Supplemental material 5

2023-06-30

The multi-city multi-species autologistic occupancy model

This model is almost exactly the same model as we had used in Magle et al. (2022). For s in $1, ..., S$ species and c in $1, ..., C$ cities, π_{sc} is the probability species s is within city c. Futher, let x_{sc} be a Bernoulli random variable that equals 1 if the species is within that city and is otherwise zero such that $x_{sc} \sim \text{Bernoulli}(\pi_{sc})$. We made π_{sc} a function of one covariate–the distance a city is from the known edge of a species extent–using the logit link. This covariate was positive if a city was within a species extent and negative if it was outside. We compiled range information from IUCN red list data (IUCN, 2020). Thus the linear predictor for this level of the model was

$$\text{logit}(\pi_{sc}) = \boldsymbol{d}_s \boldsymbol{h}_c$$

where $\boldsymbol{d}_s$ is a vector of species-specific covariates while $\boldsymbol{h}_c$ is a vector of conformable regression coefficients where the leading element is 1 to accomodate the model intercept.

As the first level of the model estimates a species presence at the city-level, the next level estimates species presence within cities conditional on their presence in a city. Given that the number of sites and sampling periods varies across cities, we add a city subscript to define i_c in $1, ..., I_c$ sites and t_c in $1, ..., T_c$ sampling periods. However, for simplicity we drop these specific subscripts while we explain the model. Additionally, let z_{scit} be a Bernoulli random variable and ψ_{scit} be the probability of occupancy. As such, $z_{scit} \sim \text{Bernoulli}(\psi_{scit} x_{sc})$. As with the previous level of the model, ψ_{scit} can be made a function of covariates via the logit link. As a depature from the Magle et al. (2022) parameterization, we added a first-order autologistic term to account for any temporal dependence in occupancy status between adjacent sampling periods within a city. Thus, for the first time period in a city, the logit-linear predictor was

$$\text{logit}(\psi_{scit=1}) = \boldsymbol{\beta}_{sc} \boldsymbol{f}_{ci}$$

where $\boldsymbol{\beta}_{sc}$ is a vector of species and city-specific parameters and $\boldsymbol{f}_{ci}$ is a vector of conformable covariates whose first value is 1 for the model intercept. After the first sampling season, we added our autologistic term

$$\text{logit}(\psi_{scit}) = \boldsymbol{\beta}_{sc} \boldsymbol{f}_{ci} + \theta_{sc} z_{scit-1}, \text{ for t > 1.}$$

For the data model, y_{scit} was the number of days species s was detected at city c and site i on sampling season t. Given j_{cit} days of sampling, we assumed y_{scit} is a binomial random variable conditional on species presence

$$y_{scit} | z_{scit} \sim \text{Binomial}(j_{cit}, \rho_{scit} z_{scit})$$

where ρ_{scit} is the daily probability of detection that can be made a function of covariates with the logit link,

$$\text{logit}(\rho_{scit}) = \boldsymbol{\alpha}_{sc} \boldsymbol{g}_{ci}$$

where $\boldsymbol{\alpha}_{sc}$ is a vector of parameters and $\boldsymbol{g}_{ci}$ is a vector of conformable covariates where the leading value is 1 to accommodate the intercept.

Hierarchical parameterization of model parameters

Given that some species occur across multiple cities, there are multiple levels in which species could partially share information. At the top-level of the model we have the simplest hierarchical parameterization for parameters associated to π_{sc}, namely that there is a community mean for each parameter of which species-level coefficients vary around. We show this for the model intercept with the understanding that the same parameterization applies to all logit-scale covariates in this part of the model.

$$\begin{aligned}
\bar{d}_{0s} &\sim \text{Cauchy}(0, 2.5) \\
\sigma_{d_0} &\sim \text{Inv-Gamma}(1, 1) \\
d_{0s} &\sim \text{Normal}(\bar{d}_{0s}, \sigma_{d_0})
\end{aligned}$$

