CHAPTER 25

Introduction to Primer and Statistical Methods in Ecological Data Analysis

J. JAYASANKAR
ICAR-Central Marine Fisheries Research Institute

Introduction

Although analytical methods in statistics have all along been generic and evolutionary in the first half of past century, the developments happening in the field of computational statistics in the past couple of decades are more need based and custom tuned. A lot of effort is being put in by researchers in bundling methods, theory and procedures in classical statistical literature on their common applicability to a targeted exploration. It is common place to collate various univariate, multivariate, parametric, non-parametric, frequentist and non-frequentist methods, which have applications in different domains like ecology, clinical trials, bioinformatics etc. and tag them as per the domain subject matter. Thus the generic and specific procedures which are of relevance in exploratory and confirmatory analyses in the field of ecological studies of communities have been grouped under a common pivot. During the course of this discussion a couple of such statistical methods used in community structure studies would be dwelled upon.

On the ecological datasets

The typical community structure dataset would have either or both the tags, viz. temporal and spatial. The data could have been collated over multiple sampling spots in a region and also over a period of time. This makes these data to be looked upon from the time series as well as space-series points of view. And another ubiquitous feature of such datasets are their being multivariate. Communities, comprising many species at various levels of abundance, are always recorded as n-tuples at each sampling session and hence are multivariate at core. Although there are possibilities of isolating responses and causes from the bunch and possible univariate procedures could be applied upon, thereafter.

Multivariate tools

Analysis of ecological data involves almost the entire gamut of multivariate data analytical tools. The pivot based (could be labelled region or cluster) comparison of the community abundance has its roots in Hotelling's T square(d) thereafter raising to the multiple comparisons using MANOVA using Wilk's Lambda, Pillai's trace etc. Needless to add, a set of single response multiple regression analysis and univariate ANOVA get subsumed in the multivariate projection and analysis. The common thread in most of these analyses is the polarization of near independent components which have a telling impact on the response variables or the system tracking as a whole.

Another important area in multivariate analysis is the clustering and discrimination domain. The basic thrust in this sector is about measuring the closeness or remoteness of the multiple streaks of expressions of communities, which then gets utilized in grouping or clustering the similarly placed or paced dynamics or also for contrasting the most orthogonal or independent of bunches of variables which could sufficiently project the overall variability in the system. In a way these types of procedures aim at reducing the dimensionality of the bouquet of variables in such a way that inferences and depictions of scenario can be made with two or three dimensional projections. The community datasets often indicate similarity in pattern amongst their subsets, which when zoomed in would yield more interesting bio-climatic cause- effect mechanisms. Tools like Principal Component Analysis (PCA), ordinations by Principal Coordinate Analysis (PCoA) and Redundancy Analysis fall broadly under this conceptualization. Of this the RDA can be viewed as the multivariate extrapolation of univariate multiple regression analysis and it yields the proportion of variance of a set of variables that could be explained by a set of causative factors. PCoA has its action rooting on the distances (preferably Euclidean) between the multi-dimensional points and routing a starting point with its nearest neighbor in as much less a dimension possible so that the resultant scatter of these points clearly shows clusters based on which further PCA type recasting can be done. This is otherwise referred to as Multi-Dimensional Scaling (MDS), the metric variant of it. Also in the context of abundance of communities datasets, the dissimilarities (distances) between the observations can be estimated more nonparametrically (with less leanings on the traditional orthodox assumptions on the values thrown out by the study variables, aka distribution) by using a “Stress” reducing monotonic transformation which simultaneously takes care of point-point contrast as well as distances between the realized observations.

The major bottleneck or invisible opportunity with ecological datasets is that they are counts based with a large possibility of null entries. Also, at times the community sampling boils down to presence or absence type of information. Hence under these circumstances parametric exploration and testing on orthodox moulds would be highly inefficient and error prone. Hence a whole lot of quasi parametric or non-parametric tools have been conceptualized by resonating or tweaking the existing parametric options. One such set of tools is available in the Plymouth Routine In Multivariate Ecological Research (PRIMER).

PRIMER- a curtain raiser

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 Linnaean classification. Suppose the species belong to the same family, the steps 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 steps to be taken is equated to 100 and all the pairwise distances between the species are recalculated to match the standardized longest distance.

