Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 783–791 doi: 10.1111/2041-210X.12536

A multiregion community model for inference about geographic variation in species richness

Chris Sutherland[1,†], Mattia Brambilla[2,3], Paolo Pedrini[2] and Simone Tenan[2*,†]

[1]*Department of Environmental Conservation, University of Massachusetts, Amherst, MA 01003, USA;* [2]*Vertebrate Zoology Section, MUSE - Museo delle Scienze, Corso del Lavoro e della Scienza 3, I-38122 Trento, Italy; and* [3]*Fondazione Lombardia per l'Ambiente, Settore biodiversità e aree protette, Largo 10 luglio 1976 1, I-20822 Seveso (MB), Italy*

Summary

1. 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.

2. 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).

3. We report improved model performance of the joint multiregion approach when compared to the more traditional two-stage approach of modelling spatial variation in species richness using simulations. Then, applying the model to data collected from eight avian communities in northern Italy, we demonstrate how species richness can be modeled as a spatially varying function of habitat complexity.

4. Extending the multi-species occupancy model to accommodate data collected across multiple regions of interest (e.g., reserves or biomes) allows for joint estimation of region-specific community size and permits species richness to be modeled as a function of region-specific covariates. Our approach provides a mechanism for testing hypotheses about why and how species richness varies across space.

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 understanding what determines whether areas 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 an area is that, due to imperfect detection probabilities, some species may go completely undetected (Boulinier, Nichols & Sauer 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, Nichols & Sauer 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 for multiple areas, then point estimates 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, Deroba & Wilberg 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; Zipkin *et al.* 2010), 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). As a result, hierarchical models that integrate both components of the 'two-stage' approach, and in doing so jointly estimate all of

*Correspondence author. E-mail: csutherland@umass.edu
[†]These authors equally contributed to this work.

the parameters accounting for the multiple sources of variation, are often promoted as preferred alternatives (Royle & Dorazio 2008).

The recent development of multispecies occupancy (or community) models (Dorazio & Royle 2005; Dorazio *et al.* 2006; Royle, Dorazio & Link 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, Dorazio & Link 2007; see Methods), providing a direct estimate of total species richness of a community (Kéry & Royle 2008). The multispecies 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 jointly model species richness across multiple communities precludes the formal testing of hypotheses about drivers of spatial variation in species richness across multiple regions. Moreover, the concept of a 'metacommunity' of Dorazio *et al.* (2006), that is, all species are potentially shared across all sites, is less appropriate when the true species composition is not the same across distinct regions, for example, when certain species do not occur in one or more of the regions like carnivores in South America vs. South-East Asia. This issue could partly be resolved using the approach of Royle & Converse (2014) who propose a 'stratified population' model for estimating which of several populations a particular individual belongs. Their approach assumes that each unique individual belongs to exactly one population, which is restrictive in the context of species and communities because often distinct communities can share one or more species. Instead, it would be preferable to jointly model communities as geographically distinct species assemblages in a more general way that allows unique regional species composition that is independent of other regions.

In this paper, we develop a 'multiregion community model' (MRCM), a general and flexible framework for joint estimation of species richness across geographically distinct communities that yields improved statistical performance and which, by preserving species identity across regions, can be applied regardless of whether communities share species or not, thus resolving the issues outlined above. In essence, the model links multiple community models via common parameters and requires no more than the familiar multispecies data structure but collected at more than one region, that is, multispecies detection histories for many sampling sites within multiple regions of interest, and naturally facilitates explicit models for regional differences in species richness. We first describe a general formulation for estimating region-specific species richness as a random effect (see also Tobler *et al.* 2015), before describing the important extension of the model that allows variation in species richness to be modelled as a function of region-level covariates, and thus a mechanism for investigating drivers of geographic variation in the size and structure of communities. We demonstrate this novel model development and compare its performance with that of a two-stage approach using both simulated data and then point count data collected from eight geographically distinct avian communities.

Materials and methods

A MULTIREGION MODEL FOR SPECIES RICHNESS

As a starting point, we consider data collected for analysis of a Bayesian multispecies occupancy model (Dorazio & Royle 2005). Multispecies occupancy models are an extension of the single-species site occupancy model (MacKenzie *et al.* 2002) that can be thought of as multiple single-species occupancy models that are linked statistically through the sharing of parameters across species (Dorazio & Royle 2005). The benefit is that by explicitly accounting for imperfect detection at the species level, the number of unobserved species can be estimated. The hierarchical structure of the multispecies model combines community- and species-level attributes within a single analytical framework. The sampling protocol requires repeated observations at different sampling locations that leads to encounter frequencies (y_{ij}) of species $i = 1, \ldots, n$ at each of $j = 1, \ldots, J$ sites, which are visited $k = 1, \ldots, K$ times.

