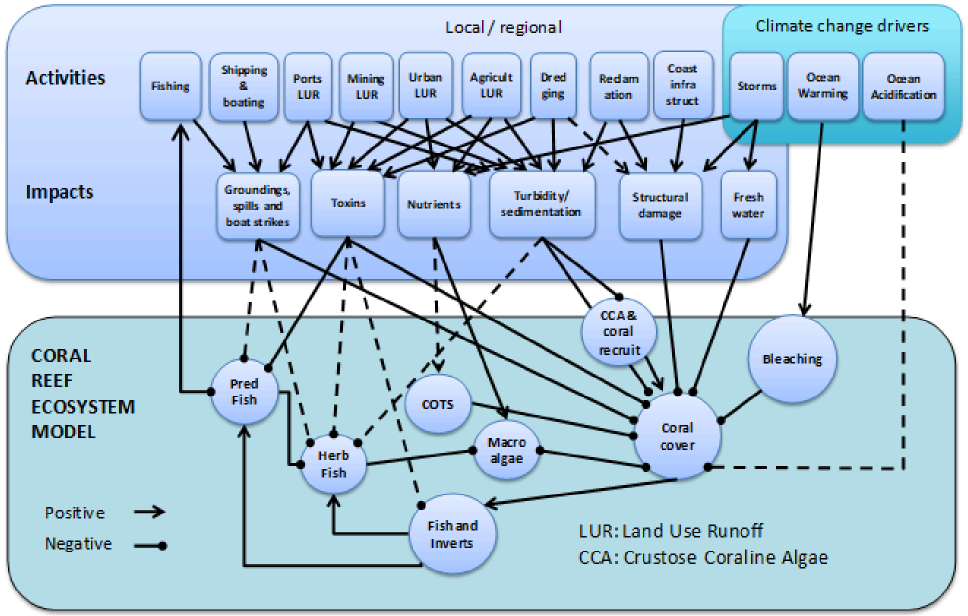
Species archetype models for presence-only data

Skipton Woolley, Piers Dunstan, David Warton & Scott Foster

25-11-2025

Managing complex ecosystems

- Managing anthropogentic pressures often requires information on the distribution of key natural values.
- Presence-only datasets provide a valuable source of information for describing these natural values.
- At broad spatial scales we often need simple, but clear ways to quantify these values.



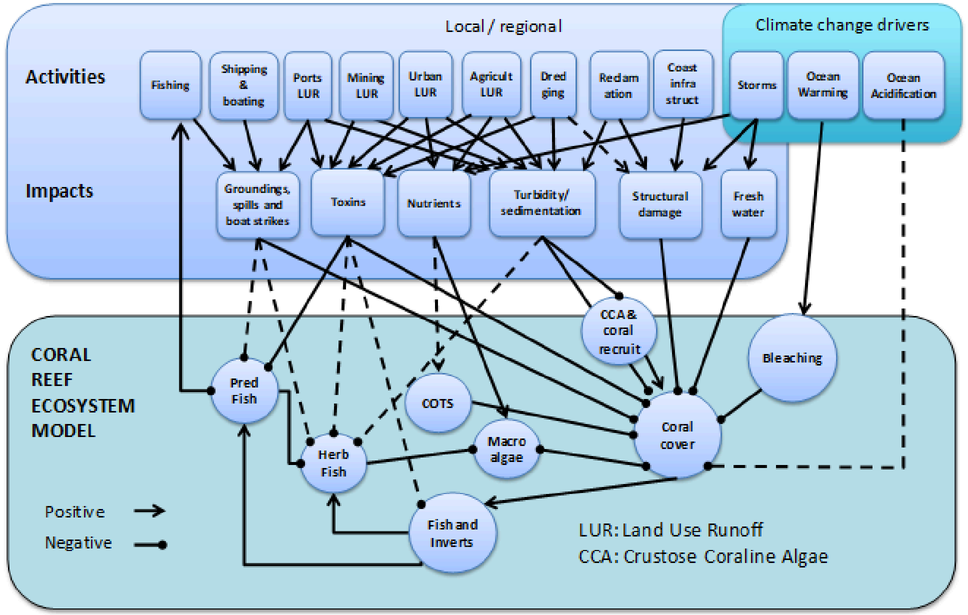
Ecosystems are complex

Managing complex ecosystems

For managing the environment inferences about ecosystem are required. Often questions are about unobserved properties.

