Organizing committee:

Prof (Dr.) Mohan Joseph Modayil
Director, CMFRI, Kochi

Dr. P.C. Thomas
Winter School Director

Co-ordinators:

Dr. R. Paul Raj, Head, PNPD
Dr. K.C. George, Principal Scientist
Dr. P. Jayasankar, Senior Scientist
Dr. D. Noble, Senior Scientist

WINTER SCHOOL

ON

Recent Advances in Mariculture Genetics and Biotechnology

4th to 24th November 2003

Course Manual

INDIAN COUNCIL OF AGRICULTURAL RESEARCH
CENTRAL MARINE FISHERIES RESEARCH INSTITUTE
P.B. No. 1603, Tatapuram P.O.,
Kochi – 682 014

TRUSS NETWORK ANALYSIS FOR FISH GENETIC STOCK DISCRIMINATION

T.V.Sathianandan
Central Marine Fisheries Research Institute, Cochin

Introduction

Groups of potentially interbreeding natural populations, which are reproductively isolated from other such groups, is referred to as an animal species. Both genotypic and phenotypic homogeneity among groups belonging to the same species are seldom seen due to factors like environmental differences, isolation by distance and natural selection. These distinctive groups are known as races and referred to as stocks in the case of fish species.

Stock

- A self-sustaining group of individuals sharing a common unrestricted gene pool.
- Genetically distinct populations within a species, which are unique biological entities.
- It is a panmictic sub unit of a species that is generally in Hardy Weinberg equilibrium.

Stock variability is important to a species for continued successful reproduction and adaptation. Fishery biologists are interested in stocks to understand the spatial and temporal dynamics of stock differentiation and to use this information for conservation and management of the species. In fisheries it is important to identify the geographical distribution and genetic characteristics of stocks. The two popular methods of stock identification are

i. Identification based on gene frequencies through Protein gel electrophoretic studies.
ii. Identification based on morphometric studies.

Morphometrics

Morphology is a primary and direct means by which organisms interact with environment. In experimental biology it is useful to know whether two populations of organisms/organs have the same typical body form to indicate

- size allometry.
- shape changes accompanying size increase over the life span.
- to characterize the difference between sexes.
- response of form to therapeutic intervention.
- response to environmental variation etc.

- Morphometrics is the study of the geometrical form of organisms, which combines themes from ***biology***, ***geometry*** and ***statistics***. Here the ***geometric form*** of organisms is analysed.

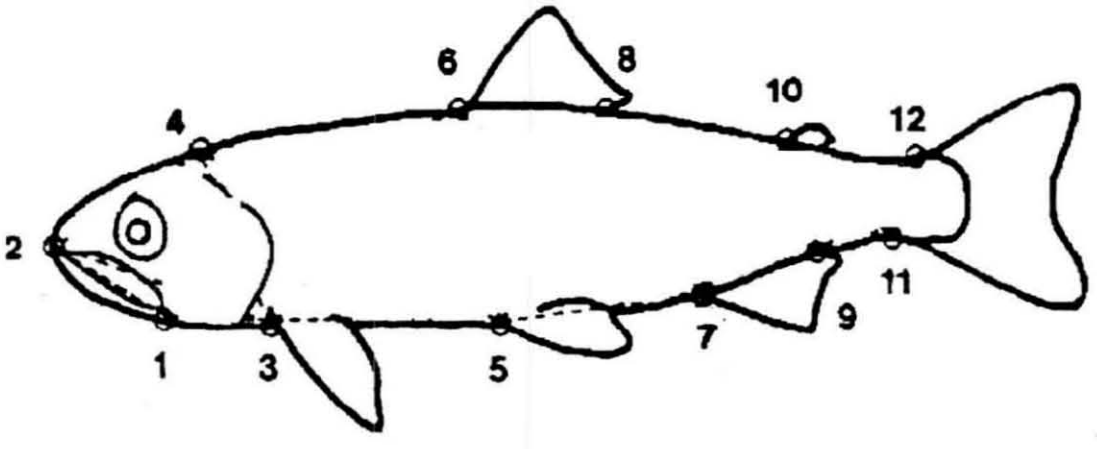
- Morphometric studies require information from ***biological homology*** and ***geometric location***.

