Lichen Question Response

Imagine that you are a consulting statistician and you receive the email shown below from a colleague who is not a statistician. The email describes a debate about what is an appropriate statistical analysis to conduct. The background is that the researchers are interested in comparing beta diversity of lichen composition on trees among three different sites. Diversity is an important concept in ecology that, unfortunately, has multiple incompatible definitions which makes communication and discussion about it challenging. For the purpose of this assignment, it is adequate to consider beta diversity as a measure of compositional dissimilarity in the following sense: if two objects have the same set of species (or some other grouping variable such as morphotype) in the same proportions, then their beta diversity is zero; if there is no overlap in their composition, then the beta diversity is one. Values between zero and one measure the relative lack of overlap between two compositions. You may focus on a measure called Bray-Curtis dissimilarity which may be computed as

\[ BC = \frac{\sum_i |X_i - Y_i|}{\sum_i X_i + \sum_i Y_i} \]

if the observed counts for each possible type in one sample are \( X = (X_1, \ldots, X_k) \) and in a second sample are \( Y = (Y_1, \ldots, Y_k) \). Note that in the special case where \( \sum_i X_i = \sum_i Y_i = n \) and if we define \( p_i = X_i/n \) and \( r_i = Y_i/n \), then

\[ BC = \frac{1}{2} \sum_i |p_i - r_i| \]

Also in the email is the information that there are 10 trees sampled per site and that there are three morphotypes of interest: cructose, foliose, and fruticose. So, for each tree, the data is a factor which indicates the site the tree came from and three counts, one for each morphotype, that sum to 150.

Your assignment is to craft a response to the colleague. Your response should include answers to her two specific questions and should address the following points: (1) Is it appropriate to model the counts as Poisson distributed? (2) What do you think about treating the 45 pairwise comparisons between trees as separate observations for each site? (Specifically, compute BC for each pair of trees in a site and consider this to be a sample of 45 beta diversity measures for each site.) Furthermore, ask any additional questions that you would need answers to in order to give advice regarding an appropriate analysis.

Here is the email

I have come up against two statistics issues in an ecological study about which I need an expert opinion.

The study design is this:

The lichen community has been sampled on 10 trees selected randomly within
a chosen trunk size range, at each of 3 different sites. On each tree a single 150cm long vertical transect was placed; the number of cm occupied by each species along the transect was recorded with the maximum sum of 150 for each transect.

In analysis, a beta diversity measure (1- coefficient of community or Bray–Curtis distance) was calculated between all pairs of trees within site. Also, transect data were summarized as the sum of cm occupied for all crustose, foliose, or fruticose lichen morphotypes in a transect. Sum for each transect is still 150.

The goal of analysis is to compare beta-diversity values or abundance of a single morphotype between sites, with the set of pairwise beta-diversity values between trees (45) or abundance of a morphotype on a single tree (10) considered replicates within site.

Both kinds of data were found to have heterogeneous variances between sites.

The statistical analysis questions are:

Question 1: When using a generalized linear models routine that chooses the strongest of several possible models, is it appropriate to apply a log-linear model to compare beta-diversity values between sites? I read that a log-linear model should be applied only to binary data, but is that likely to be accounted for when the software identifies the best model? If so, what should we look for/report to support that the analysis was appropriate? If not, what should we do instead?

Question 2: When using a generalized linear models routine that chooses the strongest of several possible models, is it appropriate to apply a Poisson distribution model to compare abundances, in cm out of 150, of a single morphotype between sites? I have always been taught that a Poisson model applies to count data, for instance presence of a species in x number of independent quadrats. I thought that the abundance of a morphotype within transect of a fixed length should be considered proportional data. Does the independence of the 10 trees and the expression of abundance as \# of cm out of 150 make these data appropriate as is for a Poisson distribution model? If so that makes our analyses easier; again, what should we look for/report to support that this analysis is appropriate? If not, what should we do instead?

Thanks so much for your opinions. There has been a heated discussion about the statistics that is holding up submission of a time-sensitive ms. I do not want to stand in the way any further if my statistics concerns are invalid or not important.