleux, 1969). With sympatric populations, the possibility of causative environmental factors is minimized, thus enhancing the likelihood that phenotypic differences are genetically controlled. There is also the opportunity for crossbreeding. However, this option is apparently not exercised to any extent (Kendall, 1927; Delisle, 1969). This indicates that there is some factor (probably genetic) which acts to maintain the separate identities of the two spawning runs. Finally, the fact that ten separate lakes support continuing sympatric forms suggest that the phenomenon is not simply the result of chance and that the critical factor possesses an element of permanence. These attributes are satisfied by the nonrandom continuing action of a genetic factor.

Esterase systems have been employed by many workers involved in the search for "biological tags" with which to identify various populations of organisms. Usually, this pursuit has been linked with an attempt to uncover a genetic model (classically a one locus two codominant allele system). These models can be used to discuss the differences between populations in terms of shifts in the frequencies of the alleles. Shifts in frequency are usually investigated

using the Hardy-Weinburg equilibrium equation (Sick *et al.* 1965; Sick, 1965a, b).

In this study, the bands were examined with a view to employing these methods. However, they were abandoned on the grounds that: (1) with the exception of esterase bands 5 & 6, all other pairs of bands were absent in various individuals, (2) the testing of genetic hypotheses ideally involves controlled breeding experiments which would have been impossible with the existing facilities and are also outside the range of this work, (3) these methods do not necessarily provide the most complete answer to the question of the degree of diversity exhibited between the samples on the basis of the esterase information. The approaches which were used, attempted to include all the information from the esterase patterns to determine the degrees of difference between the samples, on the basis of this system.

All the results of the Chi-Square tests were important in this regard, for different reasons. The overall Chi-Square showed beyond doubt that the hypothesis that all samples showed the same banding patterns was highly improbable ($p < .0001$) and therefore, must be rejected. It was then necessary to test this overall heterogeneity of patterns to determine; (1) if it was a general phenomenon or (2) if its source was more localized between certain pairs of samples. When these questions were tested,

the Chi-Square values between the salt water and the freshwater groups were generally greater than the values obtained between the samples within each group (Table 14). Between the salt water and freshwater smelts, the indication was that a basic difference existed in the esterase patterns.

Other biochemical differences have been noted between the freshwater and salt water forms. Freshwater smelts contain a higher level of thiaminase activity than do the salt water forms. This higher thiaminase activity has resulted in vitamin B_1 deficiency in dolphins which were fed on a diet of freshwater smelt rather than their usual diet of salt water smelt (Delisle, personal communication).

The results of the tests for the independence of the various esterase bands were important for two reasons. Firstly, the use of either of the previous Chi-Square tests is dependent upon the assumption of the independence of occurrence of the individual bands. On the basis of the second Chi-Square test (Equation 6), the assumption of independence seems justified and cannot be rejected. Thus, the results of the tests of overall homogeneity and between pairs homogeneity did not rest on an incorrect assumption.

Secondly, the test has shown the validity of applying this approach to the question of the degree of difference between samples on the basis of esterase data. Even if there were genetic systems in effect which produced phenotypes which would have fit the Hardy-Weinberg equation,

their influence was not great enough to produce a result that was contrary to an assumption of band independence. Thus, even if groups of components of the esterase system were the result of only a few loci, the expression of the esterase patterns appears to be such that it could also be explained by the hypothesis that the expression of each band is individually controlled. In studies on the American and European eel populations, it has been found (Pantelouris *et al.*, 1971) that many of the phenotypes of the biochemical systems could not be explained on the basis of a Hardy-Weinberg equilibrium. However, in this case, significant differences between the populations were demonstrated by running Chi-Square tests on a few selected bands taken in pairs. Therefore, the analytical methods developed in this thesis may find use in the interpretation of data which are not amenable to the standard techniques of population genetics.

The table of the significance levels between samples on the basis of the esterase patterns (Table 15, Results) gives some indication of the factors governing the expression of the esterase patterns. It has been noted that the levels of significance are much lower among the salt water and freshwater samples than between them (Table 14, Results). Also, the Lake Erie sample was highly significantly different from all other samples in the study. The immediate suggestion is that environmental factors play a large role in

determining the esterase phenotype. The marine environment is different from the freshwater environment not only in its chemical and physical properties but also in the biological complement. Thus, the fish living in the two environments are subjected to different stresses and selective forces. Also, the Lake Erie environment, especially that of the western basin from which the sample was taken, is highly eutrophic. This is in contrast to the other lakes from which samples were obtained. The plot of the esterase data along the first two canonical axes (Figure 8[A]) shows the remote position of the Lake Erie sample with respect to the other freshwater samples. Thus, eggs, larvae and juveniles living in Lake Erie would mature under a different set of selective pressures from those acting on fish in the remainder of the Great Lakes sample.

There are other results in Table 14 (Results) which indicate that selective forces produce an effect on the esterase system. The two groups of samples that were captured on the same day in the same region are not significantly different ($p > .100$) in each case, on the basis of the esterase system (Table 15, Results). As both these groups matured under similar conditions, they were probably subjected to rather similar selective forces. Chi-Square analysis shows that the resulting phenotypes of the sexually mature fish of each group do not appear to be significantly different. Similarly, among the Great Lakes samples, the

two more oligotrophic lakes, Huron and Superior, support smelt populations which are not highly significantly different ($p < .100$), especially when compared to the Lake Erie sample.