- Biological homology is a spatial or developmental correspondence among definable structures or parts. (E.g. separate bones, nerves, and muscles). In the context of morphometrics it becomes a correspondence not of parts to parts but of points to points called a homology mapping.

- In morphometrics we study the [illegible] of biological [illegible]. The variation in the relative locations of sets of homologous points over a sample of forms.

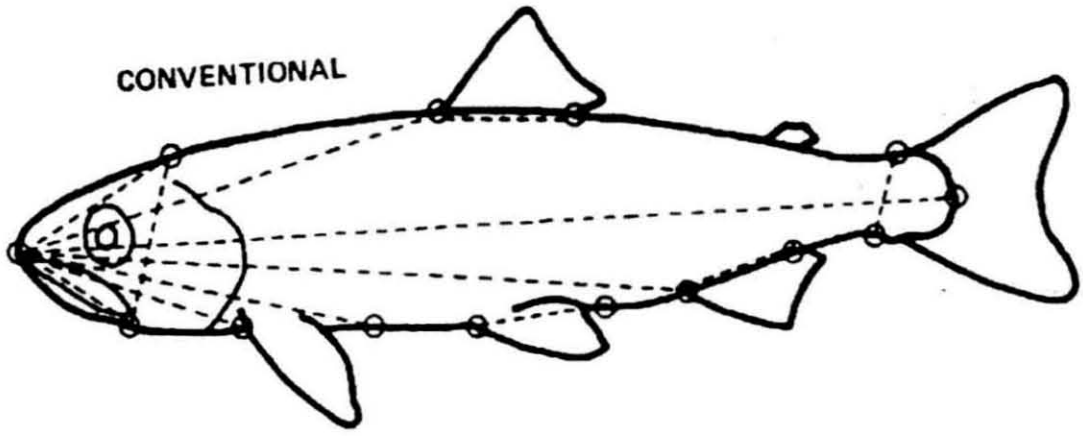
- The map of the organism is normally sampled at small number of discrete points called ***landmarks***.

Landmarks



Landmarks are defined intrinsically in terms of the anatomy in their vicinity. These are points pointed out by biologists when we talk about form of an organism. Some of the landmarks are located by juxtaposition of different identifiable structures (E.g. Anterior fin base and posterior fin base delimit the fin upon the body outline). Other landmarks are located by geometric properties (E.g. Point where the curvature of an edge is maximum).

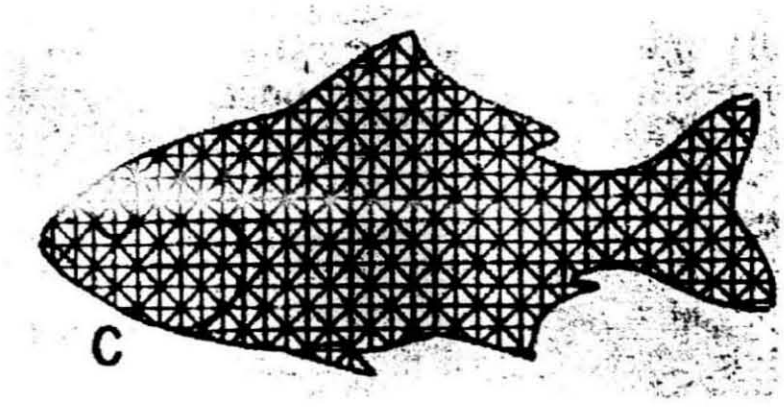
Truss Network Analysis: In systematics the interest is often in quantifying differences in form among different species or conspecific populations. When these are studied using conventional measurements (shown below) the amount of information available for analysis are repetitious and lack variation in oblique directions.



There are several biases and weaknesses inherent in traditional character set used to study stock differences in systematics.

