Improving inference of metacommunity structure using multi-species occupancy models

Authors Joseph R. Mihaljevic[1*], Maxwell B. Joseph[1], Pieter T.J. Johnson[1]

[1] *Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, 80309*

[*] *Corresponding Author*

Abstract

Blah blah blah

Introduction

Broadly speaking, the metacommunity concept seeks to understand how spatial patterns of community composition emerge as a product of both local (e.g. competition) and regional (e.g. dispersal) dynamics (Leibold et al. 2004, Holyoak et al. 2005, Chase 2005). To date, metacommunity research can be separated into two nearly distinct categories: mechanism-based approaches and pattern-based approaches. The mechanism-based approach employs mechanistic modeling or controlled experiments to generate and test hypotheses related to, for example, the coexistence of competitors across local sites (Holyoak et al. 2005, Cottenie 2005, Urban and De Meester 2009, Economo 2011, Logue et al. 2011, Pillai et al. 2011, Carrara et al. 2012). The pattern-based approach uses an inverse process, attempting to relate empirically observed patterns of species occurrences across a landscape to hypotesized biological processes (Leibold and Mikkelson 2002, Presley and Willig 2010, Presley et al. 2010, López-González et al. 2012, Henriques-Silva et al. 2013, Meynard et al. 2013). While both approaches have advanced understanding of community structure and its underpinnings, methodological improvements could facilitabe a synthesis of these often disparate lines of research. Improving metacommunity analytic tools could therefore lead to a more complete understanding of community patterns and structuring mechanisms.

The pattern-based metacommunity approach, often referred to as the 'elements of metacommunity structure' (EMS) paradigm, utilizes observed occurrences of species among 'patches' of habitat in the metacommunity (e.g. field sampling sites). This data is typically compiled into a species-by-site incidence matrix and statistical analyses determine if the metacommunity exhibits any of twelve unique structures (Leibold and Mikkelson 2002, Presley et al. 2010). Based on the specific structure observed, inferences are made as to how the metacommunity assembles along dominant environmental axes. However, the efficacy of this pattern-to-process approach depends heavily upon the quality of the observed data [**(???)**]. For instance, problems with species detection could lead to inaccurate incidence matrices and, therefore, inaccurate assessments of metacommunity structure. Fortunately, decades of advancements in occupancy modeling have led to powerful methods to deal with problems such as species detectability. Integrating occupancy modeling with the pattern-based EMS paradigm should improve our ability to assess metacommunity structure and our ability to infer structuring processes.

Occupancy modeling relies on repeated sampling surveys to distinguish between the probability of a species occurring at a site and the probability of a species being detected at a site in which it occurs (MacKenzie et al. 2002). These models also facilitate estimation of covariate effects on detection and occurrence probabilities. Recently, dynamic multi-species occupancy models have been developed for use with longitudinal data sets to better estimate metacommunity composition, such as alpha, beta and gamma diversity metrics, as well as species-, site- and time-specific covariate effects on occurrence probabilities (Dorazio et al. 2010, Burton et al. 2012). Dorazio et al. (2010) specifically identify a closer union of their modeling framework with metacommunity theory as a priority in the field. While the integration of occupancy modeling with metacommunity theory is nascent, there is enormous potential for this to improve the study of metacommunity dynamics.

In this article, we will highlight some of the difficulties with current EMS methods and

emphasize how integrating EMS with occupancy modeling will help to resolve or ameliorate these issues. We will also illustrate unique advantages of occupancy modeling that will complement the EMS paradigm. First, we will begin by briefly reviewing the methods involved in the analysis of EMS and the structuring of multi-species occupancy models. Next, we will verbally outline various problems or advantages and couple these explanations with illustrative examples generated with simulated data sets. Our motivation for this article is to highlight the breadth of utilities gained by integrating these two approaches, rather than to explore any specific question or topic in full detail. We have also made our modeling and simulation code publicly available and fully annotated so that these methods can be appropriately and broadly utilized. We believe our approach will facilitate a more complete understanding of metacommunity structure and its underpinning mechanisms.

Elements of metacommunity structure

The EMS paradigm follows a step-wise procedure to determine which of twelve potential metacommunity structures are exhibited by a data set of species occurrences observed across multiple sites. Although the procedure can determine that no orderly structure exists (i.e. random structure), most metacommunities seem to exhibit detectable structure (Leibold and Mikkelson 2002). In this section we will only briefly describe these methods, as others have done a more thorough job elsewhere (Leibold and Mikkelson 2002, Presley et al. 2010).