Another interesting result is the relatively high level of significance between the N.A. (69) and N.A. (70) samples. Although these samples are both marine and from the same sampling area, they differ in their overall esterase pattern. Although the times of the spawning runs of these two differed by almost a month and the second sample was taken slightly later during its run, the differences in esterase pattern between the two samples could still have been the result of different selective pressures during their development. However, it should also be noted that the N.A. (70) sample on the whole shows greater differences from the marine samples taken in 1969 than does the N.A. (69) sample (Table 14, Results). It therefore appears that there is a "year of capture" factor present in the esterase data. It is also interesting that the sample from Grand Bank, although collected in 1970, was more similar to the marine samples taken in 1969 than it was to those taken in 1970. Thus, if the environmental selective pressures hypothesis has any credability, it appears that the conditions, in which the 1970 Grand Bank sample developed, were similar to those of the 1969 samples from the northern and western coasts of Newfoundland. These

relationships are summarised in Table 25(A). Amongst the salt water samples, the average of the between sample Chi-Square values are much larger when the 1969 samples are compared with the 1970 samples, excluding the G.B. sample (43.4). Both the 1969 and 1970 samples (G.B. excluded) show smaller average values (18.4 and 8.9), indicating greater similarity within each group than between them. The closer relationship between the G.B. sample and the 1969 samples (11.5 versus 22.7) is also apparent.

An "area of capture" factor can also be seen from the values in Table 14 (Results). If the between sample Chi-Square value of the Notre Dame Bay group is compared with those of the other salt water samples, the values in Table 25(B) are obtained. Each of the average Chi-Squares within each group (15.4 and 9.3) is considerably smaller than that between the groups (31.9). Again, this indicates that there is less diversity within the esterase phenotypes of each group than those between groups.

The genetic control of esterase systems has been demonstrated many times, whereas the control mechanisms for morphometrics have not been well investigated. However, selection operating on the appropriate genetic complements is undoubtedly involved in the modification of those factors which control these traits. In smelt, there is tremendous potential available to selective agents. An average spawning female smelt contains about 42,000 eggs

TABLE 25

AVERAGE BETWEEN SAMPLES CHI-SQUARE VALUES COMPUTED FROM ESTERASE DATA FOR DIFFERENT GROUPS OF SALT WATER SAMPLES OF SMELT

(A) All 1969 Salt Water Samples, Grand Bank Sample, 1970 Salt Water Samples Excluding Grand Bank

(B) Notre Dame Bay Samples, Other Salt Water Samples

	All 1969 Salt Water Samples	Grand Bank Sample	1970 Salt Water Samples Excluding G.B.
All 1969 Salt Water Samples	18.4		
G.B. Sample	11.5	———	
1970 Salt Water Samples Excluding G.B.	43.4	22.7	8.9

	Notre Dame Bay Samples	Other Salt Water Samples
Notre Dame Bay Samples	15.4	
Other Salt Water Samples	31.9	9.3

(Legault and Delisle, 1968). Of these, on the average, one to two will return as adults (depending on whether a smelt spawns once or twice in its life), in a stable population. Thus, more than 99.99% of the eggs do not give rise to adults. If one percent of the eggs survive to the post hatching stage, this results in 420 larvae per female, (Rupp, 1965; McKenzie, 1947). This number can suffer a 50% reduction (mortality) 8 times and still produce adults for spawning in two to three years time. This reduction, whether in the form of predation, competition for resources or the physical and chemical environmental stresses, results in the survival of between 0.25% and 0.50% of smelt larvae. The small number of eggs which eventually result in spawning adults indicates that the successful individuals have survived many critical tests; some possibly are due to chance (not being eaten) while others may be due to a more desirable genetic complement (possessing digestive enzymes complementary to the available food supply). However, the results of this study have indicated that the factors controlling form and biochemistry are very greatly influenced by the selective forces of the environment, not only from place to place but from year to year.

The patterns of variation exhibited by the esterase system are especially interesting when contrasted with the results from the myogen system. Both between and within samples, the myogens exhibited a pattern which for the most

part remained constant. There was some apparent variation in the "globulin" region of the pattern (myogen bands 3 and 4, Figure 7[A], Results), but the electrophoretic mobilities of these bands were quite sensitive to small changes in gel consistency and pH. Apparent differences between runs had been noted, but when the samples were rerun, the standard pattern was again observed.

The species specificity of myogen patterns has been noted for a number of different species of fish (Cowan, 1968). It was also noted during the present study that the M.D.H. and L.D.H. systems in the smelt were constant between samples. As divergence between populations can only be evaluated by comparing the differences between them, these "constant" biochemical systems are of little use towards this end. However, the knowledge of their existence is important as they provide species specific characters by which the integrity of the group can be demonstrated when compared with other species. These constant systems also serve as a cohesive factor when examining populations of a species. That is, while more than 90% of the stunted forms can be distinguished from all other samples on the basis of one character (Σ G.R.), the inclination to regard these groups as two species is off-set by the between sample constancy seen in the biochemical systems. It should be noted that the esterase system also showed the same bands in all groups despite the fact that their frequencies

of occurrence between groups were variable. Therefore, the shifting of the esterase frequencies between samples indicates different groups, whereas, the presence of the same esterase bands in all samples reaffirms the integrity of the unit.

These different systems in the smelt, are probably controlled by different genes. The variability of some systems of a species and the constancy of others is a very interesting phenomenon. Why should some systems such as L.D.H., M.D.H. and myogens exhibit a constant pattern from organism to organism while the meristic, morphometric and esterase characters vary? The nexus hypothesis states that every character is likely to be affected by more than one gene. This means that for these steady systems, there are probably many genes which are constant throughout all individuals of a species. These constant systems seem to have "crystallized" in the species. This suggests that the individuals contain uniform genetic complements for the control of these crystallized systems.

Conversely, the other variables examined show a great deal of plasticity in their expression. These characters are under the control of systems which allow for much variation. Many of the esterase bands occurring in the house mouse (*Mus musculus*) have been studied and the genetic control of these bands has been determined. Therefore, the variation in the occurrence of specific

patterns between samples probably reflects the variation in the genetic complements controlling these systems. Considerably less work has been done on the much more difficult topic of the inheritance of body form. It should be noted that all the members of a smelt population inhabit the same area, probably in a loose school. As a result, any differences in form between members of a population, are probably not due to markedly different environmental stresses. This would tend to indicate that differences in form between members of a smelt population are probably due to differences in the genetic complement between the various individuals.

