# 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 distribution of species 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 structuring mechanisms (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 can be developed to better integrate these 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, relies on 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 procedures are run to 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 that 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 the associated inferences about 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 facilitate estimation of the effects of covariates on detection and occurrence probabilities. Recently, dynamic multi-species occupancy models have been developed for use with longitudinal data sets to better estimate aspects of metacommunity structure, 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). For instance, Dorazio et al. (2010) also 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 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 assessment of EMS with occupancy modeling will help to resolve 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 figures 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 great 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 lead to a more complete exploration of metacommunity structure and its underpinning mechanisms from occurrence data sets.

## 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. In this section we will only briefly describe the methods used to identify metacommunity structure within the EMS paradigm, as these methods have been detailed extensively elsewhere (Leibold and Mikkelson 2002, Presley et al. 2010).

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 dominant environmental gradient 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 distribution of embedded absences generated from simulated null matrices. When the metacommunity exhibits fewer embedded absences than expected from random chance (as defined by the method of simulating 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 with simulated null matrices. If the number of replacements is fewer than expected, the metacommunity exhibits negative turnover, which is indicative of nested subsets. If the number of replacements is grater than expected, the metacommunity exhibits clumps of species. Finally, if there is no significant turnover, various quasi-structures are assigned to the metacommunity, depending on the outcome of the final metacommunity structure metric, boundary clumping (*sensu* Presley et al. 2010).

Boundary clumping, estimated with Morista's index, helps to distinguish between clumped structures or helps to determine the way nested subsets are organized. If a metacommunity is positively coherent and clumped, and there is positive boundary clumping, the metacommunity is said to exhibit a Clementsian structure. This means that distinct species groupings emerge along the gradient (e.g. intermediate and climax communities). In this case, the ordinated structure may be the result of multimodal species-specific responses to an environmental gradient. In other words, subsets of species respond similarly as one another to the gradient, but differently from other subsets of species. 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, for example, as a unimodal distribution of responses. Negative boundary clumping is indicative of evenly spaced species assemblages.

If a metacommunity is positively coherent and exhibits nested subsets, and there is positive boundary clumping, the metacommunity shows clumped species losses along the environmental gradient. For example, a plant assemblage may show subsets along an environmental gradient that crosses distinct biomes, leading to clumped assemblages (Presley et al. 2010). If there is no evidence for boundary clumping, this is indicative of random species losses along the gradient. Finally, if there is negative boundary clumping, this shows evidence for hyperdispersed (evenly spaced) species losses. In this case, species loss is more or less predictable along the environmental gradient.

After the metacommunity structure is discerned using these three metrics, studies typically seek to determine which environmental covariates might be the responsible structuring gradient. This analyses takes various forms. In most cases, the ordination score for each sampled site is extracted from the incidence matrix. Then correlation coefficients are calculated 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

ordination scores 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**).

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 II 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. Next, we will highlight various issues with the EMS that can be ameliorated using occupancy modeling and then describe the unique features of the models that can complement the EMS paradigm.

## 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 there is inherent error in detection, and therefore the species may be present but go undetected in a survey. To overcome this hurdle, occupancy models using data from multi-observation surveys can disentangle detection and occurrence probabilities to better estimate species presences and absences. 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 estimate parameters in the Bayesian framework. 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 (e.g. multiple surveys within one month). This same model will be used for all of our subsequent simulations and analyses.

Let $z_{i,k}$ represent the true occurrence of species $i$ at site $k$, where $z_{1,1} = 1$ means that Species 1 is present at Site 1. 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, $\alpha_{ic}$ is the species-specific effect of covariate $c$, $\alpha_{i0}$ is the species-specific baseline occurrence probability, and $X_{kc}$ is the value of covariate $c$ at site $k$.

Observations of each species at each site are accumulated over a series of observations within the given time-frame. 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. 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:

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

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 are 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 framework. Thus, species- and site-level parameters are drawn from the metacommunity-level distribution of parameter values. In all of the follwoing analyses, we used R (**citation**) and JAGS (**citation**), via the R package 'rjags' (**citation**).

