CMFRI

Winter School on
Impact of Climate Change on Indian Marine Fisheries

Lecture Notes

Part 2

Compiled and Edited by

E. Vivekanandan and J. Jayasankar

Central Marine Fisheries Research Institute (CMFRI),
(*Indian Council of Agricultural Research*)
P.B. No. 1603, Cochin - 682 018, Kerala

(18.01.2008 - 07.02.2008)

भाकृ अनुप
ICAR

केंद्रीय समुद्री मात्स्यिकी अनुसंधान संस्थान
CENTRAL MARINE FISHERIES RESEARCH INSTITUTE

PRIMER- A STATISTICAL CURTAIN RAISER

52

J. Jayasankar
Central Marine Fisheries Research Institute, Kochi 682 018
(jjsankar@gmail.com)

PRIMER (Plymouth Routines In Multivariate Ecological Research) is a software aimed at analyzing data arising out of ecological and environmental investigations. But the scope of the software does not stop there. It is amenable to farther ranges and more applications once customized along with subtle pre and post processing maneuvers. While it can be grouped alongside any other multi-utility statistical software like SPSS, SYSTAT etc., it differs significantly from the bunch on its typicality of usage and the output generated by it followed by its interpretation. It is one of the few select software that prioritizes multivariate data analysis as deemed fit for environmental and ecological studies.

In statistical conceptualization multivariate data analysis occupies a place of pride as most of the univariate statistical rigours can be viewed as particular specific cases of the same. In other words multivariate tests and structures happen to be generalizations of many univariate tests and setups wherein an added advantage of co-habitation of more than one variable is available. Further the multivariate setups mimic the nature far more closer than the univariate setups. Environmental causes when dealt purely as a time series or as a causal setup for a range of derived effects, which often is the case in climate change related investigations, are best portrayed as sets instead of being viewed separately. The intricacy of relation between various major environmental parameters like Sea Surface Temperature (SST), Sea Surface Precipitation (SSP), Chlorophyll Content (CC), El Nino and Southern Oscillations Index (ENSO),Sea Surface Wind Velocity (SSW) etc are better respected than dealing them independently. As often is the case, these parameters recorded for a particular zone over a period of successive time states, prove to be more informative and useful than treating them as separate happenings and trying to regress effects upon them.

As of today the PRIMER software is now being used world wide for all types of marine community surveys and experiments, of benthic fauna, algae, corals, plankton, fish diet studies etc. It is being used even in pure multivariate studies of physico-chemical characteristics. In fact the routines used by PRIMER are so unique in nature that the underlying statistical foundations are always a shade trickier than ordinary statistical formulations, hence need cautious spadework. The package is the culmination of strategic maneuvers that have been perfected by the Community/ Ecology/ Biodiversity group at Plymouth Marine Laboratory (PML) over years and has a proven track record.

The methods employed by the routines can be broadly categorized into three groups.

(i) Univariate methods:

These are the much focused and widely practiced statistical tools which have been well documented. But in face of multiple causes and effects warranting attention, these single dimensional phenomena need proper justification at the initial stages. Once we start employing these methods, what we involuntarily commit is the fact that the variables under focus are relatively independent of any other factor of co-existence. For example when we study the abundance of a species of fish in isolation it has the inseparable assumption that the influence of other species of fish on the species under focus has been negligible. Hence these set of tools need a very crucial decision to be made even before venturing into data preparation. One of the justifiable usages of these techniques is the calculation and comparison of various indices like species diversity index which might be some measure of the numbers of different species for a fixed number of individuals (species richness). Another similar univariate measure is the biodiversity index which measures the degree to which species or organisms in a sample are taxonomically or phylogenetically relate to each

other. Another scenario which can be fitted into the univariate mode is while studying the response of single taxon indicator species to particular environmental gradient.

(ii) Distributional techniques:

In exploratory statistical tools plotting of summary data assumes immense value, especially when very less is known of the variable under study. These contrast from the univariate methods on the count that multiple streams of data can be processed simultaneously. One good example would be the case of plotting counts of species from samples converted into percentage abundance relative to total number of individuals in the sample, and plot the cumulated percentages against the rank of the species. Another useful application of this group of applications is plotting the number of species falling in different abundance ranges against geometrically scaled abundance classes. Here the emphasis is more on the simultaneous depiction of summary values of more than one variable at a time.

(iii) Multivariate methods:

Statistically placing, multivariate techniques deal with summarizing and inferring with more than one variable being considered simultaneously. To put in terms of marine researchers it amounts to something like comparing two samples taken at two different time intervals or two locations on the extent to which these samples share particular communalities like species. The measure of likeness or unlikeness leads to a measure of similarity/ dissimilarity calculated between pair of samples. These types of similarity coefficients lead to classification or clustering of the samples as well as ordination plot in which the samples are mapped in such a way that the distances between pairs of samples reflect their relative dissimilarity of species composition. In other words the manifestations expressed in terms of multiple dimensions have been reduced to singular values which can be ranked. PRIMER provides operations based on these lines like hierarchical clustering, multi dimensional scaling and principal components analysis.