Kraus and Choi (1958), working on the human foetus were able to extract four patterns of growth from a principal component analysis based on twelve skeletal measurements. They were able to obtain teratological specimens in which the failure of a gene, by mutation, resulted in the suppression of a pattern of growth. They were thus able to speculate that each of the observed patterns of growth was controlled by a single gene. This suggests that the complex form of the individual may be under the control of fewer genes than might otherwise have been expected.

In a canonical analysis, the number of axes of variation is usually limited by the number of samples examined, that is, the number of axes of dispersion is one less than the number of samples considered. In all

the analyses which contained either raw or transformed morphometric and meristic data, the major distinction was between the two stunted and the remaining normal samples. It is interesting to speculate on the relevance of the work of Kraus and Choi to this study. If the differences between large and small smelt are the result of a single gene, it would be very helpful in explaining many of the problems associated with these two groups. There is conflicting evidence on the results of transferring smelt from stunted populations to new lakes. Apparently they remain stunted in some host lakes while in others they eventually attain a more normal size. If each of the many characters which separate the two groups was controlled by a separate gene, a massive amount of selection would be involved in order to produce normal smelts from the eggs of the stunted forms. However, if there was a single primary controlling factor, which governed the expression of many characters, the number of sites which would be affected by selection would be greatly reduced. Also, the degree of required heterozygosity would be reduced, for there would be no need for the population to be large-small heterozygous at all the loci involved. It might be argued that the difference between the normal and stunted forms must be the result of the action of a single locus. The transfers of smelt from one lake to another are often successful. If many loci were involved in a transformation of form, very many individuals

would be selected against thereby greatly reducing the number of individuals that could survive in the new environment. This would greatly diminish the chances of success.

A much simpler model would be one in which a whole group of genes was common to all members of the species (homozygous). The expression of these genes would be under the control of a few pleiotropic controller genes which could be thought of as heterozygous and available to be acted upon by selection.

The nature, and for that matter, the existence of the controller genes is unknown. However, there is evidence from multivariate methods that this type of situation exists. This model is also of assistance in the formulation of an understanding of how the large and small smelts can exist separately and at the same time maintain the potential for conversion of their form. Pleiotropic controller genes interacting with environmental stresses produce the resulting variation in the morphological traits.

Slobodkin (1966) has noted that certain morphological traits are highly variable in response to environmental change, while others are not. The difference between these kinds of traits is functionally related to the pattern of fluctuation of the environmental variables. If we now imagine an environmental change which would transform what was a rapidly fluctuating feature of the environment to a relatively stable one, it would be expected that in a fairly

short time, a change in the size of the variation between the genotype and the phenotype of organisms in that environment would occur. It would also be expected that some characteristics which had apparently been under physiological or environmental control would now appear to be under tight genetic control, manifested by a reduction in the variation of characters. This change would look very much like "genetic assimilation."

When the small and large forms live sympatrically, they probably act as mutual selective factors at different stages of their respective life cycles. Adult small smelt would be in competition with the juveniles of the large form. On the other hand, adult small smelt would become the prey of the large form if they strayed into the latter's environment. Thus, it would be advantageous for the large form to quickly pass through the stage in which they would be in competition with the small form and for the small form to maintain a niche distinct from that of the large form. This then would have the effect of reducing the number of environmental variables to which each group is subjected as the interactions between the sympatric forms would result in more precise niche specialisation. This would probably be more pronounced with the small form, as it would be preyed upon by the larger form for a longer period of time. Evidence in support of this is seen in Table 9. The sympatric stunted sample (L.H.S.) forms a

considerably tighter cluster (small σ_I & σ_{II}) when plotted along canonical axes I and II, indicating a greater uniformity in form and meristics than was seen with the other (non-sympatric) populations. Therefore, sympatric populations, subjected to less fluctuation in the environmental variables, would face greater selective pressures to maintain one special territory and the resulting differences between them would have a more genetic appearance. In this regard it is interesting to note that the L.H.S. and L.H.G. samples were 20.47 units apart in the generalised distance matrix (Table 11, Results) from ALLALL. This was the second greatest distance that this stunted sample had with the other samples in this study. The greatest distance (20.63) was with N.A. (70) and the third greatest distance (20.14) was with N.A. (69). Therefore, only one salt water sample resulted in a greater distance with L.H.S. than did its own sympatric population.

By the use of canonical variates, one is able to determine the number of different patterns of development which are represented by distinct contrasts in form of the smelt samples. Secondly, it is necessary to determine the biological significance of these contrasts in form. If two related forms differ in shape in the adult stage, they must have followed different patterns of development (Blackith and Blackith, 1969).

The first two canonical axes have been shown to

separate the stunted from normal and the freshwater from salt water forms, respectively. The importance and function of the third canonical variate alone is not as clear but from the three dimensional model (Figure 4[C], Results), it became apparent that a combination of the information contained in the first, second and third variates separate the normal samples on the basis of length. This trend is seen for the Great Lakes samples. In order of increasing mean length these were L.H., L.S. and the L.E. sample. The respective values along the third canonical variate were 2.22, 0.91 and -0.44. Therefore, the third variate is also size related.

Among the salt water groups, the fourth and fifth variates appear to separate the samples upon the basis of their year of capture. The only exception to this is the C.B. sample along the fourth variate. However, this relationship holds true for all the other salt water samples over these two variates. Among the freshwater samples, the fourth variate disperses the group means in a manner which reflects the south to north geographical position of the samples. Both the fourth and fifth variates indicate the same distinctions among the freshwater groups, in that the L.S. sample is distinguishable from the other two on the basis of these variates. This same distinction can also be seen in the values of the group means along the second canonical axis (Figure 3).