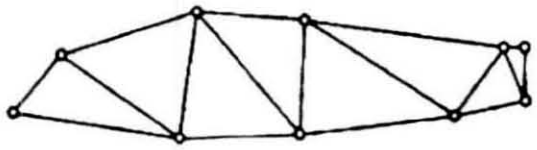
- They tend to be in one direction only (longitudinal) lacking information of depth and breadth.
- Coverage is highly uneven both by region and orientation
- Some landmarks like tip of the snout and posterior end of vertebral column are used repeatedly.
- Many landmarks are external rather than anatomical and their placement may not be homologous placement may not be homologous from form to form.
- Many measurements extend over much of the body.
- When measurements are taken on soft-bodied organisms, the amount of distortion due to preservation cannot be easily estimated.

The most ideal measurements, which overcome these problems, is as in the picture down below.

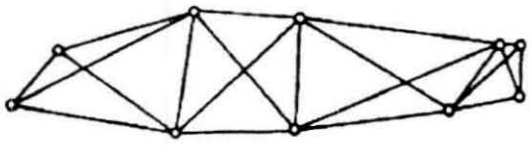


Alternative types of measurements are:

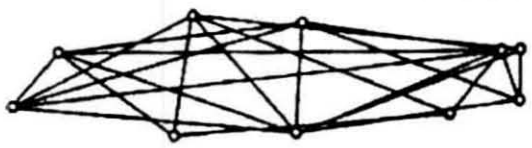
A. TRIANGULATION 2n-3 distances



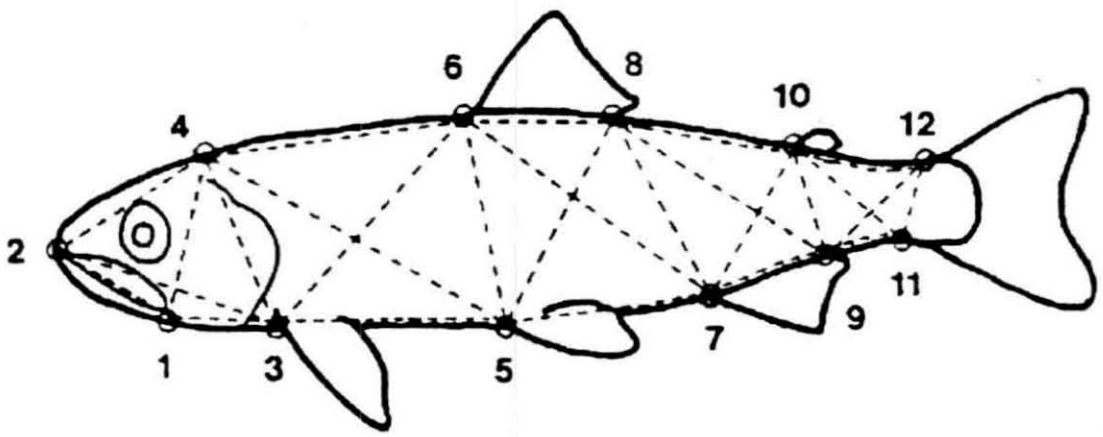
B. TRUSS 5n/2 - 4 distances



C. GLOBAL REDUNDANCY 3(n-2) distances



Truss is a geometric protocol for character selection, which largely overcomes the disadvantages of conventional data sets, and it leads to certain style of analysis. In *truss* system, homologous landmarks on the boundary of the form are divided into two tiers and paired. The distance measures connect these landmarks into an over determinate truss network which is a series of quadrilaterals each having internal diagonals. Each quadrilateral shares one side with each succeeding and preceding quadrilaterals (see figure below).



The following are the properties of a truss network measurements.

- It enforces systematic coverage across the form

- It exhaustively and redundantly archives the form
- The degree of measurement error in data can be measured and corrected
- Forms may be standardized to one or more common reference sizes by representing measured distances on some composite measure of body size and reconstructing the form using the distance values predicted at some standard body size.
- Principal components can be given geometrical interpretations. Component scores are measures of configuration while loadings are descriptors of shape change.
- Composite mapped forms are suitable for biorthogonal analysis of shape differences between forms.