First, species occurrence data are assembled into a site x species incidence matrix with rows as sites and columns as species. This matrix is then ordinated, typically using reciprocal averaging, to simultaneously group sites with common species assemblages and species with common distributions among sites. This ordinated incidence matrix theoretically represents how species assemblages are structured along a dominant environmental axis (i.e. gradient) across sites. From this matrix statistics are then calculated to summarize the elements of metacommunity structure.

The first and most important metric of metacommunity structure is coherence, which represents whether the majority of species in the metacommunity respond to the same enviornmental axis of variation as a cohesive unit, the foundation of structure. Coherence is estimated using the number of embedded absences in the ordinated matrix. Embedded absences occur in areas of the matrix where a species is absent in a site in which it would be expected to occur based on the ordination. The observed number of embedded absences are then compared to a null distribution of embedded absences generated from simulated matrices. When the metacommunity exhibits fewer embedded absences than expected from random chance (as defined by the chosen method of generating null matrices), the metacommunity is said to exhibit positive coherence. If the metacommunity has more embedded absences than expected, there is evidence for a checkerboard pattern. Finally, if the number of embedded absences is not significantly different from random, the metacommunity is not coherent and exhibits no discernible structure. This is often interpreted as meaning that species in the metacommunity respond to different dominant environmental gradients or that the size of the matrix (in terms of the number of sites or species or observed occurrences within the matrix) is too small to detect patterns.

If a metacommunity exhibits positive coherence, two more metrics are calculated to further describe the metacommunity structure, the first being turnover. Turnover represents how species composition changes along the theoretical environmental gradient, estimated using the number of species replacements observed in the ordinated matrix. Again, this observed number is compared to a distribution generated by null matrices. If the number of replacements is fewer than expected, which is indicative of nested subsets. If the number of replacements is greater than expected, the metacommunity exhibits positive turnover, which must be futher analyzed (below).

Boundary clumping, estimated with Morista's index, helps to further distinguish structures. If a metacommunity is positively coherent, has positive turnover, and there is positive boundary clumping, then the metacommunity is said to exhibit a Clementsian structure [**(???);(???)**]. Alternatively, if there is no discernible boundary clumping, the metacommunity is said to exhibit a Gleasonian structure, where species respond idiosyncratically to the environmental gradient[**(???)**]. Finally, if there is positive coherence, but no significant turnover, various quasi-structures are assigned to the metacommunity, depending on the trend observed in turnover and boundary clumping (*sensu* Presley et al. 2010).

After the metacommunity structure is discerned using the three metrics above, studies typically seek to determine which environmental covariates might be driving the primary axis of variation in the ordinated community. This analysis takes various forms. In most cases, the ordination score of the primary axis is extracted for each sampled site. Then univariate correlation tests are run for each covariate of interest, and subsequently researchers theorize how significantly correlated covariates might be responsible for structuring the metacommunity (**citations**). Another, related approach is to use, for example, canonical correspondence analysis to relate site compositions to multiple covariates simultaneously (**citations**). A more recent approach is to combine one of these previously discussed analyses with a variance partitioning analysis to evaluate the relative contribution of classes of covariates, such as 'local' and 'spatial' or 'abiotic' and 'biotic' (**citations**) to species compositions across sites.

Although these methods have yielded various insights into the structuring of metacommunities, at each step of this process, there are problems that can arise due to the methods themselves, such as type I error inflation, or due to inherent issues in data quality, such as detection errors. Below, we will briefly describe the structure of multi-species occupancy models, which estimate detection and occurrence probabilities and associate these probabilities with covariates, all within the same model. After this, we will highlight various issues with the EMS paradigm that can be ameliorated using occupancy modeling.

Multi-species occupancy models

Species occupancy models were developed to estimate a species' probability of occurring within a site while correcting for the fact that species may go undetected in a survey. Occupancy models use data from multi-observation surveys, which allows us to disentangle detection and occurrence probabilities to better estimate true species occupancy. More recently, these models have been extended to multi-species and multi-timepoint (longitudinal) surveys. In these models, species-, site- and time-specific estimates of detection probability, occurrence

probability and covariate effects can be estimated. Additionally, with longitudinal surveys, the probabilities of persistence at a site and colonization of previously unoccupied sites can be estimated. All of these models typically estimate parameters using Bayesian methods. Here, for simplicity, we will describe the structure of a single timepoint, multi-species occupancy model with mulitple observations at each site over the single time period.

