

1 **A multi-region community model for inference about geographic**
2 **variation in species richness**

3 Chris Sutherland^{1,†}, Mattia Brambilla^{2,3}, Paolo Pedrini², and Simone Tenan^{2,*,†}

4 ¹ New York Cooperative Fish and Wildlife Research Unit, Department of Natural Resources, Cornell
5 University, Ithaca, New York, USA;

6 ² MUSE - Museo delle Scienze, Vertebrate Zoology Section, Corso del Lavoro e della Scienza 3, I-38122
7 Trento, Italy

8 ³ Fondazione Lombardia per l'Ambiente, Settore biodiversità e aree protette, Largo 10 luglio 1976 1,
9 I-20822 Seveso (MB), Italy

10 * E-mail: Corresponding `simone.tenan@muse.it`

11 **Running title:** A multi-region community model

12 **Word count:** XXXX

13

† These authors equally contributed to this work.

Abstract

An enduring challenge in ecology is to understand what drives spatial variation in the size and structure of communities. The ability to count the number of species present at a location is hindered by the fact that not all species are equally detectable, and invariably some go completely undetected. This makes comparing species richness across distinct spatial units (or regions) problematic as sources of error are usually unaccounted for in simple enumerations of species. Multi-species occupancy models explicitly incorporate a model for this observation uncertainty and provide a framework for estimating community size when detection is imperfect. Currently, however, the model is restricted to estimating the number of species at only a single region of interest. In this paper we extend the multi-species occupancy model to accommodate data collected across multiple regions of interest (e.g., reserves or biomes) allowing for simultaneous estimation of region specific community size. Moreover, using this approach species richness can be modeled as a function of region specific covariates providing a mechanism for testing hypotheses about why and how species richness varies in space. Here, we first demonstrate the value of the integrated multi-region approach using simulations to compare model performance to more traditional two-step approaches for modelling spatial variation in species richness. Then, applying the model to data collected from eight avian communities in northern Italy, we show how species richness varies in space as a function of habitat complexity.

Key-words: *Bayesian analysis, Biodiversity, Biogeography, Community structure, Data augmentation, Geographic variation, Site occupancy models, Species richness.*

Introduction

For decades, ecologists have been interested in geographic variation in the size and structure of communities (MacArthur & Wilson 1967; Stevens 1989; Field *et al.* 2009), and maximizing the number of species in a region of interest continues to be the central focus of many conservation management strategies (May 1988; Kerr 1997; Myers *et al.* 2000). Species richness is therefore an important state variable, and under-

standing what determines whether regions are species-rich or species-poor remains an active area of ecological research (Purvis & Hector 2000; Field *et al.* 2009). A major challenge associated with estimating the total number of species present in a region is that not all species are observed with equal probability and, as a consequence, some species go undetected (Boulinier *et al.* 1998). When species detection is imperfect, comparisons of community size and structure across multiple regions to test theoretical predictions about species richness using species counts can be difficult or even misleading (Boulinier *et al.* 1998; Nichols *et al.* 1998; Gotelli & Colwell 2001; Cam *et al.* 2002).

Traditional methods for investigating geographic variation in species richness typically adopt a two-stage approach whereby species richness is first estimated, then used as data in subsequent analyses (e.g., Field *et al.* 2009). This approach makes it difficult to account for statistical uncertainty in parameter estimates (Link 1999; Royle & Dorazio 2008; Brooks *et al.* 2015). Therefore, the ability to model geographic variation in community size, structure, and composition within a single framework should be of great interest. For example: evaluating the performance of alternative reserve design or management practices relative to conservation targets (Cabeza & Moilanen 2001), assessing link between biodiversity and ecosystem function and service (Balvanera *et al.* 2006), or testing long standing theories of island biogeography such as the species-area relationship (MacArthur & Wilson 1967).

The recent development of multi-species occupancy (or community) models (Dorazio & Royle 2005; Dorazio *et al.* 2006; Royle *et al.* 2007; Iknayan *et al.* 2014) provides a hierarchical framework that, by accounting for imperfect detection, produces estimates of site occupancy for multiple species simultaneously, including those never detected. Using data augmentation, the number of species present in the community that were unobserved (i.e. that had ‘all-zero’ encounter histories) can be estimated (Dorazio & Royle 2005; Dorazio *et al.* 2006; Royle *et al.* 2007, see methods), thus providing a direct estimate of total species richness of a community (Kéry & Royle 2008). The multi-species occupancy model is useful for estimating community size at a single spatial unit of interest (which we refer to as a ‘region’ from here), although, the inability to simultaneously model species richness across multiple communities precludes the formal testing of hypotheses about drivers of spatial variation in species richness across multiple regions.

In this paper we extend the ‘single-region community model’ (SRCM) to allow for direct and simultaneous estimation of species richness across communities from multiple regions. We first describe a general

67 ‘multi-region community model’ (MRCM), an extension of the community model that allows species rich-
68 ness to vary by region as a region-specific random effect (see also Tobler *et al.* 2015). We then describe
69 an important extension of the model that allows variation in species richness to be modeled as a function
70 of region-level covariates, thus providing a framework for investigating drivers of geographic variation in
71 the size and structure of communities. We demonstrate this novel model development and compare its
72 performance with that of a two-step approach using both simulated data, and then actual point count data
73 collected from eight geographically distinct avian communities.

74 **Methods**

75 **A multi-region model for species richness**

76 As a starting point, we consider data collected for analysis of a Bayesian multi-species occupancy model
77 (Dorazio & Royle 2005), where y_{ij} represents the encounter frequencies of species $i = 1, \dots, n$ at each of
78 $j = 1, \dots, J$ sites, which are visited $k = 1, \dots, K$ times. The data can be formatted as a 2-dimensional
79 $n \times J$ matrix \mathbf{Y} , and are (potentially imperfect) observations of the corresponding 2-dimensional $N \times J$
80 matrix \mathbf{Z} , the true but unknown occupancy states of each species in each site, z_{ij} (Fig. 1). Here, n is
81 the total number of species *observed*, N is the total community size, and thus $N - n$ is the number of
82 species in the community *never* detected, which, by definition, have ‘all zero’ encounter histories. The
83 inference objective is to estimate the ‘true’ occupancy states, \mathbf{Z} , by accounting for imperfect detection
84 which is estimated using detection data from repeated sites visits, i.e. from \mathbf{Y} . Sites are assumed to be a
85 representative sample of a larger, single geographic area for which species-specific occupancy and overall
86 species richness are of interest, which is to say there is $R = 1$ region.

87 We extend this single-community model to accommodate data collected at multiple spatially indepen-
88 dent regions, i.e. $r = 1, \dots, R$. This provides some distinct practical benefits; specifically, the integration
89 of data collected across geographic gradients and scales into a single analytical framework, e.g. reserve
90 networks (Sierra *et al.* 2002), biogeographic regions (Dorazio *et al.* 2010) or, more generally, sites where
91 communities of interest are sampled and compared at a continental or global scale (Ahumada *et al.* 2011).