Collection of Truss Measurements Data

Different Methods of collection of truss measurements data are:

- Position the specimen on light plastic in the field and take photograph with a scale in the frame. Prepare slide and project on to a digitizing tablet attached to a graphic terminal. With appropriate software locate the landmarks using the cross hairs of mouse and store the co-ordinates of landmarks.

- Place the specimen on water-resistant paper and tease the body posture and fins into a natural position. Around the outline of the form identify the landmarks. Record each landmark by making a hole in the water resistant paper with a dissecting needle. *Transfer the co-ordinates of landmarks by placing the paper on a digitizing pad and depressing the attached digitizing stylus into each hole.*

- Make the truss measurements using digital calipers connected to a Polycorder data logger. Using scanner and digitizer connected to a computer, images of specimens can be digitized and stored. With the help of an image processing software the landmarks can be identified and the truss measurements can be made.

Data Analysis

Classification problems exist in numerical taxonomy in biology and many other branches of Science. The interest here is to classify objects into one of many existing classes and is based on [illegible] taken on a set of [illegible] (called variables). Hence classification is [illegible] two to [illegible].

- We have multiple measurements data from a number of individuals belonging to known groups. Also we have data collected on individuals whose group membership is not known and is to be determined using the measurements made on them. This problem in statistical terminology comes under Discriminant Analysis.
- Another type is the case when the groups are them selves unknown and a primary purpose of the analysis is to find groups so that those belonging to same group are similar than those belonging to different groups. This in statistics come under the heading of cluster analysis or pattern recognition.

Cluster Analysis: This involves the search through multivariate data for observations that are similar enough to each other to be usefully identified as part of a common cluster. Clusters consist of observations that are close together and that the clusters themselves are separated. If each observation is associated with only one cluster, then the clusters form a partition of the data. Finding the partition into clusters is not always easy. There are numerous methods for clustering. Some methods of making clusters starts with models like mixture models of clusters. Examples of application of cluster analysis are studying genetic diversity within and between populations of and endangered fish species, clustering species of bees into higher-level taxonomic groups, developing clusters of patients based on physiological variables, constructing a speaker-independent word recognition system etc. Numerical methods of clustering with out any model can be into three major types; *hierarchical, partitioning* and *over lapping.*

Principal Component Analysis (PCA)

The objective here is to find linear combinations of the variables so that the first linear combination accounts for maximum possible variation in the data, the second linear combination accounts for the next highest possible variation and so on.

- PC analysis produces another set of variables that are linear combinations of the original variables. The new set will have the property that they will be mutually uncorrelated (orthogonal) and by considering few of them we will be able to explain a major portion of the variability in the population.
- If there are only a few clusters, the leading principal axes will tend to pick projections with good separations.
- PC analysis tend to act as a variation reducing technique relegating most of the random noise to the trailing components and collecting the systematic structure into the leading ones.

In principal component analysis we have a sample of observations taken on a set of variables and the objective is to find linear combinations of the variables so that the first linear combination accounts for maximum possible variation in the data, the second linear combination accounts for the next highest possible variation and so on. By this we get another set of transformed variables, which are linear combinations of the original variables and they, new set will have the property that by considering few of them we will be able to explain a major portion of the variability in the population. The approach in principal component analysis is to reduce dimensions by calculating the eigen values and eigen vectors of the covariance or correlation matrix and project the data orthogonally into the space spanned by the eigen vectors belonging to the largest eigen values. These projections are interesting due to the following reasons

- If projection is an aggregate of several clusters, then these can become individually visible only if the separation between clusters is larger than the internal scatter of the clusters. Thus, if there are only a few clusters, the leading principal axes will tend to pick projections with good separations.
- It tend to act as a variation reducing technique relegating most of the random noise to the trailing components and collecting the systematic structure into the leading ones.