The data can be formatted as a 2-dimensional $n \times J$ matrix $\boldsymbol{Y}$ and are (potentially imperfect) observations of the corresponding 2-dimensional $N \times J$ matrix $\boldsymbol{Z}$, the true but unknown occupancy states of each species in each site, z_{ij} (Fig. 1). Here, n is the total number of species *observed*, N is the (latent) total community size, and thus, $N - n$ is the number of species in the community *never* detected, which, by definition, have 'all-zero' encounter histories. The inference objective is to estimate the 'true' occupancy states, $\boldsymbol{Z}$, by accounting for imperfect detection which is estimated using detection data from repeated sites visits, that is from $\boldsymbol{Y}$. Sites are assumed to be a representative sample of a larger, single geographic area for which species-specific occupancy and overall species richness are of interest, which is to say there is $R = 1$ region.

We extend this single-region community model (SRCM) to accommodate data collected at multiple spatially independent regions, that is $r = 1, \ldots, R$ where $R > 1$. This provides some distinct practical benefits; specifically, the integration of data collected across geographic gradients and scales into a single analytical framework, for example reserve networks (Sierra, Campos & Chamberlin 2002), biogeographic regions (Dorazio *et al.* 2010) or, more generally, sites where 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.

Multispecies encounter frequency data from multiple regions are summarized in a 3-dimensional array $\boldsymbol{Y}$ with elements y_{ijr}, where the subscript r now indexes region (Fig. 1). The matrix of true occupancy states $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ is formulated such that the site- and region-specific occupancy states of each species are Bernoulli random variables ($z_{ijr} = 1$ is occupied, and $z_{ijr} = 0$ is empty):

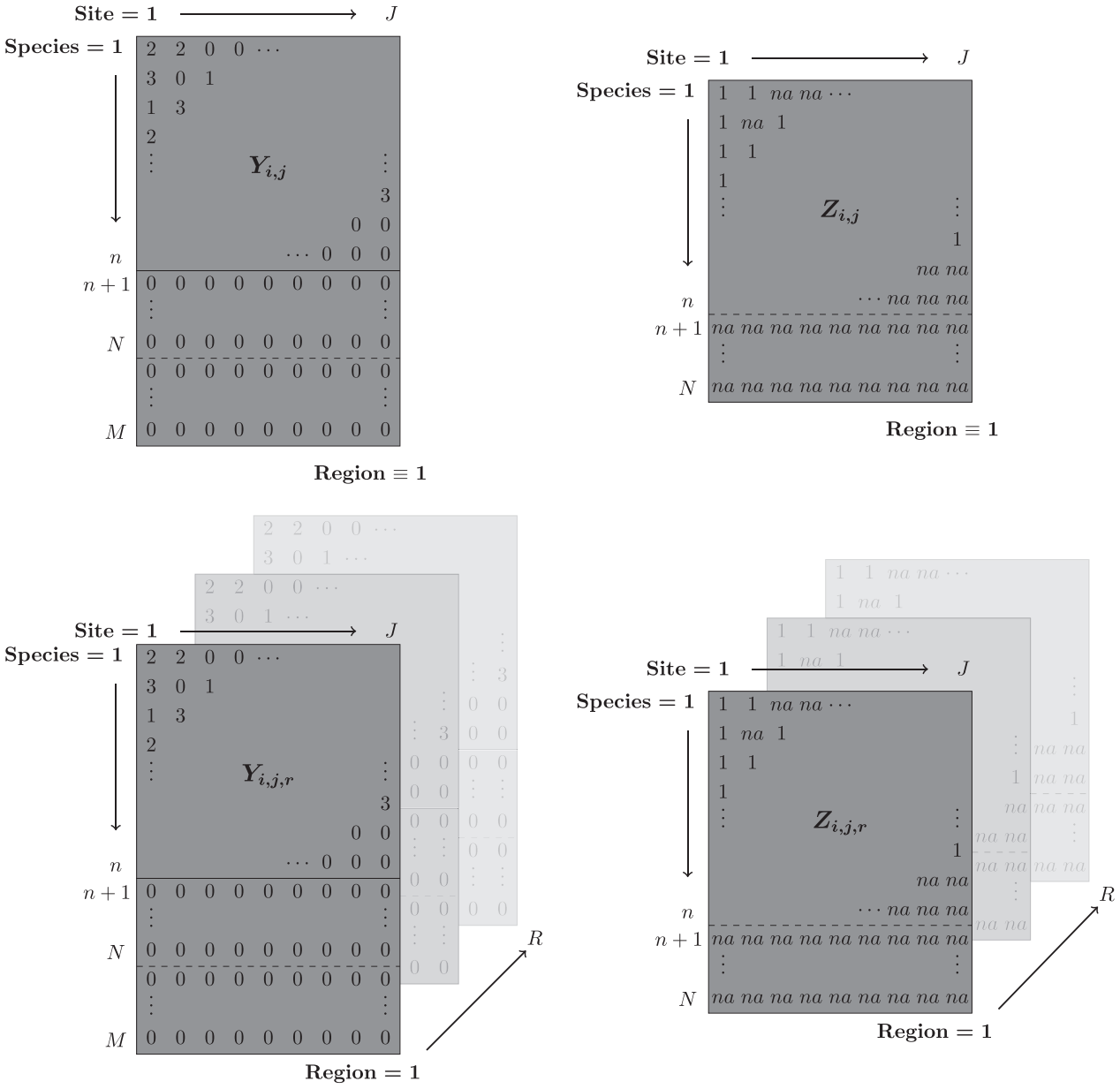

Fig. 1. Outline of the relationship between the observed data (***Y*** matrix, with the number of detections in K visits for each of the n species) and the partially observed true state (***Z*** matrix) in a classical multispecies occupancy model (above) extended to a multiregion framework (below). The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

