Working Group on Biological Problems Using Binary Matrices
Final Report
March 2011

Many fundamental questions in ecology cannot be addressed experimentally because at the relevant large spatial and temporal scales, experimentation is impractical, unethical, or impossible. Instead, to investigate these questions inferences must be made from observational data. Null model testing comprises a key tool for making these inferences, allowing large-scale effects of processes such as environmental filtering, competition, and facilitation to be inferred from observations of species ranges, abundance distributions, body sizes, and other similar traits. Three types of ecological data commonly analyzed using null models include binary presence-absence matrices, which give the distribution of species over a set of sites; ecological networks such as food webs and pollinator networks; and phylogenetic patterns in community composition. All of these data can be coded in a binary form.

We assembled a team of scientists to examine the methods to make statistical inferences from binary data relating to patterns of species co-occurrence, ecological networks (food webs and pollinator networks), and community phylogenetics. We succeeded in retaining the interest of our working group members whose interests included species co-occurrence patterns and ecological networks, but not in the area of community phylogenetics.

Our group divided into sub-groups, and working group members could participate in more than one subgroup.

Species Co-occurrence Group

Incidence Approach – Joshua Ladau, William Godsoe, Dan Simberloff, Nicholas Gotelli, Richard Barker, Steven Schwager, Edward Connor

Abundance Approach – Robert Dorazio, Nicholas Gottelli, Edward Connor

Ecological Networks (Foodwebs and Pollinator Networks) – Stefano Allesina, Daniel Stouffer, Diego Vasquez

Community Phylogenetics -Steven Kembel

Species Co-occurrence

Incidence Approach

To investigate ecological questions inferences must often be made from observational data. Null model tests of presence-absence data (NMTPAs) comprise a key tool for making these inferences. NMTPAs are particularly important because the necessary presence-absence data are easily obtained, widely available, and relatively reliable.

In general, an NMTPA is constructed to exclude the process of interest, and if the data are found to be inconsistent with the model, then the process is concluded to possibly occur. NMTPAs are often used instead of process-based models because they give primacy to the parsimonious hypothesis -- that the process of interest is not occurring -- and they allow the rate of falsely concluding that a process is occurring to be controlled, as in statistical hypothesis testing.

In this subgroup, we addressed two major questions regarding NMTPAs. NMTPAs are generally developed and justified based on intuition. However, multiple tests can all seem intuitively appropriate for the same data, yet yield conflicting conclusions. The first issue we considered was developing and implementing an overarching mathematical framework to guide the development and application of NMTPAs. Second, we considered models of ecological assembly under which NMTPAs can be derived. Different NMTPAs are appropriate under different models, so it is key that a model appropriate to the ecological system of interest is developed and implemented.

We considered two mathematical frameworks for developing NMTPAs: the frequentist framework and the Bayesian framework. Within the frequentist framework, we investigated and implemented a maximum likelihood approach and an optimality approach for constructing NMTPAs. The maximum likelihood approach is relatively easy to apply, and leads to tests that perform well, particularly when sample sizes are large. The optimality approach yields tests that are optimal, exceeding the performance of the maximum likelihood methods and having the lowest Type II error rates subject to controlled Type I error rates. These uniformly most powerful tests are desirable, but they exist only in certain circumstances and are difficult to derive. The Bayesian framework is particularly useful with little data, and can allow estimation of numerous parameters. However, depending on the analysis, the Bayesian methods that we considered can be difficult to implement. These approaches complement each other and together will be useful for developing improved NMTPAs.

For all of the approaches, a model of possible associations between species must be assumed. We considered two general classes of models of community assembly. The first model is a loglinear model. In the case of no interactions, this model is equivalent to independently flipping biased coins for each species at each site: if a given coin comes up heads, then the species occurs at a given site; otherwise, it does not occur. The model allows for various associations between species, wherein coin flips are not independent. In the second model, species are thought of as arriving sequentially. In the non-interacting case, the odds ratio between observing any pair of species in a particular sequence remains constant regardless of which species have previously arrived. With associations, this odds ratio varies: for instance, a given species might be less likely to colonize a location if a competitor has already arrived than if it is absent, while a species that is a symbiont of the competitor might be more likely to colonize. These models are mathematically distinct, and lead to different NMTPAs using the frequentist and Bayesian approaches. We investigated the implications of these models and inferences to which they lead.