Perhaps more importantly though, is that doing so allows information to be shared across regions as well as across sites and species as is the case with traditional community occupancy models (Kéry & Royle 2008), and permits formal comparison of community size and structure across multiple strata within a single analytical framework.

Multi-species encounter frequency data from multiple regions are summarized in a 3-dimensional array \mathbf{Y} with elements y_{ijr} where the subscript r now indexes region (Fig. 1). The matrix of true occupancy states \mathbf{Z} is also extended to 3 dimensions containing the elements z_{ijr} , the species-by-site occupancy states in each region (Fig. 1). For convenience, we assume that J , the number of sites visited, and K , the number of site visits, are constant across regions and sites, although this need not be the case. A hierarchical model for \mathbf{Z} is formulated such that the site- and region-specific occupancy states of each species are Bernoulli distributed binary state variables ($z_{ijr} = 1$ is occupied, and $z_{ijr} = 0$ is empty):

$$z_{ijr} \sim \text{Bern}(\psi_{ijr}\omega_{ir}), \quad (1)$$

and the observation model relates the truth, i.e. \mathbf{Z} , to the data such that

$$y_{ijr} \sim \text{Bin}(K_{jr}, p_{ijr}z_{ijr}). \quad (2)$$

Parameter, ψ_{ijr} is the species-specific occurrence probability for each site in each region, and p_{ijr} is the corresponding detection probability. ω_{ir} is a species-specific indicator variable denoting whether the species is present in the region (see below). We note that both p and ψ are indexed by i , j and r meaning it is possible to model these parameters as species-, site- or region-level fixed or random effects, or using species-, site- and region-specific covariates.

A popular and convenient way to estimate the number of unobserved species in a community is by data augmentation (Dorazio & Royle 2005; Dorazio *et al.* 2006; Royle *et al.* 2007). The approach assumes a $\text{Uniform}(0, M_r)$ prior for the ‘true’ number of species present in each community, N_r , where the choice of M_r is arbitrary, and for convenience, is kept equal across regions, i.e., $M_r \equiv M$, but must also be larger than the total number of species in the largest community, i.e., $M \gg \max(N_1, \dots, N_R)$ (Fig. 1). Alternatively,

114 if the species pool is known, as is the case for well-studied taxa and areas, the model can be conditioned
 115 on the known ensemble of species (e.g., Tobler *et al.* 2015).

116 The data are then augmented with $M - n$ ‘all-zero’ encounter histories (Fig. 1), and the aim is to
 117 estimate the proportion, Ω , of these represent species that exist in the community but that were not detected
 118 (i.e., that were sampling zeros and not structural zeros):

$$\omega_{ir} \sim \text{Bern}(\Omega_r), \quad (3)$$

119 where ω_{ir} is the species-specific indicator variable in Eq. 1 denoting whether species i was present in the
 120 r^{th} community ($\omega_{ir} = 1$) or whether it is a structural zero ($\omega_{ir} = 0$ and therefore $z_{ir} = 0$). For species that
 121 were observed in a region, $\omega_{ir} = 1$.

122 Extending the model to accommodate multiple regions allows simultaneous estimation of region spe-
 123 cific Ω (i.e. Ω_r). Recognizing that $E(N_r) = M\Omega_r$ (or alternatively, $N_r = \sum_{i=1}^M \omega_{ir}$), data augmentation
 124 thus converts the problem of estimating region-specific species-richness, N_r , to that of estimating the
 125 region-specific zero-inflation parameter, Ω_r .

126 **Modeling regional variation in species richness**

127 Given the development of a formal derivation of the expected number of species in a region, N_r , the
 128 multi-region framework lends itself to fairly straightforward extensions for explicitly modeling the ef-
 129 fect of region-specific covariates on species-richness. Recognizing that the model described above is an
 130 ‘intercept-only’ model, it can easily be extended to a logit-linear model including covariate(s) X_r , e.g.,
 131 $\text{logit}(\Omega_r) = \alpha_\Omega + \beta_\Omega X_r$. We use a single covariate here without loss of generality. A canonical exam-
 132 ple of a covariate affecting the total number of species would be the size of the region, e.g. species-area
 133 relationship (MacArthur & Wilson 1967) or habitat complexity/heterogeneity (Johnson *et al.* 2003).

134 **Demonstration by simulation**

135 To demonstrate, and to assess the performance of the multi-region community model we simulated multi-
 136 species occupancy data for multiple regions where species richness was generated as a function of a

region-specific covariate, X_r , drawn from a Uniform(-1,1) distribution. To explore the performance of our model in relation to variation in data quality, we simulated data for a **low** ($R = 6$) and **high** ($R = 15$) number of regions, and for a **low** ($J = 25$) and **high** ($J = 50$) number of sites sampled per region. To explore model performance in relation to variability in community-level detectability, we simulated data using **low** ($\bar{p} = 0.3$) and **high** ($\bar{p} = 0.6$) mean species detectability (\bar{p}), but also **low** ($\sigma_p^2 = 0.5$) and **high** ($\sigma_p^2 = 1.5$) community level detection heterogeneity (σ_p^2). Of particular interest was the ability of the model to estimate the effect of covariates on species richness, so, keeping the intercept of the relationship between species richness and covariate X_r constant ($\alpha_\Omega = -0.8$), we simulated data using **no** ($\beta_\Omega = 0$), **moderate** ($\beta_\Omega = 0.4$), and **strong** ($\beta_\Omega = 0.8$) covariate effects (remembering that $\text{logit}(\Omega_r) = \alpha_\Omega + \beta_\Omega X_r$, and $E(X) = 0$). For completeness, we simulated data under each combination of these parameter settings resulting in a total of 48 simulation scenarios. We used a single community-level mean occupancy ($\bar{\psi} = 0.3$ with, on the untransformed scale, $\sigma_\psi^2 = 1.0$).

For each scenario, we simulated 144 multi-region community data sets and analysed each using the proposed multi-region community model using Markov chain Monte Carlo, or MCMC (Robert & Casella 2004). We modelled probabilities on the normal scale such that $\text{logit}(\theta) = \mu_\theta$. We specified Normal(0,100) prior distributions for parameters μ_{α_Ω} and μ_{β_Ω} , Normal(0,2.25) distributions for μ_ψ and μ_p , and Gamma(0.1,0.1) distributions for precisions $1/\sigma_\psi^2$ and $1/\sigma_p^2$. The results presented below are based on 50,000 samples from the post-burn-in (5,000 iterations) posterior distribution of model parameters. We retained the posterior mean value of each parameter from each simulation, and recorded whether the Bayesian credible intervals of each parameter overlapped the true value (coverage form here). We report the mean and standard deviation of the means across all simulations and the proportion of simulation credible intervals that contained the true value.