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

where ω_{ir} is a species-specific indicator variable denoting whether the species is present in the region. The observation model relates the truth, that is ***Z***, to the data such that

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

Parameter, ψ_{ijr} is the species-specific occurrence probability for each site in each region conditional on species i being a member of the rth community, and p_{ijr} is the corresponding detection probability which is conditional on the species being present at site j in region r. 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 *et al.* 2006; Royle, Dorazio & Link 2007). The approach assumes a 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 but must be kept equal across regions, that is, $M_r \equiv M$, but must also be larger than the total number of species in the largest community, that is, $M \gg \max(N_1, \ldots, N_R)$ (Fig. 1). Alternatively, if the species pool is known, as is the case for well-studied taxa and areas, the model can be conditioned on the known ensemble of species (e.g. Tobler *et al.* 2015). In this case, species richness may not be the inference objective although the MRCM still provides the benefit of being able to jointly estimate parameters shared across species, sites and regions.

The data are then augmented with $M - n$ 'all-zero' encounter histories (Fig. 1), and the aim is to estimate the proportion, Ω, of these

representing species that exist in the community but that were never detected (i.e. that were sampling zeros and not structural zeros):

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

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

Extending the model to accommodate multiple regions allows simultaneous estimation of region-specific 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 thus converts the problem of estimating region-specific species richness, N_r, to that of estimating the region-specific complement of the zero-inflation parameter, Ω_r. In order for estimates of Ω to be comparable among regions, M must be the same for each region, which, when M is large, is approximately equivalent to modelling region-specific Poisson intensity parameters (Royle & Dorazio 2008).

MODELLING REGIONAL VARIATION IN SPECIES RICHNESS

Given the development of a formal derivation of the expected number of species in a region, N_r, the multiregion framework lends itself to the straightforward extension of explicitly modelling the effect of region-specific covariates on species richness. Recognizing that the model described above is an 'intercept-only' model, it can easily be extended to a logit-linear model including covariate(s) X_r, e.g. logit $(\Omega_r) = \alpha_\Omega + \beta_\Omega X_r$. We use a single covariate here without loss of generality. A canonical example of a covariate affecting the total number of species would be the size of the region, for example species-area relationship (MacArthur & Wilson 1967) or habitat complexity/heterogeneity (Johnson *et al.* 2003).

DEMONSTRATION BY SIMULATION

To demonstrate, and to assess the performance of the multiregion community model, we simulated multispecies 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** ($\bar{p} = 0 \cdot 3$) and **high** ($\bar{p} = 0 \cdot 6$) mean species detectability ($\bar{p}$), but also **low** ($\sigma_p^2 = 0 \cdot 5$) and **high** ($\sigma_p^2 = 1 \cdot 5$) community-level detection heterogeneity (σ_p^2). Species-specific detectability values were drawn from a Normal ($\bar{p}$, σ_p) distribution. 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 ($\alpha_\Omega = -0{\cdot}8$), we simulated data using **no** ($\beta_\Omega = 0$), **moderate** ($\beta_\Omega = 0{\cdot}4$) and **strong** ($\beta_\Omega = 0{\cdot}8$) covariate effects (remembering that 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 ($\bar{\psi} = 0 \cdot 3$ with, on the untransformed scale, $\sigma_\psi^2 = 1 \cdot 0$). Species-specific occupancy probabilities were drawn from a Normal ($\bar{\psi}$, σ_ψ) distribution.

For each of the 48 scenarios, we simulated 144 multiregion community data sets and analysed each using the proposed multiregion community model using Markov chain Monte Carlo, or MCMC (Robert & Casella 2004). We modelled probability parameters on the normal scale such that logit $(\theta) = \mu_0$ and 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 (burn-in = 5000 iterations) posterior distribution of model parameters. We retained the posterior mean value of each parameter from each simulation and recorded whether the 95% Bayesian credible intervals of each parameter overlapped the true value (coverage from here). In this situation, we were particularly interested in coverage to evaluate the ability of the model to accurately estimates the parameters driving variation in species richness. In addition, we report the mean and standard deviation of the posterior mean parameter estimates from all 144 simulations for each of the eight scenarios.

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, for comparison, we contrasted our results with results obtained using a two-stage procedure. To do so, we analysed the data from each region separately using single-region multispecies occupancy models (SRCM) and then regressed species richness point estimates (medians) against the community-specific covariate using a generalized linear model (stage two). Note that in order to compare across models, we used the estimated data augmentation parameters Ω_r in the regression models. We recorded the maximum likelihood estimate of each parameter from the post hoc regression, as well as recording whether the 95% confidence interval spanned the true parameter values. The point estimates used in the post hoc regression analysis were obtained by MCMC from 10 000 post-burn-in (burn-in = 5000 iterations) posterior samples.