Overall, we believe that the approaches that we have developed will greatly improve the reliability and reach of NMTPAs. Through the use of these approaches, it will be possible to better understand the large-scale effects of many key ecological processes.

Edward Connor has begun a project to compare the computational complexity, as well as the bias and variance of estimates of co-occurrence metrics derived from the variety of existing algorithms for generating a sample of binary matrices with fixed marginal totals. He is collaborating with two graduate students at San Francisco State University, Steven Li a MS student in Mathematics, and Shenhaochen Zhu a MS student in computer science. Our goal is to provide guidance on which algorithms are optimal for generating a sample of matrices upon which to estimate the distribution of test statistics.

Abundance Approach

Historically. abundance data has received little attention as a basis for inferring non-independence of species among a site of sites. However, binary matrices of presence and absence data can be viewed as a truncation of the inherent population and community ecological processes that affect abundance to one in which site occupancy is the random variable of interest. We decided to entertain the possibility of not truncating abundance data, but rather to see if by modeling the counts of species at a set of sites we could make statistical inferences about the non-independence of species.

The first paper generated by the Abundance Approach (Gotelli et al. 2010) focuses on testing for temporal trends in species abundance using both more traditional null model approaches and also by fitting hierarchical models. The null model approach reached the conclusion that temporal trends were more heterogeneous than expected, and the hierarchical modeling approach was able to determine which species had long term trends toward declining or increasing numbers, or stable abundances.

The second paper also applied a hierarchical model to data on the abundance of birds on 46 forest patches. Since many of the patches were sampled repeatedly, a binomial model of the sampling process could be fit so that the counts of birds could be adjusted for imperfect detection. The model included covariates for the species-specific detection probabilities (sound power of bird vocalizations since detections were auditory) and for the species- and site-specific abundances (forest area). The most novel component of the model was that effects of species on other related species (e.g., those within the same feeding guild or taxonomic group) were included as random effects modulated by an adjacency matrix that includes measures of interspecific relatedness. The model is fit using a Bayesian approach and the abundances of each species or groups of species can be estimated, as well as the correlations among species within guilds or taxonomic groups. The estimated correlations can be used to make inferences concerning the species effects or dependence among species. We are currently finishing this paper and expect to submit it for publication in Ecology later this spring (Dorazio and Connor 2011). In addition, we expect two other papers to arise from this project. In the course of fitting the model, Dorazio determined that a class of conditionally autoregressive models commonly used in spatial lattice models suffers from non-identifiability problems. Dorazio is working with a non-working group member (Malay Ghosh University of Florida) to finish a paper that describes the set of constraints required for Bayesian learning of the non-identified parameters in this class of spatial models (Dorazio and Ghosh 2011). Finally, Dorazio and Connor anticipate a follow up paper to the Ecology paper that will deal with Bayesian methods of model selection. We anticipate submitting this paper later this year to a statistical journal.

Ecological Networks

Historically, food-web models rely upon stochastic process to generate food webs that are “similar” to empirical food webs. In this sense, “similar” tends to imply that the model-generated food webs exhibit similar properties, such as the number of basal species or top predators, degree distributions, intervality, etc. In contrast, we have modeled the probability of connections among predators and prey as a function of various empirical covariates. In particular, we have used log-linear models to quantify the contribution of body size and phylogeny to understanding of empirically-observed predator-prey interactions.

Probabilistic Niche Model

We are currently applying log-linear models in an effort to reproduce the results obtained from the “probabilistic niche model” (Williams et al. 2010) and to model the presence of linkages in pollinator networks. We are using both ML and Bayesian approaches to model fitting and including covariates that reflect the biology the species in the food-web. We have developed 7 models (6 based on degrees + niche model) and for each model we are attempting to perform model selection, as well as to estimate the distribution for several derived parameters that represent the properties of food webs. We have implemented models that produce the posterior distributions of many of the commonly used metrics of food web properties such as the numbers of top, basal, and intermediate species, food chain length, amount of omnivory and herbivory, and several other motifs. A presentation on this work will be given at the TIES meeting this summer, and a paper outlining our approach and results will be completed shortly thereafter (Allesina 2011). We developed Bayesian methods to determine the effects of taxonomic relatedness on food web structure (Eklof et al. 2011), and we are working on a Bayesian model to disentangle environmental effects on parasitoid-host relationships (Staniczenko et al. 2011). This work has involved the participation of three postdocs and one undergraduate student from the department of Ecology and Evolutionary Biology, University of Chicago (postdocs - Anna Eklof, Phillip Staniczenko, and Matthew Helmus; undergraduate student – M. Moore). Our work is also done in collaboration with Jason Tylianakis (Department of Biological Sciences, University of Canterbury, NZ).