The two stunted populations remain quite close

together on the basis of the first three canonical variates. They become somewhat more separated on the basis of the fourth and fifth variate. By far the greatest separation between these two samples is brought about by means of their positions along the sixth canonical variates. With the inclusion of six canonical axes, 96.2% of all variation has been accounted for. It is therefore doubtful if the remaining six variates contribute much value to an understanding of the patterns of variation in the samples.* It has been noted that in any year, a population will show a gradual decrease in size over its spawning run (McKenzie, 1964; Rupp, 1959). With this in mind, an attempt was made to obtain the samples from the early part of each run in an attempt to make the samples more comparable. Of the samples for which the commencement dates of the runs were known, the samples were taken between the second and the seventh day. As smelt runs may last upwards of three weeks, these samples can be considered to be taken from the first part of the run.

Studies of populations can be of three types. Single samples from a number of populations can be compared. Secondly, a number of samples can be taken from a single population over some period of time which is of interest (e.g. day, spawning run or year). Thirdly, a single population may be repeatedly sampled over a number of years to assess any changes that might occur over this greater period of time. A study of the second type is underway on smelt (Nhwami,

---

*See page 234.

unpublished data). The present study is primarily of the first type, although samples were taken from one area, over two successive years (N.A. [69] and N.A. [70]).

These two samples are of special interest in that they present an example of the type and magnitude of the variation that exists between the spawning individuals of a population on successive years. From the Minimum Spawning Trees of the raw morphometric and meristic data from ALLALL (Figure 6[A], Results), the esterase data from ESTALL (Figure 8[A], Results), and the transformed and meristic data from SZOTME (Figure 10[A], Results), the N.A. (69) and N.A. (70) samples are connected through the I.R. sample which comes from the same geographic region. Therefore, on the basis of any of the types of data, the N.A. (69) sample more closely resembles a Notre Dame Bay sample than any other sample in the study. However, the N.A. (69) sample has been shown to be somewhat abnormal and removed from the central salt water cluster on the basis of analyses containing either raw or transformed morphometric data (Figure 6[B] and Figure 10[B], Results). It was taken to be an extreme expression of the marine form.

The size of the variation between these two samples can be seen from the generalised distance matrix resulting from the ALLALL analysis (Table 11, Results). The N.A. (69) and N.A. (70) samples are 6.05 units apart. When t-tests were run on the samples' means for the different

variables their level of significance was assessed. Of the twenty-two morphometric measurements, twenty-one were highly significant ($p < 0.001$) while one (L.Or.) was significant at a level greater than 0.100 (0.104). On the other hand the differences in the means of all the meristic counts were not significant ($p > 0.100$). It was also noted that the N.A. (69) and N.A. (70) samples were 6.06 units apart in the SZOTME analysis. Therefore, a similar level of significance was expected between the samples on the basis of the transformed data. When these t-tests were performed, it was found that only six of the twenty-one transgenerated variables were not significant ($p > 0.100$). On the other hand, five were significant ($0.050 > p > 0.001$), while ten were highly significant ($p < 0.001$). Therefore, only six of these variables can be considered as truly size dependent, while the remaining fifteen continue to exemplify the differences between the samples. The results of both these sets of t-tests are probably not as remarkable as they may first appear. The N.A. (69) sample was generally composed of larger fish than was the N.A. (70) sample. As a result, all the comparisons of the raw measurements contained a significant size difference and therefore, there was a great deal of redundancy in the data. However, the significance levels of the transformed variables were also high between the two populations. The use of ratios and similar devices has become discredited because the ratios

change in magnitude when larger individuals of a given species are compared with smaller ones by virtue of the almost universal occurrence of allometric growth (Blackith and Reyment, 1971). What actually occurs, when a ratio is used, can be seen by an examination of Figure 1 (Results). The conversion of a measurement to a ratio will decrease the absolute distance (d) between the two means. However, the standard deviation (σ) will also be decreased. The result of this is that the overlap of the two populations (shaded area) will not be effectively changed. Therefore, the variation in the data will exhibit its' affect whether or not ratios are used.

When the age determinations were done, it was noted that the scale growth interval corresponding to the year of 1967 was larger than normal. This was especially apparent in the predominant (59%) four year old fish from the N.A. (69) sample, in which this interval was seen as the second band of growth. From these qualitative observations, the indication was that during the year of 1967, greater than normal growth occurred. This may partially account for the greater size of the N.A. (69) sample as this year would only have been the first year of growth for the four year old fish in the N.A. (70) sample (3% of the sample).

The predominant (86%) three year olds of N.A. (70) did not experience this year of exceptional growth. These observations indicate that a degree of caution should be

exercised when using growth rates to characterize populations. The greater than normal growth in the second year of the N.A. (69) sample would increase the slope of the growth curve and thus give an abnormal result. Care should also be taken, in this regard, if one is involved in allometric growth studies on a species. For example, suppose a sample contained individuals who were 2, 3 and 4 years old. Suppose also that three years previously, there had been either an exceptionally good or bad growth period. This would mean that the four year olds had been exposed to the exceptional year during their second year of growth, while the three year olds would have been exposed during their first year. The two year olds would not have been exposed to the exceptional year.

Therefore, the different years could not be directly compared as representatives of the various year classes. The three and four year olds were affected by the exceptional year during different periods of their development and thus the resulting effects would be different in each group. The two year olds are probably the only group that is representative of the "normal" pattern of development while the other two year classes represent experimental groups which have been exposed to an abnormal environment, each at a different stage in its development.

Evidence supporting this hypothesis has already been seen. The means of the morphometric variables for the two

Norris Arm (N.A.) samples are, for the most part significantly different. This is also true when ratios are used in an attempt to standardise the data. The N.A. (69) sample was composed predominately of four year olds who had experienced an exceptionally good growth period during their second year while the N.A. (70) sample, composed primarily of three year olds had not been exposed to this exceptional year. Apparently, none of the other samples experienced any growth period which produced as marked an effect as in the N.A. (69) sample. Therefore, in the dendrograms representing the degree of dispersion amongst the populations, on the basis of either raw or transformed variables, the N.A. (69) sample is somewhat remote from the main body of the cluster of samples (Figures 5[B], 6[B] and 10[B], Results). Therefore, part of the atypical appearance of the N.A. (69) sample may be explained by the fact that most individuals in that sample (> 90%) were exposed to an exceptionally good period of growth. Corroborating evidence for this hypothesis is seen in the position of the I.R. sample (intermediate to N.A. [69] and N.A. [70]) both in respect to its age distribution and its position resulting from canonical analysis (Figures 6[A], 8[A] and 10[A], Results). Comparisons of the age distribution patterns (Table 3, Results) and the linkages in the M.S.T.'s, did not generally show an association for the other samples, and the presented hypothesis is more plausable.