Let $z_{i,k}$ represent the true occurrence of species i at site k, where $z_{1,2} = 1$ means that Species 1 is present at Site 2. This true occurrence is estimated as a Bernouli trial with probability, $\psi_{i,k}$, the probability of occurrence:

$$z_{i,k} \sim Bern(\psi_{i,k})$$

The probability of occurrence, $\psi_{i,k}$, can be related to any number of covariates as follows:

$$logit(\psi_{i,k}) = \alpha_{i0} + \alpha_{i1}X_{k1} + \cdots + \alpha_{iC}X_{kC}$$

, where C is the number of covariates, α_{ic} is the species-specific effect of covariate c, α_{i0} is the species-specific baseline occurrence probability, and X_{kc} is the site-specific value of covariate c at site k.

Multiple surveys are conducted at each site and compiled into a species-by-site matrix, Y. For example, if Species 1 is observed in five total surveys at Site 2, $Y_{1,2} = 5$. These observations, however, have inherent error in detectability. So, let $\rho_{i,k}$ represent the species- and site-specific probability of detection. Again, the probabilities of detection can be influenced by site-level covariates. These covariates can be the same as or unique from the covariates influencing occurrence proabilities. Here we assume the same covariates:

$$logit(\rho_{i,k}) = \beta_{i0} + \beta_{i1}X_{k1} + \cdots + \beta_{iC}X_{kC}$$

The observed occurrences are thus binomially distributed, influenced by both the detection and occurrence probabilities and the number of surveys conducted at each site, J_k:

$$Y_{i,k} \sim Binom(J_k, z_{i,k}\rho_{i,k})$$

Depending on how many sites were sampled and how many species were observed, estimating all of these species- and site-specific parameters can be computationally expensive. In order to optimize parameter estimation, metacommunity-level hyperparameters are estimated in a heirarchical Bayesian model. Thus, species- and site-level parameters are drawn from the metacommunity-level distribution of parameter values.

Difficulties with the EMS paradigm

Detection errors

A large issue with occurrence data is that these data inherently suffer from problems with detection, and the EMS paradigm so far has not attempted to account for this problem. Detection error can influence the ordination of the community matrix, the calculated EMS

metrics, and the accuracy of structural inference based on null matrices. For example, if species detection is imperfect, the calculated number of embedded absences from the ordinated matrix may be overestimated, which could lead to type II errors where metacommunities are incorrectly assigned random structures. Additionally, imperfect detection influences the form of null matrices, as most methods of null matrix generation utilize the raw data on row and/or column sums (**citation**). The occupancy modeling framework can circumvent this problem by estimating true occupancy of each species at each site, $z_{i,k}$.

With the Bayesian approach to occupancy modeling, Gibbs sampling is used to create a distribution of posterior estimates for each $z_{i,k}$. At each sampling iteration, the estimated $z_{i,k}$ values can be assembled into an incidence matrix, Z_{iter}. Each Z_{iter} can then be ordinated to provide a distribution of incidence matrices that represent estimates of true occupancy states. In this way, we can generate a posterior distribution of metacommunity structure, rather than a single estimate for a given dataset.

For example, we simulated occupancy of 12 species at 75 sites, assuming 4 re-sampling visits per site. For simplicity, we assume that base-line occurrence probabilities and detection probabilities are equal among species. In order to acheive coherence, we assumed that species' overall occupancy probability was influenced by a single continuous covariate, distributed normally among sites. Species-specific covariate effects were assigned with hyperparameters drawn from a normal distribution. Again, for simplicity, we assumed no covariate effects on detection probability. All code and documentation for this and all following simulations can be found at (**github link**). We used R (**citation**) and JAGS (**citation**), via the R package 'rjags' (**citation**), for all of our simulations and analyses.

After we verified model convergence, we randomly sampled 200 Z_{iter} from the full posterior sample. We chose to subsample the posterior for computational tractibility. We then ordinated each Z_{iter} and overlayed these ordinated matrices to generate a heat map (Figure 1). This heat map allows us to see the range of likely ordinated matrices, given that some species are likely to go undetected during data collection. Next, for each of the 200 posterior matrices, we calculated the three elements of metacommunity structure using 1000 null matrix simulations for each Z_{iter}. In this way, we were able to generate Bayesian posterior estimates for the elements of metacommunity structure (Figure 2). (**Generate pie-chart-esque figure that shows how often different structures were assigned to** Z_{iter})