Let us have a peek preview of these methods by way of focusing one module under each one of them.

(i) Univariate Techniques:

Under the univariate setup discussed in detail earlier there are different stages at which the tools can be applied. Let us focus on the determination of stress levels. Let us explore the case of average taxonomic diversity. Species richness (S) is a measure which either can be simply defined as the total number of species present or some adjusted form which attempts to allow for differing numbers of individuals. These species richness indicators form the essential part of diversity indices which give an overall view of multi-species, multi-locational data into a single index. The other aspect of standardizing samples of multi-species data is a measure of their evenness. For example if two samples comprising 100 individuals and four species had abundances of 25,25,25,25 and 97,1,1,1, it is obvious to state that the latter sample lacked evenness. Evenness can be worked out as the function of diversity index (Shannon's index) and the species richness. Though S has been an accepted index of richness of species, it has got its dose of disadvantages too, A few reasons are as follows:

(a) The observed richness is too dependent on the sample

(b) Species richness has no direct reflection of the phylogenetic diversity

(c) Statistically the test on departure of the diversity from expected values doesn't exist.

(d) Another interesting feature of richness which attributes to its disadvantage is the fact that its response to environmental annihilations is not unidirectionally correlated.

Towards addressing these problems pairing of the species abundance along with a measure of taxonomic distances was suggested by Warwick and Clarke (1995). As per that approach the taxonomic distances are standardized by the number of steps to be covered in the tree of Linnean classification. Suppose the species belong to the same family, the seps may comprise the immediate genus of first species and then

to the family and then to the genus of the second species before reaching the species itself. The maximum number of sreps to be taken is equated to 100 and all the pairwise distances between the species are recalculated to match the standardized longest distance. After the calculation of these taxonomic distances ($\grave{u}_{ij}$) between the ith and jth species whose richness is denoted by x_i and x_j, an average taxonomic diversity is defined as

$\Delta = \left[\sum\sum_{i<j} \omega_{ij} x_i x_j\right] / [N(N-1)/2]$ where $N = \sum_i x_i$ i.e. the total number of individuals in the sample.

The average taxonomic diversity has a simple interpretation that it gives the average taxonomic distance between every pair individuals in the sample. As a special case when all the species get collapsed to a single level i.e. when all of them belong to the same genus $\grave{u}_{ij}$ s take unitary value or no relevance as all the distances are same and if we express p_i as x_i/N then the previous expression can be re-written as

$\Delta^o = [2\sum\sum_{i<j} p_i p_j] / (1 - N^{-1})$ which is a form a Simpson diversity. Dividing Ä by Äº gives the average taxonomic distinctness which single out its focus on taxonomic hierarchy.

Other similar diversity measures can be derived and defined taking into account the various exigencies arising out of the nature of the study and the phylogenetic uniqueness of the population.

(ii) Distributional Techniques:

One of the major challenges facing researchers dealing with marine ecological studies is the issue of discriminating locations or sites is by comparing the data summaries on equal footing. A classical tool in statistics for this situation would be testing the null hypothesis that two or more sites (or conditions) have the same curvilinear (pattern) structure. The easiest method to effect the testing would be to perform Analysis of Variance (ANOVA). But as is known very well, ANOVA in the classical sense has more stringent assumptions about the population and the distribution. Hence if the same were to be performed on variables like Bray-Curtis similarity which have less to resemble the sample means of ANOVA concept. There range is limited and they are proportions and hence have less to do to fulfill the normality assumptions. Hence for such ordination methods the classically rooted univariate ANOVA methods and their multivariate extension MANOVA will stand less chance of justification. A valid test for such situations should be built on a simple non-parametric permutation procedure, applied to the similarity matrix underlying the ordination or classification of samples. Hence PRIMER propounds an analogous test termed as Analysis of Similarities (ANOSIM) to face such multiple comparison problems. The cue is taken from the basic methodology wherein the between categories variation is measured against within categories variation (the one which cannot be explained more). The null hypothesis (H_0) is that there are no differences in community composition at different sites (if we consider a study involving samples from different locations). The null hypothesis is examined in the following steps:

(i) The test statistic (a function involving sample observations) is computed reflecting the observed differences between sites, contrasted with the differences among replicates within sites. Using any typical methodology the distances between samples can be computed (viz Bray- Curtis similarity or MDS distance). The ideal test would then be based on the average distance between pairs corresponding to different sites and those within the sites. If $\bar{r}_W$ is defined as the average of all rank similarities among replicates within sites and $\bar{r}_B$ is the average of rank similarities arising from all pairs of replicates between different sites,

then a suitable test statistic is

$$R = \frac{(\bar{r}_B - \bar{r}_W)}{\frac{1}{2}M}$$ where M=n(n-1)/2 and n is the total number of samples under consideration. It has to be noted that the highest similarity corresponds to a rank of 1 (the lowest value), following the usual mathematical convention for assigning ranks. The denominator, M/2 ensures that R can never lie outside the range (-1,1). It also ensures that R will take the value unity only if all replicates within the site are more similar to each other than any from other sites. R will become zero only when the similarities between and within the sites will be same on average. R can seldom take sub-zero values as that may imply that the similarities between locations is far higher than those within the locations.