Typically, the approach to modelling spatial variation in species richness involves estimating the number of species present in several regions independently (stage one), then using these point estimates for inference about covariate relationships. Therefore, in order to justify our model development, we contrasted our proposed model with the two-step procedure. To do so, we analysed the data from each region separately using single region multi-species occupancy models (SRCM), and then regressed species-richness point estimates (medians) against the community-specific covariate using a Generalized Linear Models.

165 Note that in order to compare across models, we used the estimated data augmentation parameters Ω_r in
166 the regression models. As with the MRCM, we recorded the maximum likelihood estimate of each pa-
167 rameter from the post-hoc regression, as well as recording whether the 95% confidence interval spanned
168 the true parameter values. The results for the SRCM are based on the estimated regression coefficients
169 obtained using the mean parameter estimates from 10,000 post-burn-in (5,000 iterations) posterior sam-
170 ples as data in the regression. We also recorded whether each 95% confidence interval overlapped the true
171 value.

172 Models were fitted using JAGS (Plummer 2003), called from R (R Core Team 2012) using the packages
173 `rjags` (Plummer 2013) and, for parallelization, `snowfall` (Knaus 2013). A detailed description of the
174 model and the simulations study are provided in the supplemental material along with the JAGS model
175 code.

176 **Simulation Results**

177 Overall, there was little difference in the bias (simulation mean) or precision (simulation standard devi-
178 ation) between the two approaches for both α_Ω and β_Ω , and in general both appeared to perform well
179 in retrieving the data generating parameter values (Table 1). The major difference was the fact that the
180 MRCM approach provided substantial improvements in coverage ($\sim 5 - 10\%$) when compared to the
181 two-stage regression approach (Table 1, Figure 2). The MRCM achieved the nominal 95% coverage for
182 both parameters in almost every scenario, whereas the SRCM rarely (4 out of 48 scenarios) achieved that
183 level (Table 1). This suggests that our integrated modelling approach benefits greatly from cross-region
184 parameter sharing, and importantly, that this improvement was particularly evident in low data quality
185 scenarios (only 6 regions and only 25 sites, Fig. 2).

186 **Application: Avian communities and habitat heterogeneity**

187 Having evaluated the performance of the model via simulation, we provide an application of the model
188 to detection-non detection data collected from bird assemblages in $R = 8$ geographically distinct ar-

189 eas (regions) in northern Italy. Habitat composition in each region varied and the initial motivation for
 190 the sampling was to evaluate the effect of such heterogeneity on species richness (Padoa-Schioppa *et al.*
 191 2006). Habitat complexity ranged from only grassland and woodland, to habitats made up of woodland,
 192 grassland, wetland, agricultural and urbanized areas. We characterised habitat heterogeneity using the
 193 Shannon entropy index (SEI, Shannon 1948), and modelled the relationship between species richness and
 194 SEI: $\text{logit}(\Omega_r) = \alpha_\Omega + \beta_\Omega \text{SEI}_r$. We expected a positive relationship between habitat complexity and
 195 species richness. SEI for each region was calculated for the area defined by the minimum convex polygon
 196 encompassing all surveyed points plus a 100-m buffer around it.

197 In each region, incidence records of all species observed within a 100-m radius of the observed during
 198 multiple 10-minute count periods was recorded. The number of point count locations in each region
 199 ranged from a minimum of $J = 11$ to a maximum of $J = 103$ (median: 31.5), and the number of sampling
 200 occasions (10-minute point counts) ranged from $K = 2$ to $K = 10$ (median: 3).

201 We analyzed these data using the MRCM approach and report results based on 6000 samples from the
 202 posterior distribution of the parameters (3 chains, thinning of 50, and burn-in of 100,000). We augmented
 203 the data for all regions such that $M = 200$, i.e., by $200 - n_r$ species for each region. As above, we
 204 compared the estimated regression coefficients for species richness from the MRCM with those obtained
 205 using the two-stage post-hoc regression approach (SRCM). Although we do not know truth in this case, it
 206 is somewhat instructive to compare the results of the two models in light of the simulation study.

207 **Example application**

208 The number of *observed* species across the eight regions varied between 29 and 75 species (median 49.5).
 209 The estimated species richness from the multi-region model ranged from 66 (95% Bayesian CI: 53 – 86)
 210 to 93 (95% BCI: 82 – 110; Table 2, Fig. 3a). In comparison, estimated species richness from the two-stage
 211 approach of analysing estimates from the independent single community models (SRCM) ranged from
 212 55 (95% CI: 41 – 114) to 121 (95% CI: 91 – 184; Fig. 3c). Both models suggest that species detection
 213 was imperfect. Species richness estimates from the multi-region model had narrower credible intervals
 214 compared to the traditional single-region model approach (Fig. 3a and 3c).

Under the multi-region framework, the size of the eight communities was positively related to habitat complexity, and the coefficient for the relationship was different from zero ($\beta_{\Omega_{MRCM}} = 0.475$, 95% BCI: 0.017 – 0.948; Fig. 3b). On the contrary, under the post-hoc regression the 95% CI for the coefficient encompassed zero ($\beta_{\Omega_{SCM}} = 0.086$, 95% CI: –0.179 – 0.351; Fig. 3d).

Discussion

We extended the multi-species (community) occupancy model to allow for simultaneous estimation of species richness across multiple regions. Using our approach, spatial variation in species richness can be modeled directly using region-specific covariates, while explicitly accounting for species-level heterogeneity in probabilities of both occurrence and detection (Dorazio & Royle 2005). We demonstrated that the hierarchical structure of the model allows information to be shared across sites, species and regions resulting in model performance improvements in terms of ability to estimate covariate effects relative to the single season analogue. More importantly, the MRCM model provides a formal mechanism for testing hypotheses about drivers of geographic variation in community size and structure.

Dorazio *et al.* (2010) invoke the concept of a metacommunity (Leibold & Holyoak 2004) to motivate the development of multi-species occupancy models, in which the metacommunity is made up of a collection of sample locations considered to be ‘local’ communities. Our formulation of the model can also be considered a metacommunity model, although, rather than communities being defined by sampling locations, they are defined by specific regions of biological interest which are themselves repeatedly sampled in space and time (e.g. sampling plots or grids, management units or reserves). This permits the investigation of community size, structure and dynamics across a range of geographic scales such as landscape-scale patch networks (Tschamntke *et al.* 2007), national-scale reserve networks (Sierra *et al.* 2002), or global-scale biodiversity monitoring networks (Ahumada *et al.* 2011). Using an integrated multi-region approach, communities can be compared across any of these scales to investigate a wide range of important environmental influences on biodiversity. We demonstrated this using the multi-community bird data set, finding that larger communities were found in more heterogeneous habitats (Figure 2). Example of other potentially useful applications are the investigation of variation in species richness across cli-

mate gradients which can be used to infer potential impacts of projected environmental change (Araújo & Rahbek 2006), or the evaluation of conservation strategies across regions or reserves with different managements regimes or socio-ecological conditions (Pence *et al.* 2003; Murray *et al.* 1999). More generally though, this approach should be useful for investigating many aspects of spatial community ecology within a single framework that deals explicitly with issues of imperfect detection, heterogeneity in detectability and heterogeneity in occurrence probabilities (Cam *et al.* 2002).