For the rest of the latent state model we add an additional hierarchical level to the model. However, this parameterization also aligns with the data model, and as such so we only describe it here once for the model intercept of the latent-state model. For example, for the model intercepts we begin with a community-level average among species and cities ($\bar{\beta}_0$). This parameter partially informs a species-level average among cities ($\bar{\beta}_{0s}$), which then informs species-specific coefficients in all cities (β_{0sc}).

$$\begin{aligned}
\bar{\beta}_0 &\sim \text{Cauchy}(0, 2.5) \\
\sigma_{\beta_0} &\sim \text{Inv-Gamma}(1, 1) \\
\bar{\beta}_{0s} &\sim \text{Normal}(\bar{\beta}_0, \sigma_{\beta_0}) \\
a_{\beta_0} &\sim \text{Uniform}(0, 10) \\
b_{\beta_0} &\sim \text{Uniform}(0, 10) \\
\sigma_{\beta_{0s}} &\sim \text{Inv-Gamma}(a_{\beta_0}, b_{\beta_0}) \\
\beta_{0sc} &\sim \text{Normal}(\bar{\beta}_{0s}, \sigma_{\beta_{0s}})
\end{aligned}$$

We added the hyperparameters for the shape (a_{β_0}) and rate (b_{β_0}) of the Inv-Gamma distribution to account for the fact that some cities may only have one sampling period of data. Note that the above parameterization also applies to the autologistic term of the model (θ_{sc}).

Finally, the latent state and data model included one other set of parameters to account for variation in occupancy or detectability across sampling seasons. As we have already added hierarchical structure via the centered parameterization of the other model parameters, we usd a non-centered parameterization here for this level of variability. Again, we show this for the latent state, but with a swapping of subscripts this could readily be applied to the data model as well.

$$\begin{aligned}
a_{\psi} &\sim \text{Uniform}(0, 10) \\
b_{\psi} &\sim \text{Uniform}(0, 10) \\
\sigma_{\psi c} &\sim \text{Inv-Gamma}(a_{\psi}, b_{\psi}) \\
\beta_{sct} &\sim \text{Normal}(0, \sigma_{\psi c})
\end{aligned} \tag{1}$$

With this parameterization, β_{sct} is a difference term that represents the logit-scale difference in occupancy for species s at city c from their average β_{0sc} (i.e., the model intercept). Again, like with the other parameters in this part of the model, we used hyperparameters for the Inv-Gamma distribution because not every city had more than one season of data.

The alpha diversity meta-analytic model

From our occupancy model we created a posterior distribution for the latent state of each species at each site across all cities (z_{scit}). From this, we derived two quantities for this model:

1. The mean expected species richness at each site across all sampling periods (r_{ci}). To do so, we calculated the number of unique species detected in city c and site i across the t sampling seasons and took the median across 5000 posterior samples.
2. The standard deviation of the first quantity across those 5000 posterior samples (σ_{ci}). This quantifies our level of uncertainty with the first estimate.

Following this, let $\boldsymbol{\beta}_c$ be a vector of city-specific regression coefficients and $\boldsymbol{x}_{ci}$ be a vector of city and site specific covariates where the leading element is 1 to account for the intercept. Within the linear predictor we also added an additional residual variation term, ϵ_{ci}, which was given a $\sim$ Inv-Gamma$(1, 1)$ prior. Thus, the log-linear predictor was

$$\log(\mu_{sci}) = \boldsymbol{\beta}_c \boldsymbol{x_{ci}} + \epsilon_{ci}$$

and following Kery and Royle (YEAR), we accounted for variability in the response variable with an additional level of the model

$$r_{ci} \sim \text{Normal}(\mu_{sci}, \sigma_{ci})$$

We treated the intercept and slope terms as city-level random effects. For example, the prior for the model intercept was

$$\begin{aligned}\bar{\beta}_0 &\sim \text{Cauchy}(0, 2.5)\\ \sigma_{\beta_0} &\sim \text{Inv-Gamma}(1, 1)\\ \beta_{0c} &\sim \text{Normal}(\bar{\beta}_0, \sigma_{\beta_0})\end{aligned}$$

and the same specification was also applied to the slope terms as well (though not described here).