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

Simulation Results

Overall, there was little difference in the bias (simulation mean) or precision (simulation standard deviation) between the two approaches for both α_Ω and β_Ω (Appendix S1, Table S1). Parameter coverage, however, differed markedly between the two methods. To compare the coverage performance between the two approaches (cov_{MRCM} and cov_{SRCM}, respectively), we calculated the relative improvement (κ) of MRCM as $\kappa = (\text{cov}_{\text{MRCM}} - \text{cov}_{\text{SRCM}}) / \text{cov}_{\text{SRCM}}$. The major difference was that the MRCM approach provided substantial improvements in coverage (up to 10%) across practically all scenarios when compared to the two-stage regression approach (Table 1, Fig. 2).

The MRCM achieved the nominal 95% coverage for both parameters in the majority of the simulation scenarios ($\hat{\alpha}_\Omega$: 27/48, $\hat{\beta}_\Omega$: 41/48, Appendix S1, Table S1), whereas the SRCM rarely achieved that level ($\hat{\alpha}_\Omega$: 0/48, $\hat{\beta}_\Omega$: 4/48, Appendix S1, Table S1). As shown by the systematic positive relative improvement values, κ, in Fig. 2, the MRCM achieved improved coverage in all but one scenario and the improvements were more pronounced in the lower data quality scenarios (six regions, red lines in Fig. 2). To evaluate more generally how well both approaches were able to accurately estimate the

Table 1. Coverage statistics for estimates of α_Ω and β_Ω for the multi- and single-region community model (MRCM and SRCM, respectively). Coverage statistics are the proportion of the simulated 95% BCIs that contained the true, data generating value, and we also report the relative improvement of MRCM, κ, which is computed as $\kappa = (\text{cov}_{\text{MRCM}} - \text{cov}_{\text{SRCM}}) / \text{cov}_{\text{SRCM}}$. To evaluate, in general, how well both approaches were able to accurately estimate the covariate effect, coverage statistics were calculated for simulations pooled across the three values of β_Ω for scenarios with varying number of regions (*Regions*), sites per region (*Sites*), mean community-level detectability ($\bar{p}$), and heterogeneity in community-level detectability (σ_p^2)

				Coverage statistic					
				α_Ω			β_Ω		
Regions	Sites	$\bar{p}$	σ_p^2	MRCM	SRCM	κ	MRCM	SRCM	κ
6	25	0·30	0·50	0·956	0·887	0·078	0·972	0·889	0·093
			1·50	0·926	0·880	0·052	0·961	0·882	0·090
		0·60	0·50	0·968	0·900	0·076	0·975	0·887	0·099
			1·50	0·954	0·880	0·084	0·972	0·889	0·093
	50	0·30	0·50	0·954	0·882	0·082	0·965	0·884	0·092
			1·50	0·951	0·868	0·096	0·949	0·882	0·076
		0·60	0·50	0·961	0·894	0·075	0·958	0·875	0·095
			1·50	0·954	0·884	0·079	0·963	0·882	0·092
15	25	0·30	0·50	0·924	0·921	0·003	0·956	0·942	0·015
			1·50	0·840	0·873	−0·038	0·963	0·921	0·046
		0·60	0·50	0·938	0·917	0·023	0·949	0·917	0·035
			1·50	0·935	0·921	0·015	0·954	0·931	0·025
	50	0·30	0·50	0·942	0·926	0·017	0·947	0·914	0·036
			1·50	0·903	0·884	0·021	0·954	0·933	0·023
		0·60	0·50	0·942	0·933	0·010	0·954	0·905	0·054
			1·50	0·947	0·924	0·025	0·954	0·914	0·044

covariate effect, we computed coverage statistics and relative improvement values pooled across all values of β_Ω. The MRCM outperformed the SRCM in terms of coverage in every case and yielded, on average, a 9% improvement in the low data quality scenarios, that is, with only six regions, and a 3·5% improvement on average in the higher data quality scenarios, that is, with 15 regions (Table 1).

In summary, the two approaches are comparable in terms of bias and precision, however, the ability to accurately and consistently estimate covariate effects on spatial variation in species richness is greatly improved using the proposed MRCM approach, especially in situations where less data are available (Table 1, Fig. 2, Appendix S1, Table S1).

APPLICATION: AVIAN COMMUNITIES AND HABITAT HETEROGENEITY

Having evaluated the performance of the model via simulation, we provide an application of the model to detection non-detection data collected from bird assemblages in $R = 8$ geographically distinct areas (regions) in northern Italy. Habitat composition in each region varied markedly among regions and the initial motivation for the sampling was to evaluate the effect of such heterogeneity on species richness (Padoa-Schioppa *et al.* 2006). Habitat complexity ranged from only grassland and woodland, to habitats made up of woodland, grassland, wetland, agricultural and urbanized areas. We characterized habitat heterogeneity using the Shannon entropy index (SEI, Shannon 1948), and modelled the relationship between species richness and SEI: $\text{logit}(\Omega_r) = \alpha_\Omega + \beta_\Omega \text{SEI}_r$. We expected a positive relationship between habitat complexity and species richness. SEI for each region was calculated for the area defined by the minimum convex polygon encompassing all surveyed points plus a 100-m buffer around it.