Within the multi-region framework we extend the dimension of the augmented data to a third ($r = 1, \dots, R$) dimension and allow Ω to be estimated for each region (Ω_r), either as fixed or random effects (see also Tobler *et al.* 2015), but also as a function of covariates. A particular appealing feature of such an approach is the ability to retain species identity across regions which provides useful benefits such as the sharing of species level information across regions, and the ability to estimate and compare not only species richness but also community composition/similarity (Chao *et al.* 2004) and nestedness (Cam *et al.* 2000; Lomolino 1996).

In practice, relating species richness to explicit spatial or temporal covariates is often done by first obtaining region specific estimates of species richness, and then modelling this collection of estimates as if they were data. Such a two-step approach of ‘doing statistics on statistics’ has been repeatedly criticised (Brooks *et al.* 2015; Link 1999; Link & Barker 2004; Grosbois *et al.* 2008) and a hierarchical approach suggested (e.g. Cooch *et al.* 2012; Royle & Dorazio 2008). Moreover, characterizing uncertainty in estimated relationships between species richness and covariates in such a way can be misleading (Gould & Nichols 1998). Results from both simulated and real data suggest that a *de facto* hierarchical approach can be advantageous, particularly reducing the risk of finding spurious results and of overconfidence in the precision of results. Specifically, the hierarchical multi-region approach we propose provides an integrated modelling framework that partitions the total variability of species-richness into sampling and process variance.

We note that Tobler *et al.* (2015) recently proposed a ‘multi-session multi-species’ occupancy model in which sessions are treated as nested random effects, where surveys at multiple sites can be treated as sessions. This model is the equivalent of the ‘intercept-only’ model we use to motivate the more useful covariate models developed for testing hypotheses about structural variation in community size

and structure. The difference, which we argue is an important one, is that by treating sessions (regions) as random effects, shrinkage may be less informative about spatial variation when community size differs substantially, systematically, and/or predictably across group of sites as a function of measurable covariates (i.e., *not* randomly). In addition, variance for the random sessions can be difficult to estimate, or can be overestimated, when the number of sessions is small or when there are species detected in a few sessions (Gelman 2006; Tobler *et al.* 2015). Our multi-region framework is not subject to these limitations and can thus be more appropriate to investigate spatial variation in community size and structure, particularly when the number of regions is small (Fig. 2).

The multi-region model we present retains the regional level species-by-site data structure, but with a dimension expansion to allow for simultaneous modelling of multiple regions. As a result, all of the benefits and the recent model developments and extensions of the single community model (reviewed in Iknayan *et al.* 2014) apply to the multi-region case. We presented a static, or ‘closed’ community model, although it can easily be applied to ‘open’ communities allowing for colonization-extinction dynamics to occur within or between communities (e.g. Dorazio *et al.* 2010). Additionally, ecological stratification within communities, e.g., species traits such as body mass or functional guild, can be investigated using information shared across multiple communities which provides a mechanism for testing specific hypothesis about spatial variation in species traits among communities, as well community size and composition. To-date, single region community models derive region specific estimates of species richness, or in the dynamic case, region specific estimates of temporal trends. The multi-region approach, however, can be used to define meta-community-level relationships of species richness, community structure and species occurrence patterns that are spatially explicit. Because the multi-region framework allows for simultaneous modelling of multiple communities across space and time, such an approach should be extremely useful for understanding one of the most enduring challenges in ecology - what drives geographic variation in the size and structure of communities (MacArthur & Wilson 1967; Stevens 1989; Field *et al.* 2009).

Acknowledgments

We thank Aaron Iemma for IT assistance, and several anonymous referees for constructive comments on previous versions of this manuscript. Part of this research was performed using the ATLAS HPC Cluster which is supported by NSF grants (Award #1059284 and #0832782).

References

- Ahumada, J.A., Silva, C.E., Gajapersad, K., Hallam, C., Hurtado, J., Martin, E., McWilliam, A., Mugerwa, B., O'Brien, T., Rovero, F. *et al.* (2011) Community structure and diversity of tropical forest mammals: data from a global camera trap network. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2703–2711.
- Araújo, M. & Rahbek, C. (2006) How does climate change affect biodiversity? *Science*, **313**, 1396–7.
- Balvanera, P., Pfisterer, A.B., Buchmann, N., He, J.S., Nakashizuka, T., Raffaelli, D. & Schmid, B. (2006) Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology letters*, **9**, 1146–56.
- Boulinier, T., Nichols, J. & Sauer, J. (1998) Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, **79**, 1018–1028.
- Brooks, E.N., Deroba, J.J. & Wilberg, M. (2015) When ‘data’ are not data: the pitfalls of post hoc analyses that use stock assessment model output. *Canadian Journal of Fisheries and Aquatic Sciences*, **72**, 634–641.
- Cabeza, M. & Moilanen, A. (2001) Design of reserve networks and the persistence of biodiversity. *Trends in Ecology & Evolution*, **16**, 242–248.
- Cam, E., Nichols, J.D., Hines, J.E. & Sauer, J.R. (2000) Inferences about nested subsets structure when not all species are detected. *Oikos*, **91**, 428–434.

- 315 Cam, E., Nichols, J.D., Hines, J.E., Sauer, J.R., Alpizar-Jara, R. & Flather, C.H. (2002) Disentangling
316 sampling and ecological explanations underlying species-area relationships. *Ecology*, **83**, 1118–1130.
- 317 Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.J. (2004) A new statistical approach for assessing
318 similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.
- 319 Cooch, E.G., Conn, P.B., Ellner, S.P., Dobson, A.P. & Pollock, K.H. (2012) Disease dynamics in wild
320 populations: modeling and estimation: a review. *Journal of Ornithology*, **152**, 485–509.
- 321 Dorazio, R.M., Kéry, M., Royle, J.A. & Plattner, M. (2010) Models for inference in dynamic metacom-
322 munity systems. *Ecology*, **91**, 2466–2475.
- 323 Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by mod-
324 eling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.
- 325 Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accu-
326 mulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.
- 327 Field, R., Hawkins, B.A., Cornell, H.V., Currie, D.J., Diniz-Filho, J.A.F., Guégan, J.F., Kaufman, D.M.,
328 Kerr, J.T., Mittelbach, G.G., Oberdorff, T. *et al.* (2009) Spatial species-richness gradients across scales:
329 a meta-analysis. *Journal of Biogeography*, **36**, 132–147.
- 330 Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*,
331 **1**, 1–19.
- 332 Gotelli, N. & Colwell, R. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and
333 comparison of species richness. *Ecology letters*, **4**, 379–391.
- 334 Gould, W.R. & Nichols, J.D. (1998) Estimation of temporal variability of survival in animal populations.
335 *Ecology*, **79**, 2531–2538.
- 336 Grosbois, V., Gimenez, O., Gaillard, J.M., Pradel, R., Barbraud, C., Clobert, J., Møller, A. & Weimer-
337 skirch, H. (2008) Assessing the impact of climate variation on survival in vertebrate populations. *Bio-*
338 *logical Reviews*, **83**, 357–399.

