CHAPTER 42

Statistical Methods in Ecological Data Analysis

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

J Jayasankar *ICAR-Central Marine Fisheries Research Institute, Kochi, Kerala*

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 predominantly 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). 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 practically 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.

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.

Referred literature

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

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

(iii) Other classical statistical text books

Annexure:

Certain computational tools that can be put to use in Ecological data analysis

In R language

(1) 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.

(2) **CatDyn: Fishery Stock Assessment by Generalised Depletion Models**
As a recourse to viewing the stock dynamics through catch rather than the population, which is of course used as an index for the latter, routines have been developed to assess, model and predict stock health using Generalised Depletion models. The entire gamut of parametrisation, modelling and forecasting has been made handy by the R library CatDyn. As per the introduction given by the author(s) of CatDyn, the library is capable of the following:

Based on fishery Catch Dynamics instead of fish Population Dynamics (hence CatDyn) and using high-frequency or medium-frequency catch in biomass or numbers, fishing nominal effort, and mean fish body weight by time step, from one or two fishing fleets, estimate stock abundance, natural mortality rate, and fishing operational parameters. It includes methods for data organization, plotting standard exploratory and analytical plots, predictions, for 77 types of models of increasing complexity, and 56 likelihood models for the data.

The concept of depletion modelling is set into motion using the following parametrization.
The process equations in the Catch Dynamics Models in this package are of the form

$$C_t = ke^{-\frac{M}{2}}E_t^a N_t^b$$

$$N_t = N_0 e^{-Mt} - e^{\frac{M}{2}} \sum_{i<t} C_{t-1} e^{-M(t-i-1)} + \sum_j P_j e^{-M(t-j)}$$

where *C* is catch in numbers, *t*, *i* are time step indicators, *j* is perturbation index (*j*=1,2,...,100), *k* is a scaling constant, *E* is nominal fishing effort, an observed predictor of catch, *a* is a parameter of effort synergy or saturability, *N* is abundance, a latent predictor of catch, *b* is a parameter of hyperstability or hyperdepletion, and *M* is natural mortality rate per time step. The second summand of the expanded latent predictor is a discount applied to the earlier catches in order to avoid an *M*-biased estimate of initial abundance. Perturbations to depletion represent fish migrations into the fishing grounds or expansions of the fishing grounds by the fleet(s) resulting in point pulses of abundance. In transit models (limited to one fleet) there are also emigration events happening at specific time steps for each perturbation. In 2 fleet cases the fleets contribute complementary information about stock abundance, and thus operate additively; any interaction between the fleets is latent and affects the estimated values of fleet dependent parameters, such as *k*, *a*, and *b*.

The observation model can take any of the following forms: a Poisson counts process or a negative binomial counts process for catch recorded in numbers, an additive random normal term added to the continuous catch (in weight) predicted by the process (normal and adjusted profile normal), a multiplicative exponential term acting on the process-predicted catch such as the logarithm of this multiplier distributes normally (lognormal and adjusted profile lognormal), and Gamma (shape and scale parameterization).

The library CatDyn takes care of almost all the parameterisation issues and dishes out the type of output which would magnify the status of fisheries as seen from the macro dynamic level in such a way to aid the policy makers.

(3) **mefa-** Yet another package in R which specializes in data analysis using ecological information. This apart from dealing with community structure information, progresses to the extent of generating analysis based report in popular formats like LaTeX and html etc.

Other sources

(1) XLSTAT- is an MS Excel friendly data analysis package which performs canonical correspondence analysis in tandem with Excel spreadsheet and finds EC50 values etc. and omics data analysis.

(2) FLORA- is another software scripted for Windows environment, which handles the multivariate routines as applied to community structure data
Summarizing, it can be recorded that the tools mostly applied for dealing with eco- biological data sets based on communities of flora and fauna stem from multivariate analysis tools and the software variants focus mostly on the customized output and report generation.