- Assemblages.
- Ecoregions / bioregions.
- Functional groups,
- Species groups.
- Communities.
- Genetic groups

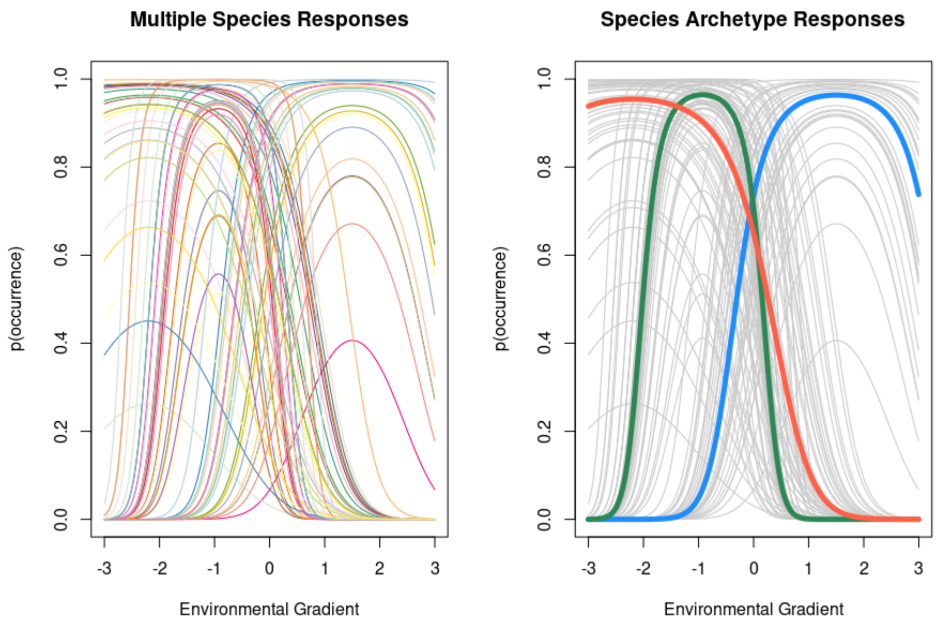
But none of these are **observed**. Our solution is to pose statistical models containing these/related constructs. Management could be targeted at a fewer **latent** groups rather than a large number of individual species.



Ecosystems are complex

Species Archetype Models (SAMs)

- Multivariate response (conditionally independent)
- Mixture of regressions, aka Species Archetype Models (SAMs)
    - Grouping *species* according to their responses to the environment
    - A relatively simple model used to understand how many species jointly respond to environmental conditions



Clustering of species along a one dimensional gradient using SAMs

Species Archetype Models (SAMs)

- Soft assignments (probabilistic)
- Intuitively:
  - Perform a regression on each species, then
  - Cluster the regression coefficients
- Finite mixture models allow for a one-step process
  - Uncertainty propagation
  - Statistical efficiency
  - Likelihood based model selection and diagnostics

Poisson process species archetype models

- Let us define $\mathbf{y}_j = (y_{1j}, \dots, y_{Nj})^\top$ as a vector of species $j \in \{1, \dots, S\}$ presence-only occurrences at observations at N_j locations in region $\mathcal{A} \in \mathbf{R}^2$.
- We assume there are P covariates observed at all sites $\mathbf{x}_i$.

The likelihood contribution of the j^{th} species is:

$$\sum_{k=i}^{K} \pi_k \prod_{i=1}^{N_j} f(\boldsymbol{\beta}_k; \mathbf{y}_j)$$

- where $f(\boldsymbol{\beta}_k; \mathbf{y}_j)$ is a function of the conditional intensity λ_{ijk} of species j at location i, conditional on archetype k.
- π_k is the mixing proportion (satisfying $\pi_k \in (0, 1)$ and $\sum_{k=1}^{K} \pi_k = 1$), determining the proportion of species classified into each of the K archetypes.