## Difficulties with the EMS paradigm

*Detection errors*

A large assumption that the EMS strategy makes is that the collected data do not suffer from problems with detection. However, presence/absence data are inherently flawed by observation error. This observation or 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 construction of null matrices. Depending on the method used to generate these null matrices, type I or type II errors are of concern (**citation**), and detection problems can enhance these issues. 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 $z_{i,k}$. These $z_{i,k}$ values can be assembled into an incidence matrix for each sampling iteration, $Z_{iter}$. Each $Z_{iter}$ can be ordinated to provide a distribution of incidence matrices that represent estimates of true occupancy states. In this way, one can generate a distribution of metacommunity structure, rather than a single estimate for a given dataset.

For example, we simulated occupancy of 10 species at 50 sites using known metacommunity-level hyperparameters for species-specific detection and occurrence probabilities. In order to acheive coherence, we assumed that species' occupancy responded to a single continuous covariate, with species-specific effects assigned with hyperparameters, forming a unimodal

distribution of covariate effects. For now we assumed no covariate effects on detection probability. All code and documentation for this and the following simulations can be found at (**github link**).

In Figure 1, we show the ordinated matrix of raw data juxtaposed to a heat map that represents realizations of the ordinated $Z_{iter}$ matrices. For computational tractability, we randomly sampled 100 $Z_{iter}$ from the full posterior sample. The heat map allows us to see the range of likely ordinated matrices, given that some species are likely to go undetected during some surveys. Next, for each of the 100 posterior matrices, we ordinated the matrix and then calculated the three elements of metacommunity structure using 1000 null matrix simulations. In this way, we were able to generate Bayesian posterior estimates for each element of structure (Figure 2). (**Maybe a figure that shows how average detection prob. influences variance of EMS estimates?**). This approach allows us to fully understand the influence of detection on our ability to assign metacommunity structure and therefore understand the extent of type I and type II errors.

*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 ordination scores for sites along the first ordination axis and associate these scores to any number of covariates using univariate correlation tests. This approach is flawed due to the inflation of the overall type I error rate caused by conducting so multiple tests. Additionally, regardless of the type of statistical test, using the ordination scores from a single ordination axis could be leaving out valuable information, and these ordination scores are extracted from a matrix that is most likely flawed by detection error. Occupancy modeling remedies these issues by estimating species-specific covariate effects on occurrence probability based on the raw data (used to estimate true occurrence), not ordination scores. Then, using Bayesian model selection and model averaging, dominant covariates and their effects can be estimated without the need of multiple tests.

Additionally, the influence of confounding covariate effects on detection and occurrence probabilities can 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? (**Leave this as a question or do some analysis?**)

## Complementary features of occupancy modeling

*Structuring mechanisms*

In more ways, multi-species occupancy modeling can complement the features of the EMS paradigm and help to better elucidate metacommunity structuring mechanisms. For example, the structuring mechanisms assigned to metacommunities that exhibit Gleasonian and Clementsian structure differ based on the species-level responses to the dominant environmental covariate. It is often stated that Gleasonian structure arises from idiosyncratic responses to a dominant environmental covariate, whereas Clementsian structure arises from groups of

species that respond similarly to an environmental gradient, although differently from other groups of species. In the occupancy modeling framework, where species-specific covariate effects are estimated, one could distinguish between these two mechanisms more formally by observing the distributions of covariate effects among species. If a clear multimodal distribution is estimated, this would be evidence for Clementsian structure indeed arising from groups of species responding similarly.

In order to demonstrate this idea, we...

Additionally, when a nested metacommunity structure is present, the extent of boundary clumping is used to infer how species are lost along the environmental gradient. For example, when boundary clumping is signicantly positive, this could mean that discrete environmental variables, such as distinct biomes, may also be influencing species composition. With occupancy modeling, the effects of both continuous and discrete enviromental covariates can be estimated simultaneously to explore their influence on metacommunity structure.

We simulated two different nested communities: one influenced by a dominant continuous covariate, and the other influenced by this covariate plus a discrete covariate with three levels (e.g. a metacommunity that spans three biomes)...

*Temporal dynamics*

With more longitudinal metacommunity data becoming available, an interesting question that arises is how does metacommunity structure vary across time points and which are the driving mechanisms for these changes. For example, (**???**)

# 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.