Therefore, evidence has been presented which indicates that the attributes of a species, through the populations of which it is composed, are dynamic, not only as a function of place and year, but also as a result of the local environmental histories of the populations. These would exhibit their affects in the production of varying patterns of growth in the different populations. The form of an individual is not just controled by the interaction of a genotype with its environment, but also by the pattern of environmental fluctuation with respect to the state of development of the phenotype in which the genotype is found.

The implications of this position could be of considerable importance to the problem of population (stock) identification. It has been seen, that members of populations can be identified with a high degree of accuracy, once the parameters of the populations have been assessed. It has also been seen that the value of a population's parameters may change over time. Therefore, if these changes over time can be evaluated and the patterns of change determined, it should be possible to identify unknown individuals with a high degree of success, at some future time. It would be highly desirable to be able to relate the patterns of change to external (environmental) changes which tend to be more accessable to measurement and might permit prediction regarding the future form of the individuals.

On two occasions, data sets were combined in an

attempt to present a more complete analysis of the relationships between the samples. Section L (Results), involves the combination of information from the raw morphometric and meristic characters and the data from the esterase investigations. The results of combining the supposed size-free transformed morphometric and meristic variables with the esterase data were seen in Section N (Results). In both these cases, the data sets were combined by solving Equation 11 (Section L, Results).

The use of Equation 11 neglects any possible correlations between the meristic and morphometric variables and the esterase variables. A second approach to combining the information from two different data sets taken from the same samples is rotational analysis, as suggested by Gower (1971). The analyses ALLALL and ESTALL have resulted in the production of two sets of points, P_i and Q_i in n-dimensional space. In rotational analysis, the set of points Q_i is moved relative to the set P_i, by means of translation and rotation, so that the two sets of points P_i and Q_i now show the best fit. The best fit is attained when the 'residual' sum of squares, $R^2 = \Sigma_{i=1}^{n} \Delta^2 (P_i Q_i)$, is minimum. [$n = \min (n_p, n_q, k_{p-1}, k_{q-1})$]. Therefore, after rotation, the set of points Q_i have new positions in space and for each point, a new set of coordinates has been generated.

The coordinates (canonical means) from ESTALL were

rotated with respect to the coordinates from ALLALL. In the analysis reported here, the data were not scaled on the grounds that ALLALL had sum of latent roots of 64.17 while ESTALL's sum was 2.66. It was reasoned that scaling the data would have had the effect of assigning undue weight to the esterase data. This did not seem justified.

The plot of the rotated coordinates of ESTALL along the first two axes of variation produced results that were similar to those seen in Figure 8(A). The principal difference was that the entire plot had been rotated approximately 45° counter clockwise so that the L.E. sample was above the body of the salt water cluster. The inter-sample distance matrix computed from the new set of co-ordinates were identical to that obtained from the original coordinates of ESTALL. As a result, the combination of this distance matrix with that obtained from ALLALL (Table 11, Results) would have produced results identical to those of Figure 9(B), (Results).

The rotation of the data from the ESTALL analysis had little effect in the final result as shown by this approach. However, further work is intended in the area of combination of different data sets. Scaled and unscaled data from different analyses will be combined and Principal Coordinate analysis will be used to resolve the resulting combined sets of coordinates. The ensuing results should prove beneficial in developing proper procedures for the

combination of different data sets taken from the same or similar samples.

SUMMARY

In this thesis, an attempt has been made to test the hypothesis that smelt populations are different from each other. This hypothesis follows from the fact that the populations tend to be isolated and therefore can be subjected to different sets of environmental (selective) pressures.

Thirteen samples were taken from twelve geographical areas over two years. Seven samples were from marine populations and six were from freshwater areas. Of the freshwater group, two samples appeared to be of the stunted form (Kendall, 1927, and Delisle, 1969). Therefore, the study contained representatives from three possible subdivisions of the species.

From twenty of the individuals in each of the thirteen samples, twenty morphometric characters and thirteen meristic counts were recorded. In the eleven samples composed of fresh frozen material, electrophoretic esterase and muscle myogen data were recorded for at least 96 individuals in each sample. Therefore, the data set was both large and varied in origin.

The first hypothesis to be tested was that all samples were homogeneous, on the basis of their meristic

and morphometric data. This was found to be highly improbably ($p << .001$) and therefore the samples were taken to be heterogeneous.

Given heterogeneity amongst the samples, the next questions posed was whether all samples were equally different or whether certain samples or groups of samples were more distinct. Canonical analysis was applied to this problem and the samples were discriminated as clearly as possible by a reduced set or coordinates. On the basis of the meristic and morphometric data, the pattern of dispersion seen in Figure 3 (Results) was obtained. The two stunted samples were very distinct and the samples from the marine and freshwater environments appeared to form separate clusters as seen in two dimensions. Therefore, on the basis of 85% of information in the data (Table 4, Results) a pattern in the variation amongst the samples was emerging.

If all the information in the data was considered, could further relations between the samples be uncovered and organised? To answer this, U.P.G.M. cluster analyses and M.S.T. analyses were performed on the intersample distance matrices computed from both the first three and all (12) coordinates resulting from the canonical analysis (ALLALL). In Figure 6(B), (Results), it was seen that amongst the salt water samples, there were natural clusters. That is, those samples which were geographically and temporally similar, clustered first and all the samples taken in 1970

clustered before joining with any other marine smelt. Again, the marine, freshwater and stunted smelt formed separate clusters. Therefore, on the basis of all the available information contained in the morphometric and meristic data, the samples formed natural groups.