Poisson process species archetype models

The **log-conditional intensity** is modelled as:

$$\log(\lambda_{ijk}) = \alpha_j + \mathbf{x}_i^\top \boldsymbol{\beta}_k + \mathbf{u}_i^\top \boldsymbol{\delta} + \nu_i$$

- $\lambda(i, s, k)$ is the intensity function for each species, as each site, conditional on each archetype k.
- α_j represents the species specific intercept.
- $\mathbf{x}_i$ represents environmental/habitat observed covariates.
- $\boldsymbol{\beta_k}$ is a vector of archetype specific coefficients associated with the environmental/habitat effects.
- $\mathbf{u}_i$ represents a observation bias data, e.g distance from roads.
- $\boldsymbol{\delta}$ represents a spatial observation bias across all species occurrence records for $\{\mathbf{y_j}\}^S$.
- ν_i is a offset at site i, and can represent differences in spatial area or even a known 'plug-in' thinned process.

Poisson process species archetype models

We estimate this model via numerical approximation, the approximate log-likelihood for the j^{th} species and k^{th} archetype as,

$$\log f_{jk}(\boldsymbol{\beta}, \boldsymbol{\delta}; \mathbf{y}_{\mathbf{p}_j}, \mathbf{y}_{\mathbf{0}_j}, \mathbf{w}_j) \approx \sum_{i=1}^{M_j} w_{ij}(z_{ij} \log(\lambda_{ijk}) - \lambda_{ijk})$$

- $\mathbf{y}_{\mathbf{p}_j}$ is a vector of species-specific presences,
- $\mathbf{y}_{\mathbf{0}_j} = \{y_{n_j+1}, \dots, y_{M_j}\}$ is a vector of q quadrature locations for each species
- $\mathbf{w}_j = (w_{j_1}, \dots, w_{j_m})$ stores the species-specific weights
- $M_j = N_j + q$ is the total number of presence *and* quadrature locations of the j^{th} species.

We used a grid based design for quadrature scheme (Berman & Turner 1992; Warton & Shepard 2010), but others could be used (e.g. Dirichlet tessellation).

Poisson process species archetype models

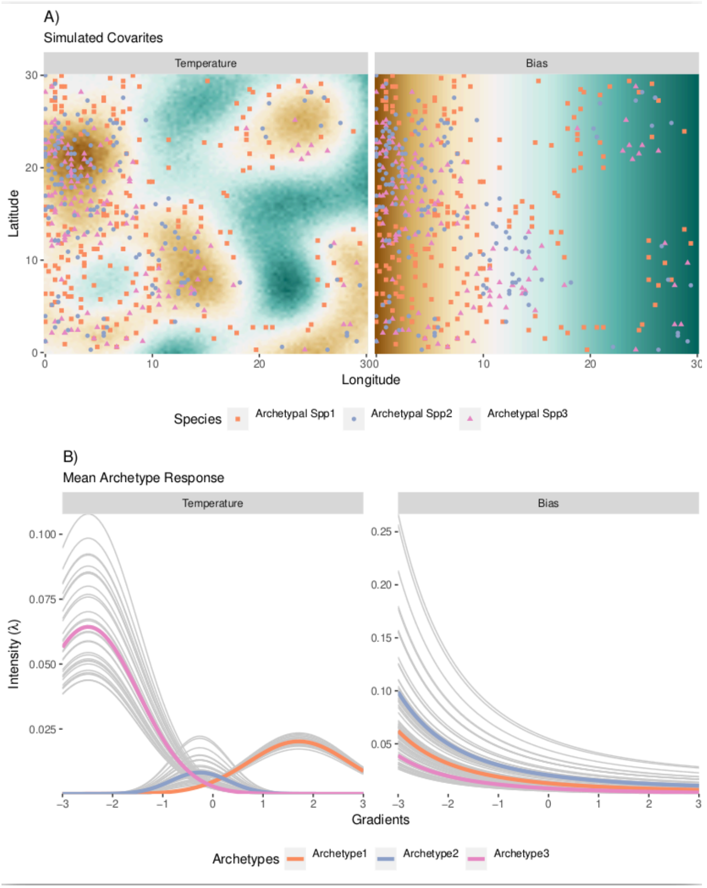
- Estimation is done via a hybrid Expectation-Conditional Maximisation and Newton-Raphson approach to estimate the above log-likelihood (Aitkin et al., 1996; Dunstan et al., 2013).
- We include in the initial starting values of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\pi}$.