Suppose that we have measurements on k variables $x_1, x_2, \text{؟}, x_k$ made on n individuals. Then we have $n \times k$ matrix of data and we can work out means for these variables which we can treat as a mean vector of length k. Also we can compute the variance covariance matrix **S** matrix using this data set. This matrix will be then used to compute the k principal components, say $z_i = a_{1i}x_1 + a_{2i}x_2 + \text{؟} + a_{ki}x_k$ for $i = 1,2, \text{؟}, k$ and the amount of variation explained by each of them will be available as $\lambda_1, \lambda_2, \text{؟}, \lambda_k$ where $\lambda_1 \geq \lambda_2 \geq \text{؟} \geq \lambda_k$.

In the analysis of multivariate data collected through truss network measurements the concept is that size and ***shape*** are the two factors, which account for the association among the distance measures. *Size* is not considered as a single variable but as a factor, which is obtained as a linear combination of the distance measures. *Shape* is considered as the geometry of the organism after information about position, scale and orientation has been removed. The *shape* discriminator should be independent of size, for it to be free from the effect of growth. Principal component (PC) analysis, which does not require any prior information about groups, is used in the analysis of truss data. A logarithmic transformation is first applied to the measurements before performing the PC analysis to reduce variance due to size variation and also because according to an allometric model diverse distance measures relate loglinearly in a homogeneous population. The first component factor of the PC analysis is then interpreted as size component (which is not fully free from shape) and subsequent component factors are designated as shape variable (not fully free from size). Then a plot of the first principal component scores against the second principal component scores will more of less show clustering for different groups. The percentage of variation explained by these two factors also should be considered before making conclusions.

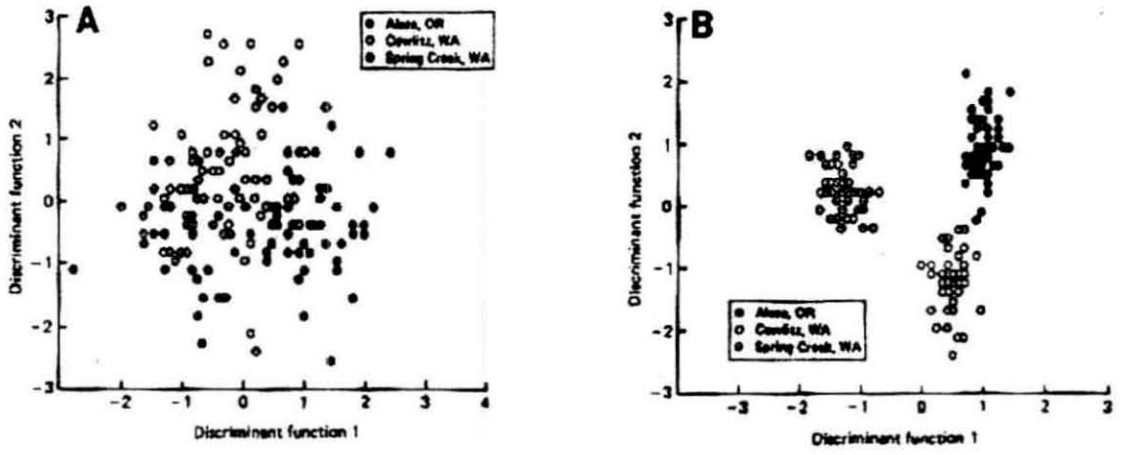
[illegible] principal component analysis of mean adjusted data by group to remove size influences from PC scores and vectors.