This new approach to estimating elements of metacommunity structure allows us to more fully quantify and describe metacommunity structure, based on estimated true occurrence. This method will also allow us to fully explore the influence of detection on our ability to assign metacommunity structure. For instance, in our simulation, the ordinated matrix created with the raw data exhibits Gleasonian structure (positive coherence, positive turnover, no significant boundary clumping). However, *. . . MORE. . .*

Covariate effects

Another issue with the EMS strategy that can be emeliorated by a merger with occupancy modeling is the way in which covariate effects are explored. As discussed above, many researchers extract the first ordination axis score for each sampled site and associate these scores to any number of site-specific covariates using multiple univariate correlation tests. This approach is flawed due to inflation of the overall type I error rate caused by conducting

multiple tests. Additionally, regardless of the type of statistical test, using the ordination scores from a single axis could be leaving out valuable information, and these ordination scores are extracted from a matrix that is most likely flawed by detection error in the first place. Occupancy modeling helps to remedy these issues by estimating species-specific covariate effects on occurrence probability based on the raw data, not ordination scores. Then, using Bayesian model selection and model averaging, dominant covariates and their effects can be estimated without the need for multiple tests.

The influence of confounding covariate effects on detection and occurrence probabilities can also be explored within the occupancy modeling framework. For example, if a single covariate, such as vegetation density, has positive effects on species' occupancy but negative effects on species' detectability, how might this influence our ability to detect or accurately assign metacommunity structure? The simulation method described above could be used to explore this question in a future study by altering the correlation between the covariate's effect on each species' detection and occurence probabilities and observing how metacommunity structure changes.

Complementary features of occupancy modeling

Structuring mechanisms

While introducing the EMS methods in their seminal paper, Leibold and Mikkelson (2002) emphasize that the EMS paradigm can identify patterns, but cannot necessarily elucidate the processes that lead to pattern. However, multi-species occupancy modeling can complement the features of the EMS paradigm and help to better elucidate metacommunity structuring mechanisms. For example, metacommunities that exhibit Gleasonian and Clementsian structure are assumed to be structured differently based on species-specific responses to a dominant environmental covariate(gradient). It is often stated that Gleasonian structure arises from idiosyncratic species responses, whereas Clementsian structure arises from groups of species that respond similarly to each other, but differently from other groups of species in the metacommunity. Alternatively, Clementsian structure could arise from negative associations between species pairs or groups that arise along the gradient (**???**). Still, these mechanistic interpretations of structure remain more or less speculative in the EMS paradigm. The occupancy modeling simulation framework that we described above can help us to understand under which conditions different metacommunity structures might arise.

We used simulation to explore how the distribution of species-specific covariate effects and the distribution of covariate values observed among sampled sites affect the resulting metacommunity structure. Again we simulated 12 species distrubted across 75 sites, with 4 re-sampling surveys at each site. Base-line occurrence and detection probabilites were fixed across species. We, however, altered the distribution and range (hyperparameters) from which the species-specific covariate effects and site-specific covariate values were drawn (Tables 1 and 2). For each set of hyperparameter values, we simulated 10 incidence matrices, ordinated each one and determined the resulting metacommuinty structure based on 1000 null matrices. Tables 1 and 2 thus show all the metacommunity structures observed over 10 simulations for each set of conditions.

Importantly, we found that the range and distribution of site-specific covariate values sampled across the landscape can alter the observed metacommunity structure, independent of the species-specific covariate effects (Table 1). In other words, the selection of sampling sites in a study could alter the observed metacommunity structure in a way that may reflect statistical phenomena rather than biological ones, complicating inferences about structuring mechanisms. Furthermore, with our chosen set of parameter values, we were overall more likely to observe Clementsian structure than Gleasonian structure, even when species-specific covariate effects were drawn from a broader distribution and were essentially idiosyncratic (Table 1). Additionally, we were able to produce nested structures when we forced covariate values to be only positive (or negative), even though covariate effects did not change from the Gleasonian/Clementsian examples (Table 2).

Our findings above suggest that in some circumstances metacommunity structure might arise from statistical phenomena in addition to or in place of biological processes. Similar to our findings, (**???**) explored the 'feasible set' of species area distributions (SAD) as constrained by total species richness and total abundance of species, showing that the ubiquitous hollow-curve SAD may be the result of a statistical phenomenon. Admittedly, our simulations did not attempt to encompass the entire feasibility space of the occupancy model, as this would require altering many parameters. However, this occupancy modeling simulation framework could be used as a type of sensitivity test to help us understand when structures might be generated by sampling issues versus biological processes.