The steps include:

1. finding starting values and adding a small amount of random noise within the standard deviation of the estimated starting values;
2. perform a limited number of initial ECM steps;
3. implement a Newton-Raphson maximiser until the model is converged
4. Doing this multiple fits per K (e.g. 10 fits per K)

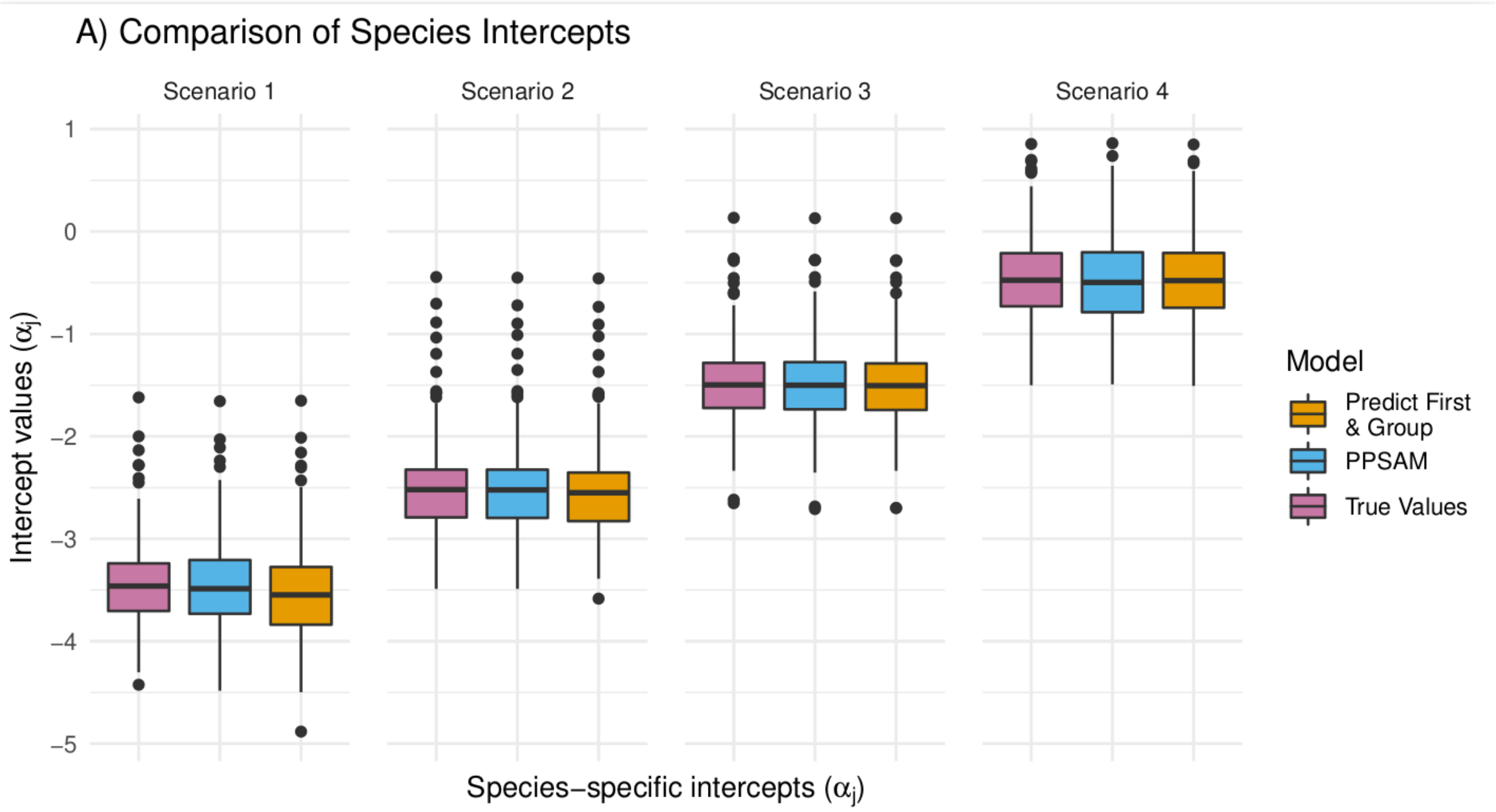
Simulation study

- We compared a two-stage approach (predicted and group) against a single PPSAM.