339 Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging meth-
340 ods to estimate species diversity. *Trends in ecology & evolution*, pp. 1–10.

341 Johnson, M.P., Frost, N.J., Mosley, M.W., Roberts, M.F. & Hawkins, S.J. (2003) The area-independent
342 effects of habitat complexity on biodiversity vary between regions. *Ecology Letters*, **6**, 126–132.

343 Kerr, J. (1997) Species richness, endemism, and the choice of areas for conservation. *Conservation Biol-*
344 *ogy*, **11**, 1094–1100.

345 Kéry, M. & Royle, J. (2008) Hierarchical bayes estimation of species richness and occupancy in spatially
346 replicated surveys. *Journal of Applied Ecology*, **45**, 589–598.

347 Knaus, J. (2013) *snowfall: Easier cluster computing (based on snow)*. R package version 1.84-6.

348 Leibold, M. & Holyoak, M. (2004) The metacommunity concept: a framework for multi-scale community
349 ecology. *Ecology Letters*, **7**, 601–613.

350 Link, W.A. (1999) Modeling pattern in collections of parameters. *The Journal of wildlife management*,
351 **63**, 1017–1027.

352 Link, W.A. & Barker, R.J. (2004) Hierarchical mark–recapture models: a framework for inference about
353 demographic processes. *Animal Biodiversity and Conservation*, **27**, 441–449.

354 Lomolino, M. (1996) Investigating causality of nestedness of insular communities: selective immigrations
355 or extinctions? *Journal of biogeography*, **23**, 699–703.

356 MacArthur, R. & Wilson, E. (1967) *The theory of island biogeography*, volume 1 of *Monographs in*
357 *population biology*. Princeton University Press.

358 May, R. (1988) How many species are there on earth?. *Science*, **241**, 1441–1449.

359 Murray, S., Ambrose, R. & Bohnsack, J. (1999) No-take reserve networks: sustaining fishery populations
360 and marine ecosystems. *Fisheries*, **8446**, 37–41.

- 361 Myers, N., Mittermeier, R.a., Mittermeier, C.G., da Fonseca, G.a. & Kent, J. (2000) Biodiversity hotspots
362 for conservation priorities. *Nature*, **403**, 853–8.
- 363 Nichols, J.D., Boulmier, T., Hines, J.E., Pollock, K.H. & Sauer, R. (1998) Inference methods for spatial
364 variation in species richness and community composition when not all species are detected. *Conserva-*
365 *tion Biology*, **12**, 1390–1398.
- 366 Padoa-Schioppa, E., Baietto, M., Massa, R. & Bottoni, L. (2006) Bird communities as bioindicators: the
367 focal species concept in agricultural landscapes. *Ecological Indicators*, **6**, 83–93.
- 368 Pence, G.Q., Botha, M.a. & Turpie, J.K. (2003) Evaluating combinations of on-and off-reserve conserva-
369 tion strategies for the Agulhas Plain, South Africa: a financial perspective. *Biological Conservation*,
370 **112**, 253–273.
- 371 Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.
372 *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.
- 373 Plummer, M. (2013) *rjags: Bayesian graphical models using MCMC*. R package version 3-10.
- 374 Purvis, A. & Hector, A. (2000) Getting the measure of biodiversity. *Nature*, **405**, 212–219.
- 375 R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statis-
376 tical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- 377 Robert, C. & Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer, New York.
- 378 Royle, J.A. & Dorazio, R. (2008) *Hierarchical modeling and inference in ecology: the analysis of data*
379 *from populations, metapopulations and communities*. Academic Press, San Diego.
- 380 Royle, J.A., Dorazio, R.M. & Link, W.A. (2007) Analysis of multinomial models with unknown index
381 using data augmentation. *Journal of Computational and Graphical Statistics*, **16**, 67–85.
- 382 Shannon, C.E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**,
383 379–423.

- 384 Sierra, R., Campos, F. & Chamberlin, J. (2002) Assessing biodiversity conservation priorities: ecosystem
385 risk and representativeness in continental Ecuador. *Landscape and Urban Planning*, **59**, 95–110.
- 386 Stevens, G. (1989) The latitudinal gradient in geographical range: how so many species coexist in the
387 tropics. *American naturalist*, **133**, 240–256.
- 388 Tobler, M.W., Zúñiga Hartley, A., Carrillo-Percestequi, S.E. & Powell, G.V. (2015) Spatiotemporal hi-
389 erarchical modelling of species richness and occupancy using camera trap data. *Journal of Applied*
390 *Ecology*.
- 391 Tscharntke, T., Bommarco, R., Clough, Y., Crist, Thomas, O., Kleijn, D., Rand, T.A., Tylianakis, J.M.,
392 van Nouhuys, S. & Vidal, S. (2007) Conservation biological control and enemy diversity on a landscape
393 scale. *Biological control*, **43**, 294–309.

Table 1: Summaries statistics from the simulation study described in the text. In total, 144 simulations were carried out for each of the 12 scenarios in which we varied the number of regions (*Region*), number of sites per region (*Sites*), the mean community-level detectability (\bar{p}), the heterogeneity in community-level detectability (σ_p^2), and the effect of a covariate on species richness (β). We were particularly interested in the ability to characterise the species richness-covariate relationship and report the mean ($\hat{\theta}$) and standard deviation ($\text{sd}(\theta)$) of the per simulation estimate of both α_Ω and β_Ω . We also provide the proportion of the 95% CIs from all simulations that spanned contained the true, data generating value. Results are presented for both the proposed multi-region community model (MRCM), and the two step regression approach based on independent single region community models (SRCM).