Steps:

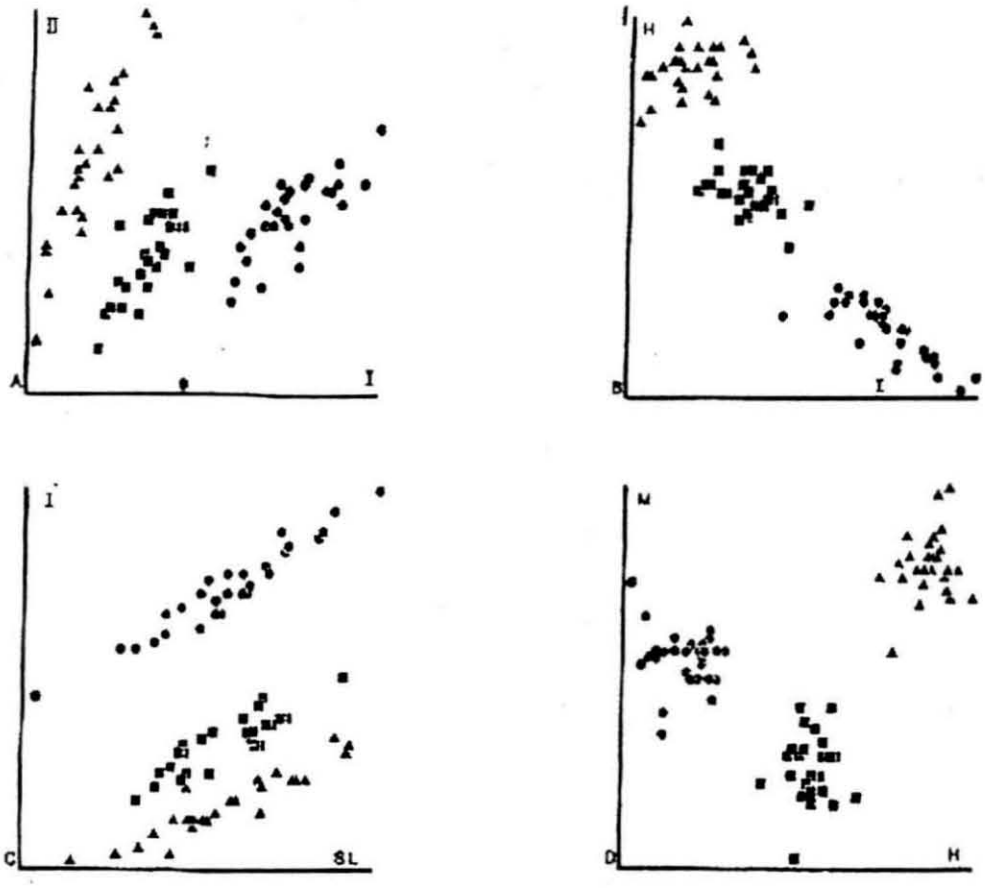
1. Transform the truss measurements data using logarithms. (According to the allometric model diverse distance measures relate loglinearly in a homogenous population).
2. Using the pooled covariance matrix ***Q*** compute the PC scores by evaluating the egen structure of matrix ***Q***.

3. From the scatter plot of the first two PC scores (say *PCI* and *PCII*) identify clusters associated with size and shape differences among populations.
4. Compute the covariance matrix Q' adjusted to zero mean for each of the identified clusters compute the PC scores by evaluating the egen structure of Q'. The first PC score, say S, will then be a within group size component.
5. Adjust the first two PC scores from the original analysis based on Q to zero mean for each of the identified clusters, say PCI_z and $PCII_z$.
6. Express the confounding of size component S with the second PC by regressing $PCII_z$ on S and denote the slope by α.
7. Estimate the portion $\hat{S}$ of S that lies in the plane of PCI_z and $PCII_z$ from a multiple regression of S on PCI_z and $PCII_z$ to yield the regression coefficients β_1 and β_2.
8. The shape discriminator H, known as the sheared factor II is then computed as, $H = -\alpha\beta_1 PCI + (1-\alpha\beta_2) PCII$. This will be uncorrelated with intracluster size and retains all discriminatory power original PC scores.