In each region, incidence records of all species observed within a 100-m radius of the observed during multiple 10-min count periods were recorded. The number of point count locations in each region ranged from a minimum of $J = 11$ to a maximum of $J = 103$ (median: 31·5), and the number of sampling occasions (10-min point counts) ranged from $K = 2$ to $K = 10$ (median: 3).

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

EXAMPLE APPLICATION

The number of *observed* species across the eight regions varied between 29 and 75 species (median 49·5). The estimated species richness from the multiregion model ranged from 66 (95% Bayesian CI: 53–86) to 93 (95% BCI: 82–110; Table 2). In comparison, estimated species richness from the two-stage approach of analysing estimates from the independent single community models (SRCM) ranged from 55 (95% BCI: 41–

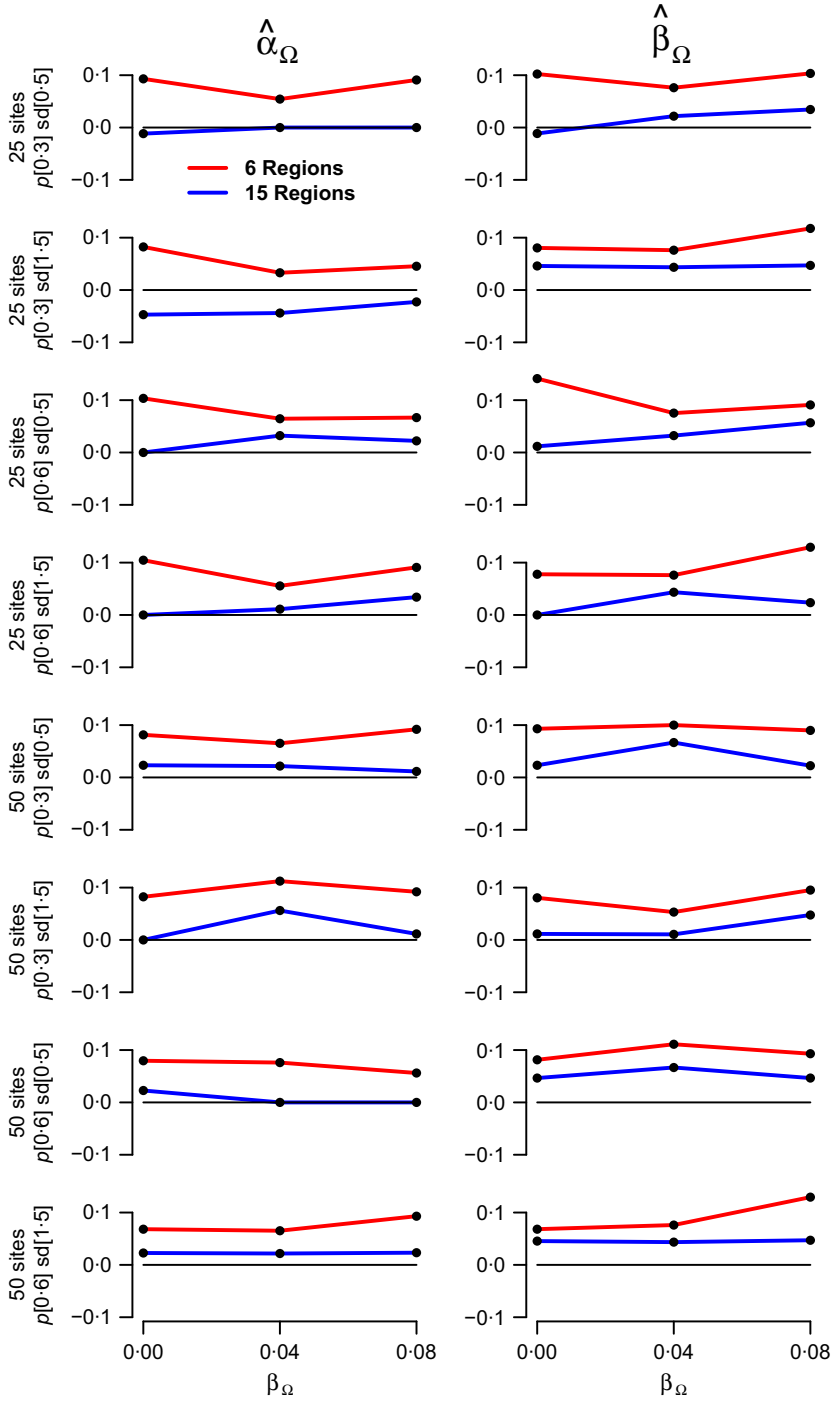

Fig. 2. Visual comparison of the relative performance of the parameter coverage of the multiregion model (MRCM) compared to the single-region model (SRCM). The values are the relative improvement of MRCM (i.e. $\kappa = (cov_{MRCM} - cov_{SRCM}) / cov_{SRCM}$), and therefore, positive values denote improved coverage using the MRCM approach, negative values denote better coverage using the SRCM approach, and 0 (shown by the black horizontal line) denotes no difference. For example, a value of 0·1 means that coverage was improved by 10% using the MRCM compared to the SRCM. Red lines are scenarios using only six regions, and blue lines are scenarios using 15. Labels on the *y*-axis describe the simulations settings for the number of sites and detectability settings (*p* sd), and the *x*-axis shows the simulated β coefficient values for the species richness relationship.