<i>Settings</i>					<i>MRCM</i>						<i>SRCM</i>					
Regions	Sites	\bar{p}	σ_p	β	$\hat{\alpha}_\Omega$	$\text{sd}(\hat{\alpha}_\Omega)$	$\text{cov}(\hat{\alpha}_\Omega)$	$\hat{\beta}_\Omega$	$\text{se}(\hat{\beta}_\Omega)$	$\text{cov}(\hat{\beta}_\Omega)$	$\hat{\alpha}_\Omega$	$\text{sd}(\hat{\alpha}_\Omega)$	$\text{cov}(\hat{\alpha}_\Omega)$	$\hat{\beta}_\Omega$	$\text{se}(\hat{\beta}_\Omega)$	$\text{cov}(\hat{\beta}_\Omega)$
6	25	0.3	0.5	0.00	-0.81	0.07	0.94	0.00	0.11	0.97	-0.81	0.07	0.86	0.00	0.11	0.88
				0.40	-0.81	0.06	0.97	0.40	0.11	0.99	-0.80	0.07	0.92	0.40	0.11	0.92
				0.80	-0.81	0.07	0.96	0.80	0.12	0.96	-0.80	0.07	0.88	0.81	0.13	0.87
			1.5	0.00	-0.84	0.07	0.92	0.00	0.12	0.94	-0.83	0.08	0.85	0.00	0.13	0.87
				0.40	-0.84	0.07	0.94	0.39	0.12	0.99	-0.82	0.07	0.91	0.39	0.13	0.92
				0.80	-0.84	0.08	0.92	0.79	0.13	0.95	-0.83	0.08	0.88	0.79	0.15	0.85
		0.6	0.5	0.00	-0.80	0.07	0.96	0.00	0.11	0.97	-0.80	0.07	0.87	0.00	0.10	0.85
				0.40	-0.80	0.06	0.99	0.40	0.11	1.00	-0.80	0.06	0.93	0.41	0.11	0.93
				0.80	-0.80	0.07	0.96	0.80	0.12	0.96	-0.80	0.07	0.90	0.80	0.12	0.88
			1.5	0.00	-0.81	0.07	0.95	0.00	0.11	0.97	-0.81	0.07	0.86	0.00	0.11	0.90
				0.40	-0.81	0.06	0.95	0.40	0.11	0.99	-0.80	0.07	0.90	0.41	0.12	0.92
				0.80	-0.81	0.07	0.96	0.80	0.12	0.96	-0.80	0.07	0.88	0.80	0.13	0.85
	50	0.3	0.5	0.00	-0.81	0.07	0.93	0.02	0.12	0.94	-0.81	0.07	0.86	0.02	0.12	0.86
				0.40	-0.81	0.06	0.98	0.40	0.10	0.99	-0.80	0.06	0.92	0.40	0.10	0.90
				0.80	-0.80	0.07	0.95	0.79	0.12	0.97	-0.80	0.07	0.87	0.80	0.12	0.89
			1.5	0.00	-0.83	0.07	0.92	0.02	0.12	0.94	-0.82	0.07	0.85	0.02	0.12	0.87
				0.40	-0.83	0.06	0.99	0.40	0.10	0.99	-0.82	0.06	0.89	0.40	0.10	0.94
				0.80	-0.82	0.07	0.95	0.79	0.13	0.92	-0.82	0.07	0.87	0.79	0.13	0.84
		0.6	0.5	0.00	-0.80	0.07	0.95	0.02	0.12	0.93	-0.81	0.07	0.88	0.02	0.12	0.86
				0.40	-0.80	0.06	0.99	0.40	0.10	1.00	-0.80	0.06	0.92	0.40	0.10	0.90
				0.80	-0.80	0.06	0.94	0.80	0.12	0.94	-0.80	0.07	0.89	0.80	0.12	0.86
			1.5	0.00	-0.81	0.07	0.94	0.02	0.12	0.94	-0.81	0.07	0.88	0.02	0.12	0.88
				0.40	-0.81	0.06	0.98	0.40	0.10	0.99	-0.80	0.06	0.92	0.40	0.10	0.92
				0.80	-0.80	0.07	0.94	0.79	0.12	0.96	-0.80	0.07	0.86	0.80	0.12	0.85
15	25	0.3	0.5	0.00	-0.81	0.04	0.92	0.00	0.06	0.97	-0.81	0.04	0.93	0.00	0.06	0.98
				0.40	-0.82	0.04	0.92	0.40	0.07	0.96	-0.81	0.04	0.92	0.41	0.07	0.94
				0.80	-0.81	0.04	0.92	0.80	0.07	0.94	-0.81	0.04	0.92	0.80	0.07	0.91
			1.5	0.00	-0.84	0.04	0.84	0.00	0.07	0.97	-0.83	0.04	0.88	0.00	0.08	0.93
				0.40	-0.84	0.04	0.83	0.40	0.07	0.96	-0.83	0.05	0.87	0.40	0.08	0.92
				0.80	-0.84	0.04	0.85	0.79	0.07	0.96	-0.83	0.04	0.87	0.79	0.08	0.92
		0.6	0.5	0.00	-0.80	0.04	0.94	0.00	0.06	0.97	-0.80	0.04	0.94	0.00	0.06	0.96
				0.40	-0.81	0.04	0.95	0.41	0.07	0.93	-0.81	0.04	0.92	0.41	0.07	0.90
				0.80	-0.80	0.04	0.92	0.80	0.07	0.95	-0.80	0.04	0.90	0.81	0.07	0.90
			1.5	0.00	-0.81	0.04	0.94	0.00	0.06	0.97	-0.81	0.04	0.94	0.00	0.07	0.97
				0.40	-0.81	0.04	0.92	0.40	0.07	0.95	-0.81	0.04	0.91	0.41	0.07	0.91
				0.80	-0.81	0.04	0.94	0.80	0.07	0.94	-0.81	0.04	0.91	0.80	0.07	0.92
	50	0.3	0.5	0.00	-0.80	0.04	0.94	0.01	0.06	0.95	-0.80	0.04	0.92	0.01	0.06	0.93
				0.40	-0.81	0.04	0.94	0.41	0.07	0.95	-0.81	0.04	0.92	0.41	0.07	0.89
				0.80	-0.81	0.04	0.94	0.80	0.07	0.94	-0.81	0.04	0.93	0.80	0.07	0.92
			1.5	0.00	-0.82	0.04	0.90	0.01	0.06	0.97	-0.82	0.04	0.90	0.01	0.07	0.96
				0.40	-0.83	0.04	0.90	0.40	0.07	0.94	-0.82	0.04	0.85	0.40	0.07	0.93
				0.80	-0.83	0.04	0.91	0.79	0.07	0.95	-0.82	0.04	0.90	0.80	0.08	0.91
		0.6	0.5	0.00	-0.80	0.04	0.94	0.01	0.06	0.97	-0.80	0.04	0.92	0.01	0.06	0.93
				0.40	-0.80	0.04	0.94	0.41	0.07	0.94	-0.80	0.04	0.94	0.41	0.07	0.88
				0.80	-0.80	0.04	0.94	0.80	0.07	0.94	-0.81	0.04	0.94	0.80	0.07	0.90
			1.5	0.00	-0.80	0.04	0.95	0.01	0.06	0.97	-0.81	0.04	0.93	0.01	0.06	0.93
				0.40	-0.81	0.04	0.94	0.41	0.06	0.94	-0.81	0.04	0.92	0.41	0.07	0.90
				0.80	-0.81	0.04	0.94	0.80	0.07	0.95	-0.81	0.04	0.92	0.80	0.07	0.91

Table 2: Summaries of posterior distributions from the multi-region community models (MRCM) applied to the avian community data from eight regions. Zero-inflation parameters α_Ω and β_Ω , and standard deviations for occurrence and detection probability (σ_ψ and σ_p , respectively) are on logit scale. β_Ω represents the slope for the relationship between species richness (via Ω) and habitat complexity. Average occurrence ($\bar{\psi}$) and detection (\bar{p}) probabilities are given on probability scale, i.e., $\bar{\psi} = \text{expit}(\mu_\psi)$ and $\bar{p} = \text{expit}(\mu_p)$, where expit is the inverse-logit function.