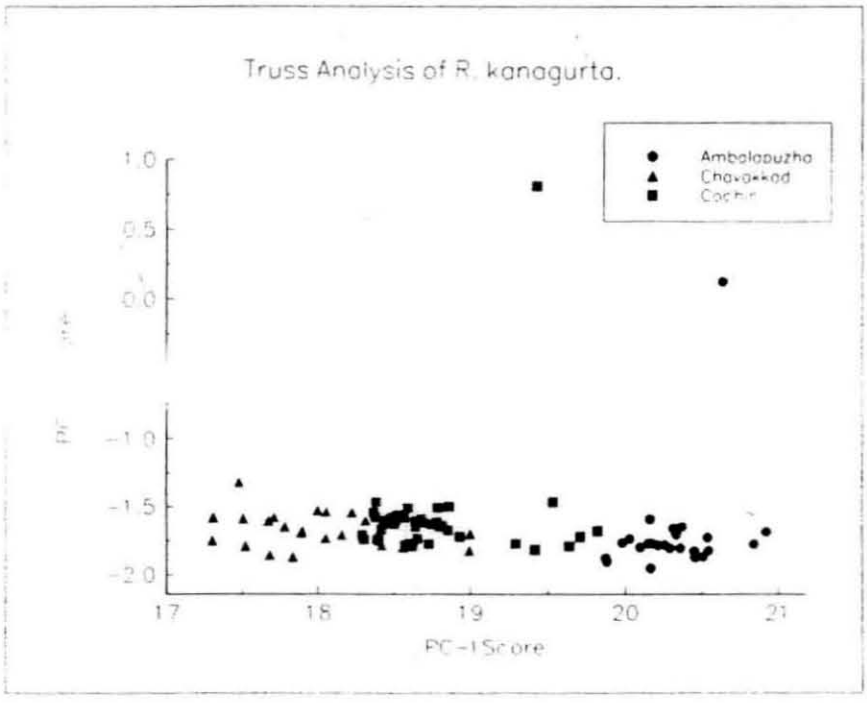
Illustrative Examples:



Scatter plot of scores based on A) conventional measurements and B) truss network measurements for chinook salmon from three locations.



Scatter plot of scores for three species of minnows A) PCI and PCII scores B) Sheared PCII against PCI C) PCI against standard length D) Sheared PCII against PC scores computed for meristic variables.



Reference

1. Anon. 1989. Discriminant Analysis and Clustering. Stat. Sci., 4(1):34-69.
2. Campbell, N.A. and Atchley, W.R. 1981. The geometry of canonical variate analysis. Syst. Zool., 30:268-280.
3. Darroch, J.N. and Mosimann, J.E. 1985. Canonical and principal components of shape. Biometrika, 72(2):241-252.
4. Dryden, I.L. and Mardia, K.V. 1992. Size and shape analysis of landmark data. Biometrika, 79(1):57-68.
5. Humphries, J.M. et. al. 1981. Multivariate Discrimination by shape in relation to size. Syst. Zool., 30:291-308.
6. Huber, P.J. 1985. Projection Pursuit. Ann. Statist., 13(2):435-475.
7. Misra, R.K. 1988. Quadratic discriminant analysis with covaraince for stock deliniation and population differentitaion. A study of Beaked Red fishes. Can. J. Fish. Aquatic Sci., 42: 1672-1676.
8. Morrison, D.F. 1990. Multivariate Statistical Methods. McGraw-Hill, New York.
9. Sampson, P.D. and Siegel, A.F. 1985. The measure of size independent of shape for multivariate lognormal populations. J. Amer. Statist. Assoc., 80(392):910-914.
10. Stocker et. al. 1984. An evaluation of morpometric and meristics for stock seperation of pacific herring. Can. J. Fish. Aquatic Sci., 41:414-422.
11. Strachan and Kell. 1995. A potential method for the differentiation between haddock fish stocks by computer vision using canonical discriminent analysis. ICES J. Mar. Sci. 52(1):145-149.
12. Strauss, R.E. and Bookstein. 1982. The truss: body form reconstruction in morphometrics. Syst. Zool., 31:113-135.
13. Winans. 1984. Multivariate morphometric variability in Pacific Salmon: Technical demonstration. Can. J. Fish. Aquatic Sci., 41:1150-1159.