114) to 121 (95% BCI: 91–184). Under the multiregion framework, the size of the eight communities was positively related to habitat complexity, and the coefficient for the relationship was different from zero ($\beta_{\Omega_{MRCM}} = 0 \cdot 475$, 95% BCI: 0·017–0·948). On the contrary, under the post hoc regression, the 95% confidence interval (CI) for the coefficient encompassed zero ($\beta_{\Omega_{SRCM}} = 0 \cdot 356$, 95% CI: −0·737–1·450).

Both models suggest that species detection was imperfect. Species richness estimates from the multiregion model had narrower credible intervals compared to the traditional single-region model approach. Notably, the approach of doing statistics on statistic (the two-stage approach), resulted in, as is expected, over confident predicted relationships relative to the uncertainty around the point estimates because this uncertainty is not taken into consideration (note the narrow intervals relative to the point estimates and associated 95% BCI, Fig. 3).

Discussion

We extended the multispecies (community) occupancy model to allow for joint estimation of species richness across multiple geographically distinct regions as a function of spatially varying covariates, while formally accounting for

Table 2. Summaries of posterior distributions from the multiregion community model (MRCM) applied to the avian community data from eight regions. 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, that is, $\bar{\psi} = \text{expit}(\mu_\psi)$ and $\bar{p} = \text{expit}(\mu_p)$, where expit is the inverse-logit function

			Quantiles		
Parameter	Mean	SD	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

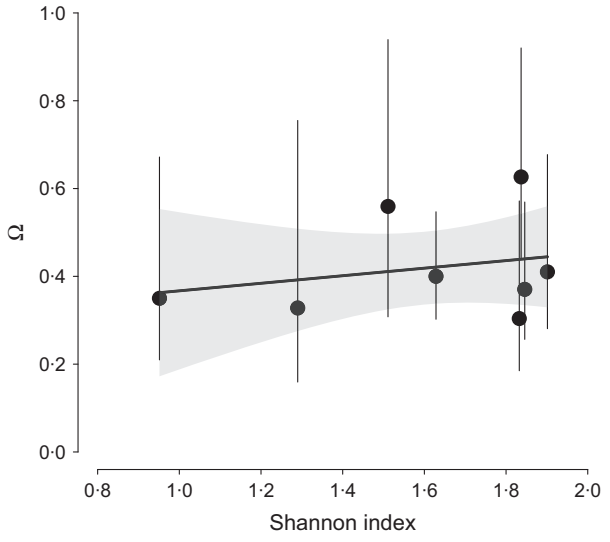

Fig. 3. Relationship between the zero-inflation parameter (Ω) and the Shannon entropy index (SEI) as a proxy of habitat complexity, derived from the post hoc regression of point estimates obtained from multiple single-region models (SRCM) separately. Shaded area indicates 95% confidence interval. Black circles in indicate Ω mean values estimates from the single region models, with 95% credible intervals (vertical lines). The solid regression line indicates the predicted relationship between Ω and the SEI from the post hoc approach.

imperfect species detection. We demonstrated, using simulation, the benefits of the multiregion community model (MRCM): (i) it can be used to jointly analyse data from multiple regions allowing formal comparison of species richness among regions that takes into account uncertainty in community size and allows sharing of information and parameters across sites, species and regions; (ii) it allows the development of models for explicitly testing hypotheses about drivers of spatial variation in species richness; (iii) the model yields improved statistical performance.

One of the advances made here is that our multiregion approach for estimating community size and composition naturally accounts for the fact that some species may not be shared by all communities, while others may occur in multiple; this is not the case for existing methods that account for imperfect detection. For example, Dorazio *et al.* (2010) invoke the concept of a metacommunity to motivate the development of multispecies occupancy models, in which the metacommunity is made up of a collection of sample locations considered to be 'local' communities nested within a 'super' community, that is the total number of species in a single region. Extending this approach to accommodate multiple regions in a single analysis, that is, by stacking the data side-by-side, is possible but lacks generality due to the fact that certain species-region combinations might not be permitted, and the fact that identity of unobserved species is not known making it difficult to identify which region unobserved species belong to. More fundamentally, however, this approach would preclude modelling species richness directly in the way we have proposed (e.g. Zipkin *et al.* 2010). The MRCM model can therefore 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). Our treatment of regions as independent in terms of species composition permits the investigation of community size, structure and dynamics across a range of geographic scales such as landscape-scale patch networks (Tscharntke *et al.* 2007), national-scale reserve networks (Sierra, Campos & Chamberlin 2002) or global-scale biodiversity monitoring networks (Ahumada *et al.* 2011). Moreover, and perhaps of larger consequence, is the fact that the MRCM provides a mechanism for explicitly modelling spatial variation in species richness.