Parameter	Mean	SD	<i>Quantiles</i>		
			0.025	0.500	0.975
α_Ω	-1.112	0.411	-1.904	-1.118	-0.288
β_Ω	0.475	0.238	0.017	0.473	0.948
$\bar{\psi}$	0.040	0.011	0.021	0.039	0.062
σ_ψ	2.193	0.163	1.900	2.185	2.542
\bar{p}	0.487	0.009	0.469	0.487	0.505
σ_p	0.439	0.031	0.381	0.438	0.502

Figures

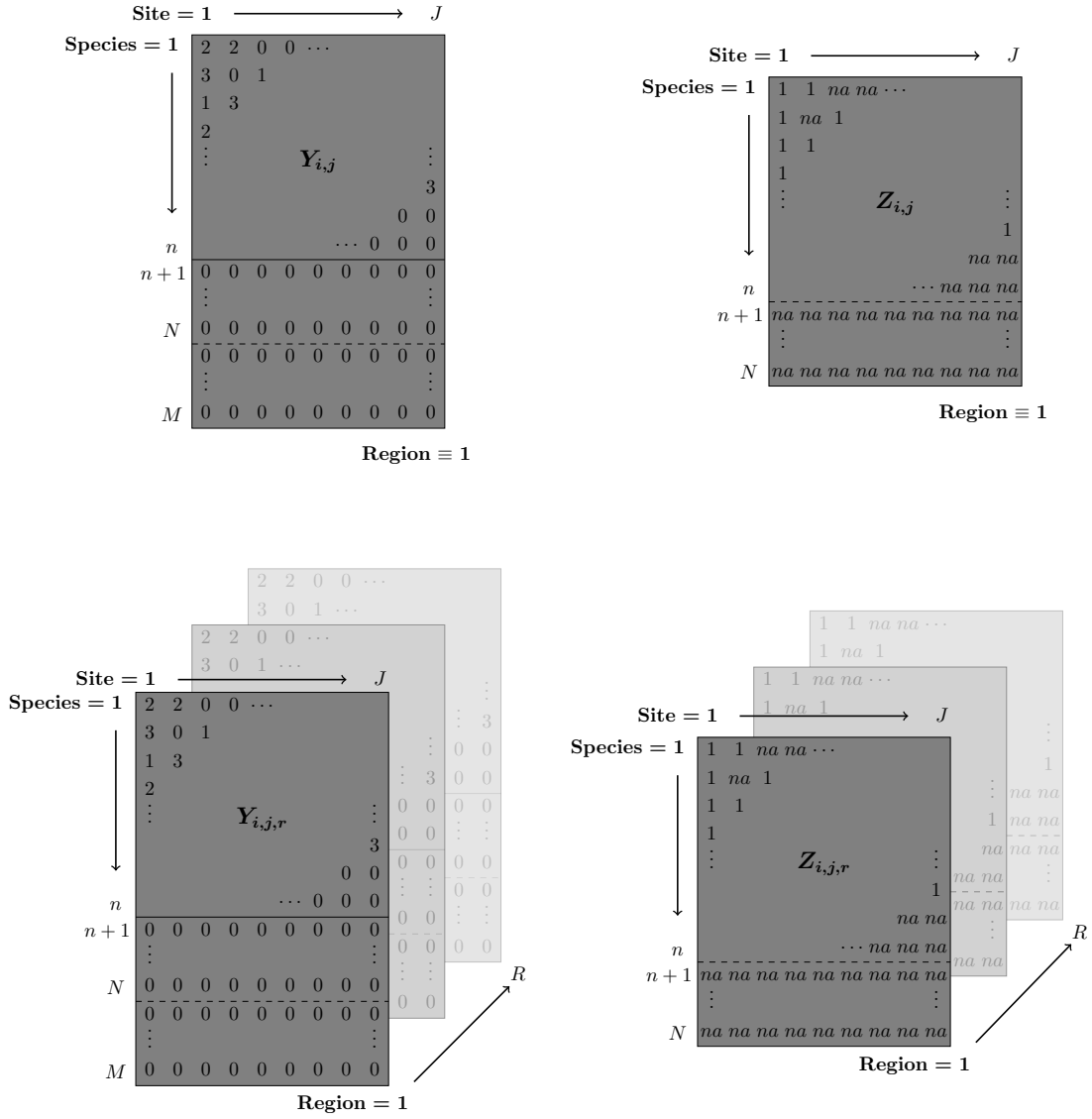


Figure 1: Outline of the relationship between the observed data (Y matrix, with the number of detections in K visits for each of the n species) and the partially observed true state (Z matrix) in a classical multi-species occupancy model (above) extended to a multi-region framework (below). The observed data matrix of each region is augmented with $M - n$ all-zero detections. Missing values are denoted by 'na'.