The beta diversity meta-analytic model

From our occupancy model, we created a posterior distribution for the latent state of each species at each site across all cities (z_{scit}). From this, we derived three quantities:

1. The mean pairwise Sorensen dissimilarity between pairs of sites within each city across all sampling periods (v_{cik}, where the subscript k denotes one of the sites in city c that is not the ith site). To do so, we calculated the number of unique species detected in city c and site i across the t sampling seasons. Following this, we calculated the Sorensen dissimilarity metric among all pairs of sites within each city using `vegan` in `R` across 5000 posterior samples. Finally, we took the median across all posterior samples for each site-pair
2. The standard deviation of the first quantity across those 5000 posterior samples (σ_{cik}). This quantifies our level of uncertainty with the first estimate.
3. The mean expected number of unique species between a site-pair (w_{cik}).

To estimate pairwise dissimilarity as a function of covariates we modified a generalized dissimilarity model to account for parametetric uncertainity of the response variable v_{cik} (GDM citation). To do so, GDMs estimate the relationship between dissimilarity and environmental or spatial differences between pairs of sites with a clog link (Mokany et al. 2022):

$$d_{cik} = 1 - \exp(-\eta_{cik})$$

where d_{cik} is the biological dissimilarity between sites i and k within city c and η_{cik} is the predicted ecological distance between (i.e., the linear predictor). Given p in 1,..,P covariates, the ecological distance between sites is

$$\eta_{cik} = b_0 + \sum_{p=1}^{P} |f_p(x_{cip}) - f_p(x_{ckp})|$$

where b_0 is the model intercept (i.e., the expected pairwise dissimilarity between sites with identical environments). Covariates are further transformed within GDMs which 1) uses I-spline basis functions (Ramsay, 1988) and 2) constrains slope terms to be non-negative. Doing so allows the effect of each covariate to non-linearly vary while also ensuring that beta diversity increases monotonically as sites are more different from one another (a core assumption of this model). More specifically, the I-spline basis function for predictor p with 3 basis functions (the default used for GDMs) is

$$fp(x_{cip}) = \sum_{j=1}^{3} a_{cpj} I_{pj}(x_{cip})$$

where a_{cpj} is a non-negative coefficient for the jth I-spline and I_{pj} is the jth I-spline of the covariate x_{cip}. For our own model, the binary gentrification status of site-pairs was not sent through an I-spline basis function. Instead, we used a dummy variable that took the value of 1 if a pair of sites differed in their gentrification status and was otherwise 0. Regardless, all slope terms were constrained to be non-negative. Thus, given η_{cik} and the total number of species between site pairs (w_{cik}), the first level of our model is

$$\begin{aligned} d_{cik} &= 1 - \exp(-\eta_{cik}) \\ \sigma_{cik} &= \sqrt{\frac{d_{cik} - (1 - d_{cik})}{w_{cik}}} \\ \mu_{cik} &\sim \text{Half-Normal}(d_{cik}, \sigma_{cik}) \end{aligned} \tag{2}$$

where σ_{cik} is the binomial variance function (Mokany et al. 2022) and the half-Normal is constrained to be non-negative. Following this, we account for variation in the measurement of our response variable

$$v_{cik} \sim \text{Half-Normal}(\mu_{cik}, \sigma_{cik})$$

where again σ_{cik} is measured based on the output of the occupancy model and provided as data to this model.

We treated all coefficients in the linear predictor as city-level random effects. For example for the model intercept the prior specification would be $b_c \sim \text{Half-Normal}(\bar{b}, \sigma_b)$ where $b \sim \text{Half-Normal}(0, 10)$ and $sigma_b \sim \text{Inv-Gamma}(1, 1)$. The Half-Normal distributions ensure that the coefficients will be non-negative.

References

Magle et al (2022) IUCN (2020)