An alternative approach would be analogous to the 'stratified population' model of Royle & Converse (2014) for estimating which population, of several, each individual, including unobserved individuals, belongs to. This approach assumes, however, that each unique individual belongs to exactly one population, which is restrictive in the context of species and communities (compared to individuals and populations), because species identity cannot be preserved across regions when, in truth, distinct communities can often share one or more species. The ability to retain species identity across regions 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 (Lomolino 1996; Cam *et al.* 2000). Interestingly, integrating the 'stratified community' model and the MRCM would allow within-region community structure to be estimated (e.g. guild structure, Yamaura *et al.* 2011), but would also permit modelling variation in strata-specific species richness across regions as a function of covariates, with total species richness being derived as the sum of strata-specific species richness (S. Tenan *et al.* in review).

Using an integrated multiregion approach, communities can be compared across any scale to investigate a wide range of

important environmental influences on biodiversity. We demonstrated this using the multicommunity bird data set, finding that larger communities were found in more heterogeneous habitats. Other potentially useful applications are the investigation of variation in species richness across climate 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 (Murray, Ambrose & Bohnsack 1999; Pence, Botha & Turpie 2003). 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).

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-stage approach of 'doing statistics on statistics' has been repeatedly criticized (Link 1999; Link & Barker 2004; Grosbois *et al.* 2008; Brooks, Deroba & Wilberg 2015) and a hierarchical approach is recommended (e.g. Royle & Dorazio 2008; Cooch *et al.* 2012). 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 hierarchical approach can be advantageous, particularly in reducing the risk of finding spurious results and of overconfidence in the precision of results. This can be seen in the analysis of the Italian bird count data: overconfidence in the relationship between species richness and the Shannon Entropy Index results from the fact that the post hoc method disregards the uncertainty in the estimates of species richness (Fig. 3) leading to confidence intervals that are markedly narrower than expected when considering the uncertainty associated with the point estimates.

We note that Tobler *et al.* (2015) recently proposed a 'multi-session multispecies' 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 groups 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 multiregion framework is not subject to these limitations and is arguably more appropriate for investigating spatial variation in community size and structure, particularly when the number of regions is small (Fig. 2, Table 1).

The multiregion 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 multiregion 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, for example, 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 hypotheses 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 multiregion approach, however, can be used to define metacommunity-level relationships of species richness, community structure and species occurrence patterns that are spatially explicit. Because the multiregion 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).

Acknowledgements

We thank Aaron Iemma for IT assistance and Jim Nichols and two anonymous referees for the 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). This research was partially funded through the project LIFE+ project 'LIFE11/IT/187 T.E.N. (Trentino Ecological Network): a focal point for a Pan-Alpine Ecological Network', by Servizio Agricoltura PAT and by 'Accordo di Programma per la Ricerca PAT- MUSE (2013-2014)'. Finally, we are grateful to A. Franzoi, F. Rizzolli, F. Rossi, for the help with fieldwork.

Data accessibility

R and BUGS script for running the model: included as an online supplement. Input data for the BUGS model: Figshare doi: http://dx.doi.org/10.6084/m9.figshare.1603456.

References

Ahumada, J.A., Silva, C.E., Gajapersad, K., Hallam, C., Hurtado, J., Martin, E. *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*, **366**, 2703–2711.

Araújo, M. & Rahbek, C. (2006) How does climate change affect biodiversity? *Science*, **313**, 1396–1397.

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**, 1146–1156.