As distinctions had been implied between the samples, it was questioned whether these were real and justified. This was answered from both theoretical and empirical considerations. When only three dimensions were employed, it was found that the 90% confidence regions for the sample means only overlapped with the two stunted samples. However, when all the variation amongst the samples was accounted for (12 dimensions) there was no overlap of the samples' means even at the 99% confidence region. Empirically, things are distinct if they can be discriminated or if new or unknown objects can be correctly assigned to their group. When discriminant function analysis was performed, it was found that on the average, 93.4% of the individuals could be assigned to their correct groups. With the marine - freshwater distinction, 99.2% of the individuals were correctly identified while the stunted samples were totally separable. Therefore, the implied distinctions between the samples were real.

An extended series of analyses were run to answer questions which arose from the primary analysis (ALLALL). The questions were:

1. Are certain "key" variables involved in the separation and identification of the samples?

2. Do morphometric and meristic data differ in their identification ability?

3. What are the most important variables involved in the separation and identification of the marine and freshwater forms, the marine forms only and the freshwater forms only?

4. What is the effect of removing the two stunted forms from the analysis?

5. If transformed data (Table 2, Results) are used instead of raw data, is there an effect on either the results from the cluster analysis or the levels of correct identification?

From the analyses it was found that:

1. Although the removal of certain variables reduces the total dispersion, the samples are still highly separable and the major patterns of association are maintained

2. Meristic characters are more important in distinguishing between marine and freshwater forms while morphometric characters are important in the identification of individual samples.

3. The three most important characters in the marine - salt water distinctions are meristic while various morphometric characters, especially of the head region are important in distinguishing members of the separate groups.

4. The removal of the stunted samples resulted in a dispersion pattern similar to that obtained from axes II - XII in ALLALL. This indicated that axis I (ALLALL) is primarily concerned with the stunted - normal distinction.

5. The use of transformed data, with or without meristics, produces results which are very similar to those obtained from raw data, both in the pattern of dispersion and the levels to which correct identification can be made.

On the biochemical level, the L.D.H., M.D.H. and muscle myogen patterns, as revealed by electrophoretic techniques, were constant between individuals of the various groups. Therefore, the integrity of the groups was maintained. The esterase system exhibited considerable variation both within and between samples. This variation was investigated through Chi-Square tests and canonical analysis with cluster analysis and discriminant analysis following. It was found that the esterase bands appeared to be independent in their expression. The patterns between samples were highly significant overall, indicating that the esterase patterns vary from group to group. It was also found that the level of significance between the marine and freshwater groups was much greater than within each group. The overall esterase pattern from the Lake Erie smelt was very different from all other samples in the study.

The canonical analysis (Figure 8[A], Results) indicated that there were three groups: the compact marine

group, the more diffuse freshwater group and the Lake Erie group. From the cluster analysis (Figure 8[B], Results), it was seen that the difference between the marine and freshwater groups were about twice as large as the differences within each separate group. However, the difference between the marine and freshwater groups was less than the differences between the freshwater group and the Lake Erie sample. Translated into terms of environmental selective pressure, the Lake Erie environment is as distinct from the 'normal' freshwater environment as the freshwater environment is from the marine, if the assumption that environmental factors have an effect on the esterase phenotype is correct.

On two occasions, an attempt was made to present results based on the total information available. That is a combination of the esterase information with that from either raw or transformed morphometrics and meristics (Figures 9 and 11, Results). In both cases, three main clusters were seen (marine, freshwater and stunted samples). Also, geographically and temporally similar samples clustered first. The marine samples from 1970 clustered before joining with any other salt water sample.

It has been suggested that when many characters are recorded from a number of similar objects, a considerable amount of redundancy will exist in the data set. This proposition was tested by running three analyses, each with information from only three characters. The

major relationships seen with the large suite of characters were also apparent with the reduced set. Therefore, redundancy in the data was clearly demonstrated. Also, a model was advanced to explain why a maximal dispersion does not necessarily result in the highest level of correct identification. It was also shown that an appreciation of the distances generated from multivariate analyses permit a concise measure of the real differences between sample which otherwise could only be investigated by a host of univariate tests.

A primary question in the biology of smelt concerns the supposed existence of two forms, one normal or large the other small or stunted. Evidence from both sides was examined and the principle opposing evidence was found wanting. Several anatomical characters can be used to separate the two groups including 'size free' morphological ratios. A mechanism is proposed, on the basis of results from multivariate analyses, to explain how it may be possible for the two types to change form on the rare occasion. Also a mechanism is suggested whereby sympatric large and small forms may act as mutual selective agents for the maintenance of both large and small forms. Evidence in support of these mechanisms was seen in the results.

Samples were taken from one geographical area (Norris Arm) on two successive years. The two samples were

distinct. This indicated that spawning populations of this species can vary markedly over a short period of time. However, both of these samples were more similar to the other sample from the same geographical area than the remaining samples, indicating that general regional similarities may exist over time.

This thesis has shown that the populations of smelt are distinct at a point in time and therefore the species itself must have wide limits. It has been shown that similarities exist between the populations on the basis of area of capture, year of capture, region of capture and environment. Therefore, the populations are both dynamic and characteristic. This has important implications to the practical problems of population (stock) management which presupposes the ability to identify the population to be managed. The development of this type of work should be continued, for use in those situations where refined answers are desirable, especially where questions have economic significance or are important to more general problems in population biology (Ehrlich and Ehrlich, 1966). There is also evidence suggesting that complex and elusive alterations in form, resulting either from selection or subtle modifications to changes in environmental parameters, can be revealed and possibly correlated with causative factors, in properly designed experiments (Hurnik *et al.*, 1973).

BIBLIOGRAPHY

Anderson, T. W. 1966. An introduction to multivariate statistical analysis. John Wiley and Sons Inc., New York.

Bartlett, M. S. 1947. Multivariate analysis. J. Roy. Statist. Soc. B, 9: 176-197.

Barlow, G. W. 1961. Causes and significance of morphological variation in fishes. Syst. Zool., 10.