Our simulations are simple in many regards and future studies could more fully explore the sensitivity of metacommunity structure to various model assumptions. One important extension would be to determine the effects of discrete versus continuous covariates and the resulting patterns when both types of covariates influence occurrence probabilities. Additionally, we held our metacommunity size constant, in terms of the number of species and the number of sites sampled. It is well established that matrix size influences the power to detect patterns using null models (**citations e.g. Gotelli**), so understanding the influence of these parameters in occupancy models will be important. For tractability, we unrealistically assumed that base-line occurrence and detection probabilities were fixed among species. Understanding how the distributions of these probabilities (and their species-specific covariance) influences overall metacommunity structure should be a leading question.

Temporal dynamics

With longitudinal metacommunity data becoming more and more available, an interesting question that arises is how does metacommunity structure vary across time points and what are the driving mechanisms for these changes. For example, (**???**) explored how metacommunity structure of stream fishes changed over time and how this structure was influenced by environmental covariates. Dynamic occupancy modeling has a few complementary features that could aid in a more cohesive approach to studying such questions.

Occupancy modeling could be used to distinguish if changes in metacommunity structure are more related to changing covariate values at sites over time, whether the dominant structuring covariate changes identity over time (e.g. from altitude to pH), or if metacommunity structure changes due to idiosyncratic changes in how species respond to the same covariates over time. **As a simple example, we simulated...**

Additionally, Dorazio et al. (2010) showed how multi-species occupancy modeling can be used to estimate species- and time-specific covariates on detection, occurrence, colonization and persistence probabilities.

References

Burton, A. C., M. K. Sam, C. Balangtaa, and J. S. Brashares. 2012. Hierarchical Multi-Species Modeling of Carnivore Responses to Hunting, Habitat and Prey in a West African Protected Area. PloS one 7.

Carrara, F., F. Altermatt, I. Rodriguez-Iturbe, and A. Rinaldo. 2012. Dendritic connectivity controls biodiversity patterns in experimental metacommunities. Proceedings of the National Academy of Sciences of the United States of America 109:5761–5766.

Chase, J. M. 2005. Towards a really unified theory for metacommunities. Functional Ecology 19:182–186.

Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community dynamics. Ecology letters 8:1175–1182.

Dorazio, R. M., M. Kéry, J. A. Royle, and M. Plattner. 2010. Models for inference in dynamic metacommunity systems. Ecology 91:2466–2475.

Economo, E. P. 2011. Biodiversity conservation in metacommunity networks: linking pattern and persistence. The American naturalist 177:E167——80.

Henriques-Silva, R., Z. Lindo, and P. R. Peres-Neto. 2013. A community of metacommunities: exploring patterns in species distributions across large geographical areas. Ecology 94:627–639.

Holyoak, M., M. A. Leibold, and R. D. Holt. 2005. Metacommunities: spatial dynamics and ecological communities. Page 513. University of Chicago Press.

Leibold, M. A., and G. M. Mikkelson. 2002. Coherence, species turnover, and boundary clumping: elements of meta-community structure. Oikos 97:237–250.

Leibold, M. a., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and a. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601–613.

Logue, J. B., N. Mouquet, H. Peter, and H. Hillebrand. 2011. Empirical approaches to metacommunities: a review and comparison with theory. Trends in ecology & evolution 26.

López-González, C., S. J. Presley, A. Lozano, R. D. Stevens, and C. L. Higgins. 2012. Metacommunity analysis of Mexican bats: environmentally mediated structure in an area of high geographic and environmental complexity. Journal of Biogeography 39:177–192.

MacKenzie, D., J. Nichols, G. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

Meynard, C. N., I. Boulangeat, L. Garraud, N. Mouquet, and W. Thuiller. 2013. Disentangling the drivers of metacommunity structure across spatial scales:1–12.

Pillai, P., A. Gonzalez, and M. Loreau. 2011. Metacommunity theory explains the emergence of food web complexity. Proceedings of the National Academy of Sciences of the United States of America 108:19293–19298.

Presley, S. J., and M. R. Willig. 2010. Bat metacommunity structure on Caribbean islands and the role of endemics. Global Ecology and Biogeography 19:185–199.

Presley, S. J., C. L. Higgins, and M. R. Willig. 2010. A comprehensive framework for the evaluation of metacommunity structure. Oikos 119:908–917.

Urban, M. C., and L. De Meester. 2009. Community monopolization: local adaptation enhances priority effects in an evolving metacommunity. Proceedings. Biological sciences / The Royal Society 276:4129–4138.