- Fitted four scenarios, each contained 1000 simulations.
- Scenarios based on the rarity of species, going from rare to more common.
- Each simulation contained 100 species within a simulated study area with a single environmental gradient, and single observation bias covariate.



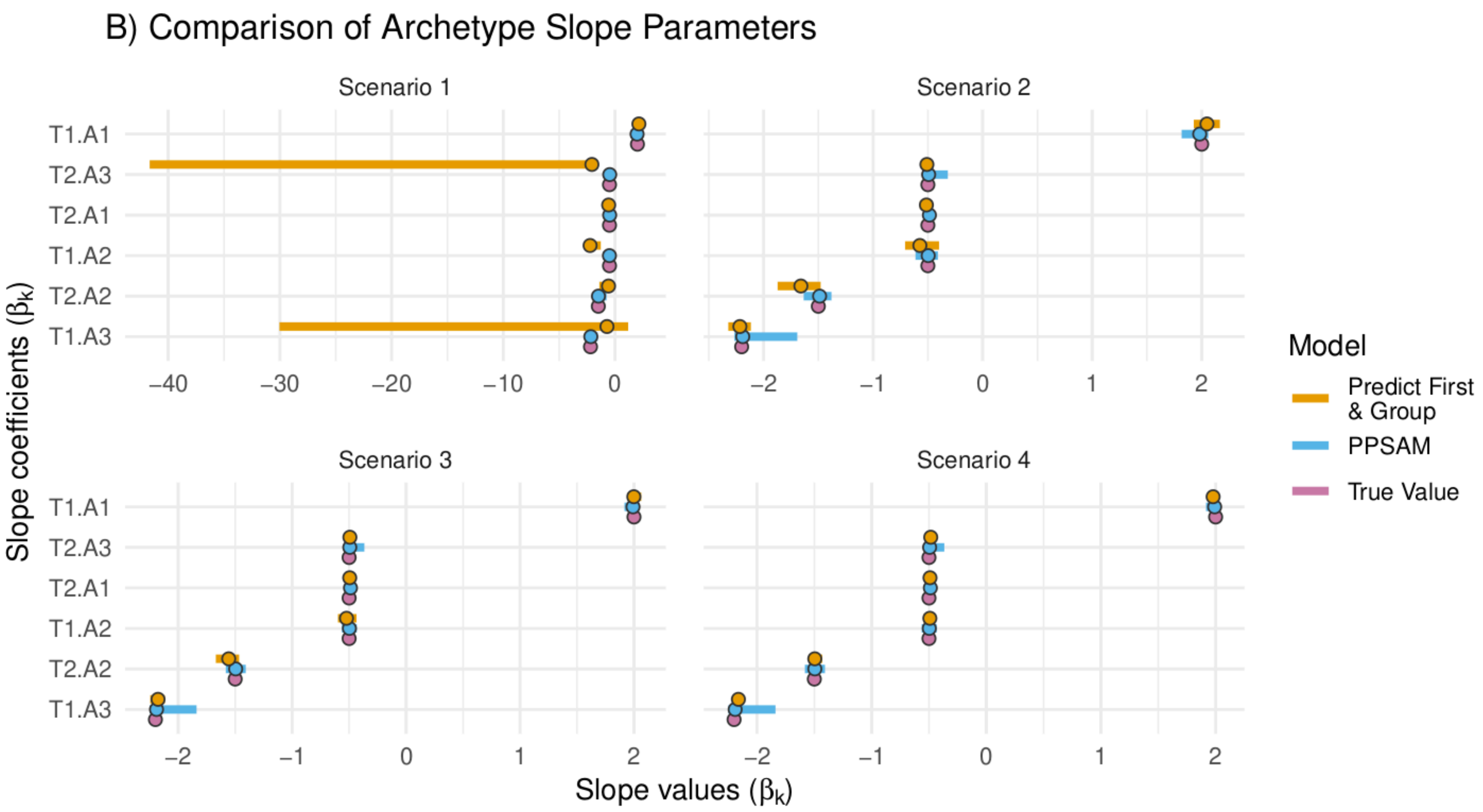
Simulated environmental and bias gradients. Three simulated archetypal responses.