Berg, L. S. 1948. Freshwater fishes of the U.S.S.R. and adjacent countries. Vol. I. Israel Program for Scientific Translations, Jerusalem.

Bigelow, H. B. and W. C. Schroeder. 1963. Fishes of the Western North Atlantic. Part III. Memoir One, Sears Foundation for Marine Research. pp. 553-572.

_____ 1953. Fishes of the Gulf of Maine. Fishery Bulletin of the Fish and Wildlife Service, Bull. #74, Vol. 53. pp. 135-139.

Blackith R. E. and R. M. Blackith. 1969. Variation of shape and of discrete anatomical characters in the morabine grasshoppers. Aust. J. Zool., 17: 697-718.

Blackith, R. E. and R. A. Reyment. 1971. Multivariate morphometrics. Academic Press, New York.

Blackwelder, R. E. 1964. Phyletic and phenetic versus omnispective classification. In Phenetic and phylogenetic classification. Edited by V. H. Heywood and J. McNeil. Systematics Assn. Publ. No. 6, London. pp. 17-28.

Brooks, J. L. and E. S. Deevey, Jr. 1963. New England. In Limnology in North America. Edited by David G. Fred. Univ. Wisconsin Press, Madison, Wis.

Calaprice, J. R. 1971. X-Ray spectrometric and multivariate analysis of sockeye salmon (Oncorhynchus nerka) from different geographic regions. J. Fish. Res. Bd. Canada, 28: 369-377.

C.R.C. Handbook of tables for probability and statistics. 2nd edition. Edited by W. H. Beyer. Chemical Rubber Co., Cleveland.

Cowan, G. I. McT. 1968. Comparative analysis of separate data sources in a systematic study of the genus Myoxocephalus (Pisces: Cottidae). Ph.D. Thesis, University of British Columbia, Vancouver.

Creaser, C. W. 1925. The establishment of the Atlantic smelt in the upper waters of the Great Lakes. Papers Michigan Academy Science, Arts and Letters, 5: 405-424.

Delany, M. J. and M. J. R. Healy. 1964. Variation in the long-tailed field-mouse (Apodemus sylvaticus (L.)) in north-west Scotland. II. Simultaneous examination of all characters. Proc. Roy. Soc. B, 161: 200-207.

_____ 1966. Variation in the white-toothed shrews (Crocidura spp.) in the British Isles. Proc. Roy. Soc. B, 164: 63-74.

Delisle, C. 1969. Écologie croissance et comportment de l'Éperlan du Lac Heney, Comté de Gatineau ainsi que la répartition en eau douce au Québec. Ph.D. Thesis, University of Ottawa, Ontario.

Delisle, C. and C. Veilleux. 1969. Répartition géographique de l'Eperlan arc-en-ceil Osmerus eperlanus mordax et et de Glugea hertwigi (Sporozoa: Microsporidia) en eau douce, au Québec. Naturaliste Can., 96: 337-358.

Ehrlich, P. R. and A. H. Ehrlich. 1966. The phenetic relationships of the butterflies. I. Adult taxonomy and the nonspecificity hypothesis. Syst. Zool., 16: 301-317.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. Lond., 7: 179-188.

Greene, C. W. 1930. The smelts of Lake Champlain. In a biological survey of the Champlain Watershed. A supplement to the 19th Ann. Rept., N.Y. Conservation Dept. pp. 105-129.

Gower, J. C. and G. J. S. Ross. 1969. Minimum spanning trees and single linkage cluster analysis. J. Roy. Stat. Soc. (C). Vol. 18, #1.

Gower, J. C. 1971. Statistical methods of comparing different multivariate analyses of the same data. In Mathematics in the archaeological and historical sciences. Edited by F. R. Hodson, D. G. Kendall and P. Tautu. Edinburgh University Press. pp. 138-149.

Hankinson, T. L. and C. L. Hubbs. 1922. The establishment of the smelt in Great Lakes waters. Copeia, No. 109.

Hotelling, H. O. 1951. A generalized T-test and measure of multivariate dispersion. 2nd Berkeley Symp. Math. Stat. Prob. California Univ. Press.

Hubbs, C. L. and K. F. Lagler. 1947. Fishes of the Great Lakes region. Cranbrook Institute of Science, Bulletin #26.

Hurnik, J. F., E. D. Bailey and F. N. Jerome. 1973. Divergent selection for retrieving behavior in mice. Can. J. Zool., 51: 21-26.

Kendall, M. G. and A. Stuart. 1966. The advanced theory of statistics. Vol. III. Griffin, London.

Kendall, W. C. 1927. The smelts. Bull. U.S. Bur. Fish., 42: 217-375.

Kraus, B. S. and S. C. Choi. 1958. A factorial analysis of the prenatal growth of the human skeleton. Growth, 22: 231-242.

Legault, R. and C. Delisle. 1968. La fraye d'une population d'eperlans géants, Osmerus eperlanus mordax, au Lac Heney, Comté de Gatineau, Québec. J. Fish. Res. Bd. Can., 25: 1813-1830.

Mitchill, Samuel L. 1814. Report, in part, of Samuel L. Mitchell, M.D., Professor of Natural History, etc. on the fishes of New York, New York. pp. 1-28.

McAllister, D. E. 1963. A revision of the smelt family, Osmeridae. National Museum of Canada Bull. No. 191, Biological Series No. 91.

McKenzie, R. A. 1947. The effect of crowding smelt eggs on the production of larvae. Fish Res. Bd. Canada, Atlantic Prog. Rept. #39, pp. 11-13.

_____ 1958. Age and growth of smelt, Osmerus mordax (Mitchill), of the Miramichi River, New Brunswick. Fish Res. Bd. Canada, 15(6): 1313-1327.

McKenzie, R. A. 1964. Smelt life history and fishery in the Miramichi, New Brunswick. Fish Res. Bd. Canada, Bulletin #144.