(ii) Once the R statistic is computed it is recomputed many times for creating a distribution of the same. This is done as R does not fall under the classical mould of a sample statistic with a well defined sampling distribution. The samples and the replicates are permuted and the R statistic is recalculated for each permutation. The rationale for this test is if the null hypothesis were to be true that will mean that there will be little effect on average to the value of R if the labels identifying which replicates belong to which sites are arbitrarily rearranged. In general there would be $(kn)!/[(n!)^k k!]$ where n replicates each at k sites are rearranged.

(iii) Once the R values for the rearranged labels were computed the locus followed by the estimated values gives an authentic glimpse of how the sampling distribution would be. From the number of recomputed R values which are equal to or greater than the R value of the original sample, the null hypothesis can be rejected at a significance level of (t+1)/(T+1) where t is the number of simulated values greater than or equal to original R out of a total T simulations.

(iii) Multivariate Methods:

Most of the multivariate routines offered by PRIMER target ordination of samples based on more than one trait considered simultaneously. The famous classical multivariate methods like Cluster analysis, Principal Component Analysis, Principal Co-ordinates analysis and Multidimensional Scaling are best utilized for such ordination of marine ecological data. For a focused elucidation let us focus on multi-dimensional scaling (MDS) as an ordination tool.

MDS is a complex numerical algorithm (can be conveniently left to suit the software's imagination!) but its base is logically very simple. The strength of this method is that it assumes very little model behaviour and the link between the final picture and that of the user's data is relatively easy to explain. By virtue of its being a basically non-parametric tool, it addresses the main criticisms hurled at Principal Components Analysis. The non-metric MDS, the purest non-parametric form that MDS can attain, starts with similarity or dissimilarity matrix among samples. This can be whatever similarity matrix that can be biologically relevant to the questions being asked of the data. In fact the superiority of this method lies in the fact that even with the similarity/dissimilarity matrices this method works on relative aspects of the pairings. MDS focuses on the rank of dissimilarity rather than the absolute measure of the same. In a nut shell MDS constructs a map of the samples in a specified number of dimensions, which attempt to satisfy all the conditions imposed by the rank similarity matrix. The two general features of MDS are

(a) The MDS plots can be arbitrarily scaled, located, rotated or inverted. Clearly the MDS does not deal with the absolute distance apart of two samples, instead relative distances have been focused.

(b) The algorithm of MDS methodology strives at reducing the distortion or stress when a multi dimensional similarity distance matrix is plotted in a reduced dimensionality meta plane. Not only the method reduces the stress but also gives a measure of the same.

A typical MDS algorithm would have the following stages:

(a) The reduced number of dimensions have to be specified.

(b) A starting mapping of the n samples have to be made , may by PCA or PCoA.

(c) Regression of the interpoint distances in the new plot over the dissimilarity measure of the original setup. The regression may be plotted based on simple linear arrangement between the new measure d and the original multivariate dissimilarity ä or the same may be based on a non-parametric paradigm.

(d) The goodness of fit of the regression happens to be the stress defined as follows:

$$\text{Stress}=\sqrt{\sum_j \sum_k (d_{jk} - \hat{d}_{jk})^2 \Big/ \sum_j \sum_k d^2_{jk}}$$

Where $\hat{d}_{jk}$ is the distance predicted from fitted regression line corresponding to dissimilarity $ä_{jk}$. If

$d_{jk}=$ for all n(n-1)/2 distinct pairs, then the stress is the least, viz 0.

(e) The next step is to choose an optimization method which will alter the stress values for changes in ordination values of the plot and finally selecting a direction where the fall in stress values will be more significant than the rest.

(f) And finally repeating steps from (c) to (e) till convergence is achieved.

Though loaded with a score of pluses MDS also has its share of drawbacks too. The main drawback is that this method is computationally more demanding and secondly convergence at a global minimum of stress is not always guatanteed. $\hat{d}_{jk}$

Though PRIMER is replete with a bunch of such specif tools which are of immense utility value in Ecological and Marine research, we have considered an objectively selected few for getting an idea about the set of routines and how they tackle inferential issues. Hence it is advised that an exhaustive hands on experience with the various modules as well as study of select references will throw more light into using this software more efficiently along with interpreting the results in a more effective manner.

Suggested readings:

Clarke, K.R. and Warwick, R.M, (1998) Similarity based testing for community pattern: the 2-way layout with no replication. *Mar. Biol.* 118: 167-176

Clarke, K.R. and Warwick, R.M, (1998) A taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels. *Mar. Ecol. Prog. Ser.* 184: 21-29

Kendall, M.G. (1970) *Rank correlation methods*. Griffin, London.

Bray, J.R. and Curtis, J.T. (1957) An ordination of the upland forest communities of South Wisconsin. *Ecol. Monogr.* 27: 325-349.

Warwick, R.M. and Clarke, K.R. (1991) A comparison of methods for analyzing changes in benthic community structure. *J. mar. Mar. Biol. Ass. U.K.* 71: 225-244.