Body Mass Ratios

In an attempt to understand the apparent similarity of the distribution of body-size ratio between predators and prey among different ecosystems, we have applied logistic models to the presence/absence of predator/prey links. We have used various aspects of the biology of the predator and prey such as their environment, metabolic category, feeding mode, and mobility as covariates. Our initial results suggest that these biological attributes of predator and prey strongly constrain the body-mass ratios of interacting species and are better predictors of the presence or absence of a link than knowledge of the species body-mass ratios. We anticipate completing a paper on this component of our work by summer (Stouffer 2011).

New collaborations

Robert Dorazio and Nick Gotelli

Nick Gotelli solicited Bob Dorazio's help with analyzing temporal trends in abundance data and in estimating species diversity from presence absence data. This collaboration has yielded one published paper, and a book chapter that is in press.

Gotelli, N.J., R.M. Dorazio, A.M, Ellison, and G.G. Grossman. 2010. Detecting temporal trends in species assemblages with bootstrapping procedures and hierarchical models. Phil. Trans. R. Soc. B 365:3621-3631.

Dorazio, R.M., N.J. Gotelli, and A.M. Ellison. 2011. Modern methods of estimating biodiversity from presence-absence surveys. In Biodiversity / Book 4, InTech (open access publisher).

Robert Dorazio and Malay Ghosh (Distinguished Professor of Statistics at Univ. Florida)

For our abundance based approach to modeling co-occurrence, we parameterized the species effects within genera using an adjacency matrix approach derived from the field of spatial statistics. Bob noticed that there appeared to be an unidentified parameter and working with Malay Ghosh was able to prove its existence and to suggest a means to identify all the parameters of this class of models. They are preparing a paper on this topic.

Dorazio, R.M. and Ghosh, M. 2011. Bayesian learning of non-identified parameters of conditionally autoregressive spatial lattice models. To be submitted to Bayesian Analysis.

Joshua Ladau

Ladau has established one new collaboration as a result of the working group, in addition to those with the participants of the working group. This new collaborative project is investigating the effects of invasive Argentine ants on the assembly processes of native ant communities in California.The project is drawing extensively on analyses using null models and loglinear models. My collaborators are: Lis Castillo Nelis (Stanford), Deborah Gordon (Stanford), Nathan Sanders (University of Tennessee), Katherine Fitzgerald (Stanford), Jessica Shors (Stanford), and Nicole Heller (Climate Central).

Literature Cited

Allesina, S. 2011. Inference in Food Webs based on Maximum Likelihood and Bayesian Approaches.

Eklof, A., Helmus, M., Moore, M. and Allesina, S., 2011. Relevance of evolutionary history for food web structure. Submitted.

Gotelli, N.J., R.M. Dorazio, A.M, Ellison, and G.G. Grossman. 2010. Detecting temporal trends in species assemblages with bootstrapping procedures and hierarchical models. Phil. Trans. R. Soc. B 365:3621-3631.

Dorazio, R.M., N.J. Gotelli, and A.M. Ellison. 2011. Modern methods of estimating biodiversity from presence-absence surveys. In Biodiversity / Book 4, InTech (open access publisher), *in press*.

Dorazio, R.M. and E.F. Connor. 2011. Estimating abundance-based patterns of species co-occurrences using guild structure, phylogenetic data and spatial covariates. To be submitted to Ecology.
Dorazio, R.M. and Ghosh, M. 2011. Bayesian learning of non-identified parameters of conditionally autoregressive spatial lattice models. To be submitted to Bayesian Analysis.
Dorazio, R.M. and E.F. Connor. 2011. Model selection in a hierarchical Bayesian model of abundance-based patterns of species co-occurrence.
Staniczenko, P., Tylianakis, J.M. and Allesina, S. 2011. Environmental Effects in Parasitoid Networks. In preparation.
Stouffer. D. 2011. What predator-prey body-mass ratios do and don’t tell us about food-web structure.
Williams, R.5, A. Anandanadesan, and D. Perves. 2010. The probabilistic niche model reveals the niche structure and role of body size in a complex food web. Plos One 5 (8): e12092.