Pantelouris, E. M., A. Arnason and F. W. Tesch. 1971. Genetic variation in the eel. III. Comparisons of Rhode Island and Icelandic populations. Implications for the Atlantic eel problem. Marine Biology, 9: 242-249.

Rao, C. R. 1952. Advanced statistical methods in biometric research. John Wiley and Sons, New York.

Ross, G. J. S. 1969a. Minimum spanning tree. J. Roy. Stat. Soc. (C). Vol. 18, #1. pp. 103-104.

_____ 1969b. Printing the minimum spanning tree. J. Roy. Stat. Soc. (C). Vol. 18, #1. pp. 105-106.

_____ 1969c. Single linkage cluster analysis. J. Roy. Stat. Soc. (C). Vol. 18, #1. pp. 106-110.

Rupp, R. S. 1959. Variation in the life history of the American smelt in inland waters of Maine. Trans. Am. Fish. Soc., 88: 241-252.

_____ 1965. Shore-spawning and survival of eggs of the American smelt. Trans. Am. Fish. Soc., 94(2): 160-168.

Rupp, R. S. and Malcolm A. Redmond. 1966. Transfer studies of ecologic and genetic variation in the American smelt. Ecology, 47(2): 253-259.

Sawyer, W. W. 1966. Prelude to mathematics. Penguin Books Ltd., Harmondsworth, England.

Seal, Hilary L. 1964. Multivariate statistical analysis for biologists. Metheun, London.

Sick, K., O. Fpydenberg, D. Moller and G. Naevdal. 1965. Haemoglobin polymorphism in Norwegian cod population. Hereditas, 53: 257-275.

Sick, K. 1965a. Haemoglobin polymorphism of cod in the Baltic Danish belt sea. Hereditas, 54: 19-48.

_____ 1965b. Haemoglobin polymorphism of cod in North Atlantic Ocean and North Sea. Hereditas, 54: 49-73.

Slobodkin, L. B. 1966. The light and the way in evolution. Q. Review of Biology, 41(2): 191-194.

Sneath, P. H. 1957. Computers in taxonomy. J. gen. Microbiol., 17: 201-226.

Sokal, R. R. and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., 38: 1409-1438.

Sokal, R. R. and F. J. Rohlf. 1969. Biometry: the principles and practice of statistics in biological research. W. H. Freeman and Co., San Francisco.

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

Templeman, W., H. J. Squires and A. M. Fleming. 1957. Nematodes in fillets of cod and other fishes in Newfoundland and neighbouring areas. J. Fish. Res. Bd. Canada, Vol. 14, #6. p. 831.

Templeman, W. 1966. Marine resources of Newfoundland. Bull. Fish. Res. Bd. Canada, #154. pp. 108-111.

Tsuyuki, H., E. Roberts, R. H. Kerr and A. P. Ronald. 1966. Micro starch gel electrophoresis. J. Fish. Res. Bd. Canada, Vol. 23, #6. pp. 929-933.

Wright, S. 1949. Adaptation and selection. In Genetics Paleontology and Evolution. Edited by G. L. Jaspen, E. Mayr and G. G. Simpson. Atheneum, New York.

Zilliox, R. G. and W. D. Youngs. 1959. Further studies on the smelts of Lake Champlain. N.Y. Fish and Game Journal, Vol. 5, #2.

ADDENDUM

The following should appear after the places indicated in the text and serve as further explanation for the points being discussed.

Page 30

The word "average" is used throughout the text and refers to the "arithmetic mean."

Page 36

Canonical analysis rests on the assumption that all the multivariate normal universes have the same variance-covariance matrix. A test for this proposition is given by Seal (1964). When this test was performed with the raw data, the resulting Chi-Square was 358 with 6732 degrees of freedom. The value of Chi-Square at $p = 0.10$ is 6924 for 6732 degrees of freedom. As the obtained Chi-Square was less than the value for $p = 0.10$, there was no reason to reject the null hypothesis that the variance-covariance matrices were equal. Therefore, the principal assumption was verified.

Page 46

As the assumptions (Seal, 1964) underlying the use of discriminant functions did not appear to be violated, there appeared to be no advantage in the use of a distribution

free method for identification. A second important point in this regard was that a computer programme for this type of approach was unknown. Although very little calculation would be involved, a considerable amount of energy would be consumed in the ranking of variables for each comparison. As each comparison might take a week to perform and as there would be 78 comparisons to make, it was questionable whether the gain in correctness would offset the loss in time. The strength of this position was also related to the robustness of the parametric operations. Sokal (1963) commented that the various techniques of numerical taxonomy were statistically robust.

Page 137

Three different values were used as estimates for the missing distance components from the esterase data for the L.H.S. and L.H.G. samples. These were the mean within group distance and the mean between group distance (Table 21) and the overall mean, obtained from Table 20. When the maximum value (3.053) was used, a single change in the linkages in the dendrogram occurred (Fig. 9 [B]). The marine and freshwater clusters joined before L.H.G. was added. This does not require any change in interpretation if one is also aware of the information in Fig. 9 (A). However, with estimated distances equal to or less than the mean distance, the linkages were as in Fig. 9 (B). The crucial distance

in this change is that between the L.H.G. sample and the freshwater cluster. As L.H.G. was a freshwater sample, it might be suggested that a within group distance estimate should have been used (≈ 1.39). However, it was felt that the mean value (2.31) was larger and therefore more conservative. There was little reason to expect that the distance between the L.H.G. sample and the other freshwater samples would be as great as the mean distance between freshwater and marine samples. Therefore, the overall mean was used.

Page 209

The length of the discussion on the emprical significance of the canonical variates might be considered limited. Although some authors have enthusiastically interpreted the variates others have been considerably more cautious. Although I am not opposed to the former practice, I personally would like to withhold my participation until I am convinced that the arrangement of the samples is not the result of a fortuitous sorting effect. However, I don't doubt that all axes associated with non-zero latent roots supply information with regard to the levels of difference between samples. Therefore, it was common practice in my work to incorporate this information in the various distance matrices and resulting dendrograms.