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.
Cam, E., Nichols, J.D., Hines, J.E., Sauer, J.R., Alpizar-Jara, R. & Flather, C.H. (2002) Disentangling sampling and ecological explanations underlying species-area relationships. *Ecology*, **83**, 1118–1130.
Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.J. (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.
Cooch, E.G., Conn, P.B., Ellner, S.P., Dobson, A.P. & Pollock, K.H. (2012) Disease dynamics in wild populations: modeling and estimation: a review. *Journal of Ornithology*, **152**, 485–509.
Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.
Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.
Dorazio, R.M., Kéry, M., Royle, J.A. & Plattner, M. (2010) Models for inference in dynamic metacommunity systems. *Ecology*, **91**, 2466–2475.
Field, R., Hawkins, B.A., Cornell, H.V., Currie, D.J., Diniz-Filho, J.A.F., Guégan, J.F. *et al.* (2009) Spatial species-richness gradients across scales: a meta-analysis. *Journal of Biogeography*, **36**, 132–147.
Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 1–19.
Gotelli, N. & Colwell, R. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.
Gould, W.R. & Nichols, J.D. (1998) Estimation of temporal variability of survival in animal populations. *Ecology*, **79**, 2531–2538.
Grosbois, V., Gimenez, O., Gaillard, J.M., Pradel, R., Barbraud, C., Clobert, J., Møller, A. & Weimerskirch, H. (2008) Assessing the impact of climate variation on survival in vertebrate populations. *Biological Reviews*, **83**, 357–399.
Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging methods to estimate species diversity. *Trends in Ecology and Evolution*, **29**, 1–10.
Johnson, M.P., Frost, N.J., Mosley, M.W., Roberts, M.F. & Hawkins, S.J. (2003) The area-independent effects of habitat complexity on biodiversity vary between regions. *Ecology Letters*, **6**, 126–132.
Kerr, J. (1997) Species richness, endemism, and the choice of areas for conservation. *Conservation Biology*, **11**, 1094–1100.
Kéry, M. & Royle, J. (2008) Hierarchical bayes estimation of species richness and occupancy in spatially replicated surveys. *Journal of Applied Ecology*, **45**, 589–598.
Knaus, J. (2013) snowfall: Easier cluster computing (based on snow). R package version 1.84-6.
Link, W.A. (1999) Modeling pattern in collections of parameters. *The Journal of Wildlife Management*, **63**, 1017–1027.
Link, W.A. & Barker, R.J. (2004) Hierarchial mark-recapture models: a framework for inference about demographic processes. *Animal Biodiversity and Conservation*, **27**, 441–449.
Lomolino, M. (1996) Investigating causality of nestedness of insular communities: selective immigrations or extinctions? *Journal of Biogeography*, **23**, 699–703.
MacArthur, R. & Wilson, E. (1967) *The Theory of Island Biogeography, Volume 1 of Monographs in Population Biology*. Princeton University Press, Princeton, New Jersey, USA.
MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.
May, R. (1988) How many species are there on earth? *Science*, **241**, 1441–1449.
Murray, S., Ambrose, R. & Bohnsack, J. (1999) No-take reserve networks: sustaining fishery populations and marine ecosystems. *Fisheries*, **8446**, 37–41.
Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.
Nichols, J.D., Boulinier, T., Hines, J.E., Pollock, K.H. & Sauer, R. (1998) Inference methods for spatial variation in species richness and community composition when not all species are detected. *Conservation Biology*, **12**, 1390–1398.
Padoa-Schioppa, E., Baietto, M., Massa, R. & Bottoni, L. (2006) Bird communities as bioindicators: the focal species concept in agricultural landscapes. *Ecological Indicators*, **6**, 83–93.
Pence, G.Q., Botha, M.A. & Turpie, J.K. (2003) Evaluating combinations of on- and off-reserve conservation strategies for the Agulhas Plain, South Africa: a financial perspective. *Biological Conservation*, **112**, 253–273.
Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
Plummer, M. (2013) rjags: Bayesian graphical models using MCMC. R package version 3-10.
Purvis, A. & Hector, A. (2000) Getting the measure of biodiversity. *Nature*, **405**, 212–219.
R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Robert, C. & Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer, New York.
Royle, J.A. & Converse, S.J. (2014) Hierarchical spatial capture-recapture models: modelling population density in stratified populations. *Methods in Ecology and Evolution*, **5**, 37–43.
Royle, J.A. & Dorazio, R. (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, San Diego.
Royle, J.A., Dorazio, R.M. & Link, W.A. (2007) Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, **16**, 67–85.
Shannon, C.E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.
Sierra, R., Campos, F. & Chamberlin, J. (2002) Assessing biodiversity conservation priorities: ecosystem risk and representativeness in continental Ecuador. *Landscape and Urban Planning*, **59**, 95–110.
Stevens, G. (1989) The latitudinal gradient in geographical range: how so many species coexist in the tropics. *The American Naturalist*, **133**, 240–256.
Tobler, M.W., Zúñiga Hartley, A., Carrillo-Percastegui, S.E. & Powell, G.V. (2015) Spatiotemporal hierarchical modelling of species richness and occupancy using camera trap data. *Journal of Applied Ecology*, **52**, 413–421.
Tscharntke, T., Bommarco, R., Clough, Y., Crist, T.O., Kleijn, D., Rand, T.A., Tylianakis, J.M., van Nouhuys, S. & Vidal, S. (2007) Conservation biological control and enemy diversity on a landscape scale. *Biological Control*, **43**, 294–309.
Yamaura, Y., Andrew Royle, J., Kuboi, K., Tada, T., Ikeno, S. & Makino, S. (2011) Modelling community dynamics based on species-level abundance models from detection/nondetection data. *Journal of Applied Ecology*, **48**, 67–75.
Zipkin, E.F., Andrew Royle, J., Dawson, D.K. & Bates, S. (2010) Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation*, **143**, 479–484.

Received 4 July 2015; accepted 4 January 2016
Handling Editor: Nick Isaac

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Simulation results table documenting the performance of the multi- and single-region community models.

Appendix S2. R and JAGS script for fitting the multiregion model to real data.

Appendix S3. Data used in the application of the multiregion community model to detection non-detection data collected from bird assemblages in $R = 8$ geographically distinct areas (regions) in northern Italy are available from Figshare at http://dx.doi.org/10.6084/m9.figshare.1603456. We augmented the data for all regions such that $M = M_r = 200$, that is, by 200-n_r species for each region. The data file contains the following six objects: `dataREG`: region-specific covariates; `K_tot`: number of sampling occasions for each site in each region; `M`: total number of species in the augmented data set; `nsites`: total number of sites per region; `nspecies`: total number of observed species for each region; `Yaug_tot`: augmented species-by-sites-by-region detection frequency matrix.