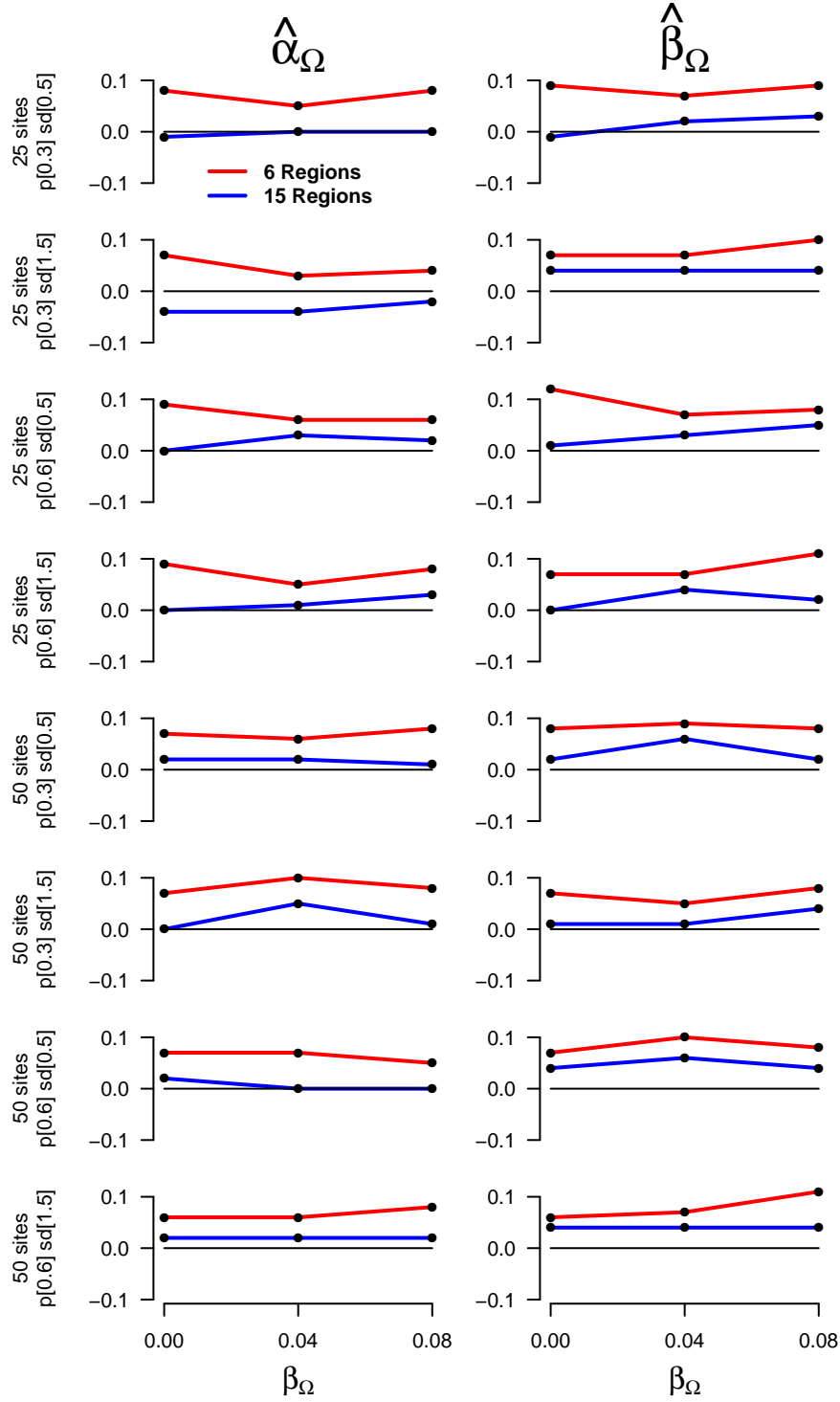


Figure 2: Visual comparison of the relative performance of the multi region (MRCM) and single region (SRCM) community models based on the simulation study described in the text. The values are the difference in coverage between the MRCM and the SRCM (i.e., $\text{cov}(\theta_{MRCM}) - \text{cov}(\theta_{SRCM})$), and therefore more positive values denote better coverage using the MRCM approach, more negative values denote better coverage using the SRCM approach, and 0 (shown by the black horizontal line) denote no difference. For example, a value of 0.1 means that coverage using the MRCM was 10% better than when using SRCM. Red lines are scenarios using only 6 regions, blue lines are scenarios using 15. X and Y axis labels describe the simulations settings for detectability and the number of sites (Y axes), and the simulated β coefficient for the species richness relationship (X axes).

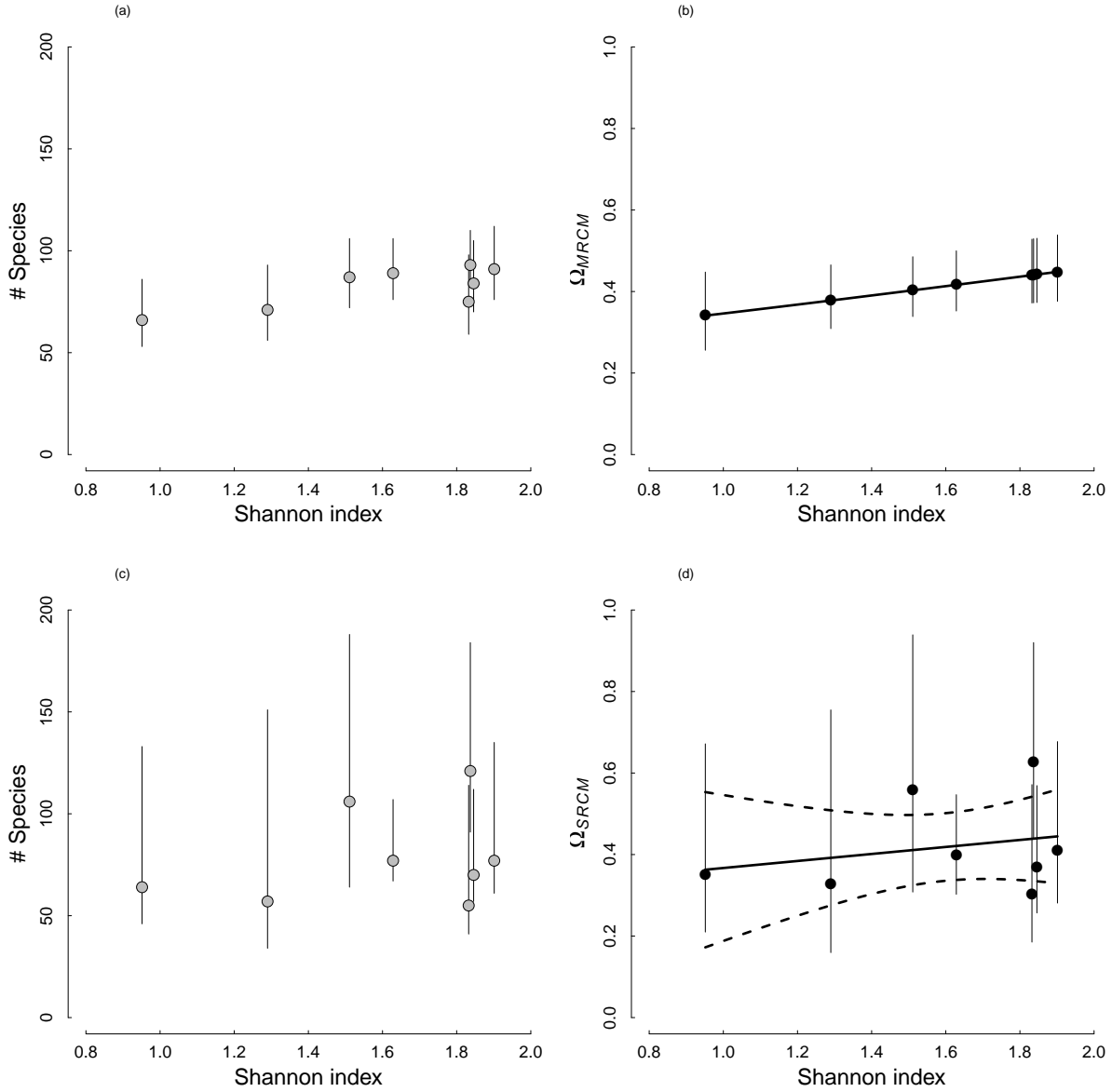


Figure 3: Relationship between species richness (N) and the Shannon entropy index (SEI) as a proxy of habitat complexity, derived from the multi region model (MRCM) (a) and the single region models (SRCM) separately fitted (c). Median values (grey-filled circles) and 95% credible intervals (vertical lines) reported. In (b) and (d) the analogue relationship with the zero-inflation parameter is reported for the MRCM and SRCMs, respectively. Black circles indicate Ω mean values, with 95% credible intervals in (b) and 95% confidence intervals in (d). The solid regression line in (d) indicates the predicted relationship between Ω_{SRCM} and the SEI from the post-hoc approach (dashed lines indicate 95% confidence interval).