(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 (H0) 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

$$R = \frac{(\bar{r}_B - \bar{r}_W)}{0.5 * M}$$

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!)kk!] 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

(1) 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.

(2) 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 a 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 using a statistic called stress which is the function involving squared differences between each unique pair distance and the regression based distance. If for all possible unique n(n-1)/2pairs the distances happen to be same 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.

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.

The following routines enshrined in the software are quite useful in numerically testing and robustly inferring and graphically assimilating large sets of community sample sets.

(i) CLUSTER (grouping) (ii) MDS (Ordination) (iii) PCA (recast visualisation) (iv) ANOSIM (hypothesis testing) (v) SIMPER (sample discrimination) (vi) BEST (trend correlations) (vii) BIOENV (paired group comparison) and (viii) PERMANOVA (permutational multivariate analysis of variance) among others. PRIMER also has extensive routines for estimating various beta, alpha and gamma diversity measuring indices. All these routines are built on a near total non-parametric platform thereby warding off the presumption and assumption blues. A classic routine worth focusing on is ANOSIM. Smartly worded to sound akin ANOVA this routine has a refreshingly different set of approach rooted deeply on all generated by the data alone. Under this procedure the samples are treated as arrays whose rows are samples and columns are the component resources like planktons etc. Based on the intensity of the resources available in each location, a rank-based similarity matrix is generated equivalent to the sample dimension. This index popularly known as Bray- Curtis similarity is then subjected to the inter and intra factor comparison yielding a functional known as R statistic. The value falling between 0 to 1 with lower limit indicating perfect similarity in divergence within factor groups and between them and the upper limit indicating near perfect similarity between pairs within groups as compared to those between them, thereby indicating significant inter group heterogeneity. The measure of the R value's robustness is also arrived at by estimating the R estimate on prior number of large recombinations of the sample data and noting down the values of R falling above the one realized from the original sample. Thus the non-parametric conceptualization right from estimating the group similarity to studying its distributional aspect is complete in this approach.

CRAN- R language's Vegan:

Vegan- A contributed package totally dedicated to the procedures and methods discussed by Clarke and Warwick (2001), whose software version is Primer-E. This contains most of the common tools like dissimilarity measures, Anosim, BioEnv etc.

Other modeling options with ecological data sets

To start with even the simple multiple regression itself is a model in the strict statistical sense which depicts the role and measure of causal factor upon explaining the variability of the response variables. These regression models fall under the category of linear models with normality assumptions. However with the responses being binary at times and highly skewed and noisy counts on the other end of the spectrum, the classical assumptions of normality which validates the tests of significance are most inapplicable in these datasets. Hence the more liberated and broader versions of the linear model called Generalised Additive Models (GAM) are the most aptly poised set of paradigms to fit into such situations. With a wide range of link functions, smooth functions and a range of distributions including non Gaussian like Poisson etc. GAMs can practically link any type of causative variable with any type of response sets which can be foreseen in ecological studies. With many measures for their rates of success based on Information criterion, the best of such group of models can always be zeroed in on.

The developments made in the time series modeling area including the methods to split the time spanned datasets into components of trend, cyclicity etc. have come in handy while dealing with the biotic and temporal factors and their influence on the community structures. The direction oriented process based decomposition of time series like Asymmetric Eigenvector Mapping and the direction free mapping like Morgan/s Eigenvector Mapping have given a specific thrust towards modeling the data with a view to focus on temporal and spatial angles.

Tools like Local contributions to beta diversity (LCBD) help in arriving at comparative measures of ecological uniqueness of samples which would go a long way in studying and inferring about the community structures.

To conclude, it can be safely assumed that the rate of development of computational statistics has lead a sort of newer opportunities and horizons in locating and studying the hitherto unknown camouflaged patterns and undercurrents existing in community structure datasets. With the rate of innovation higher on the computational front the treading of hitherto unheralded territory is becoming all the more in vogue thing for researchers.

Referres

Legendre P, Gauthier O. 2014 Statistical methods for temporal and space–time analysis of community composition data. Proc. R. Soc. B .

Clarke, KR, Warwick RM (2001). Change in marine communities: an approach to statistical analysis and interpretation, 2nd edition. **PRIMER**-E.

Other classical statistical text books.