Simulation study



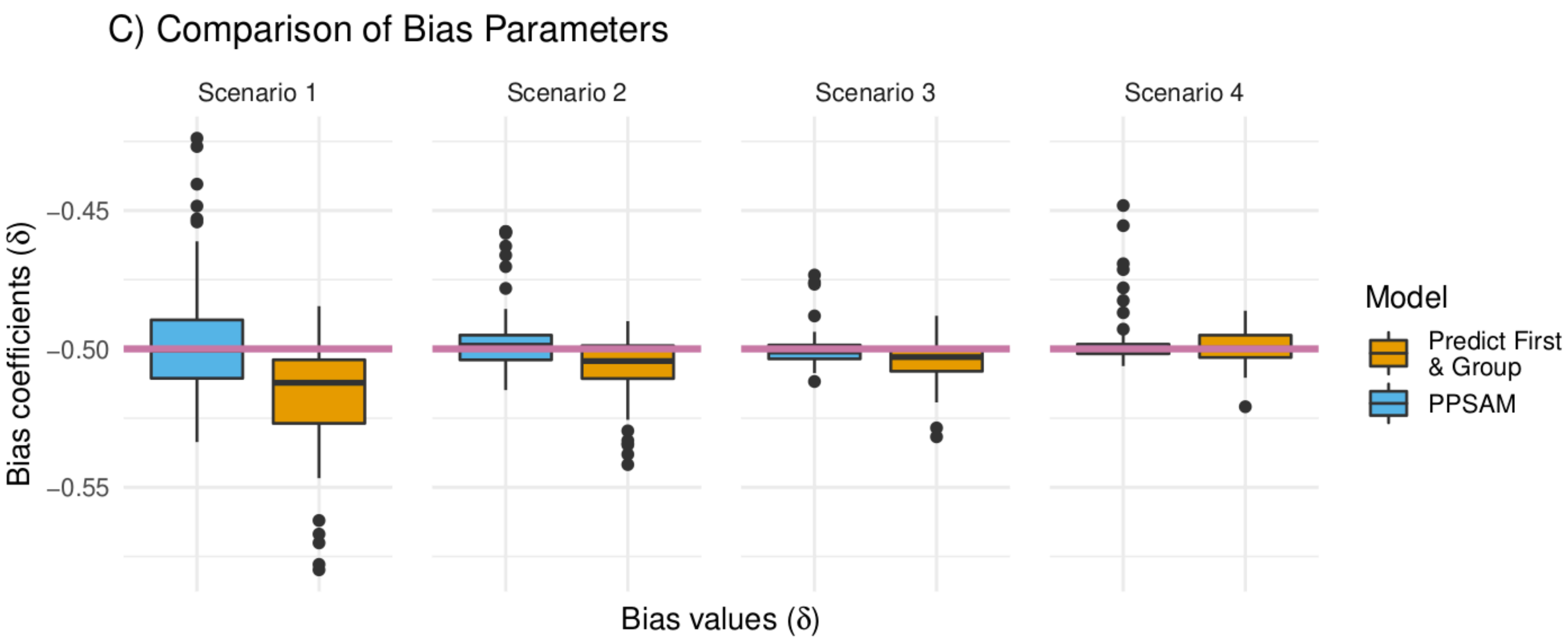
Estimated species-specific intercepts `$\alpha_j$` from the individual species-specific Poisson Processes and Poisson Process Species Archetype Models.

Simulation study



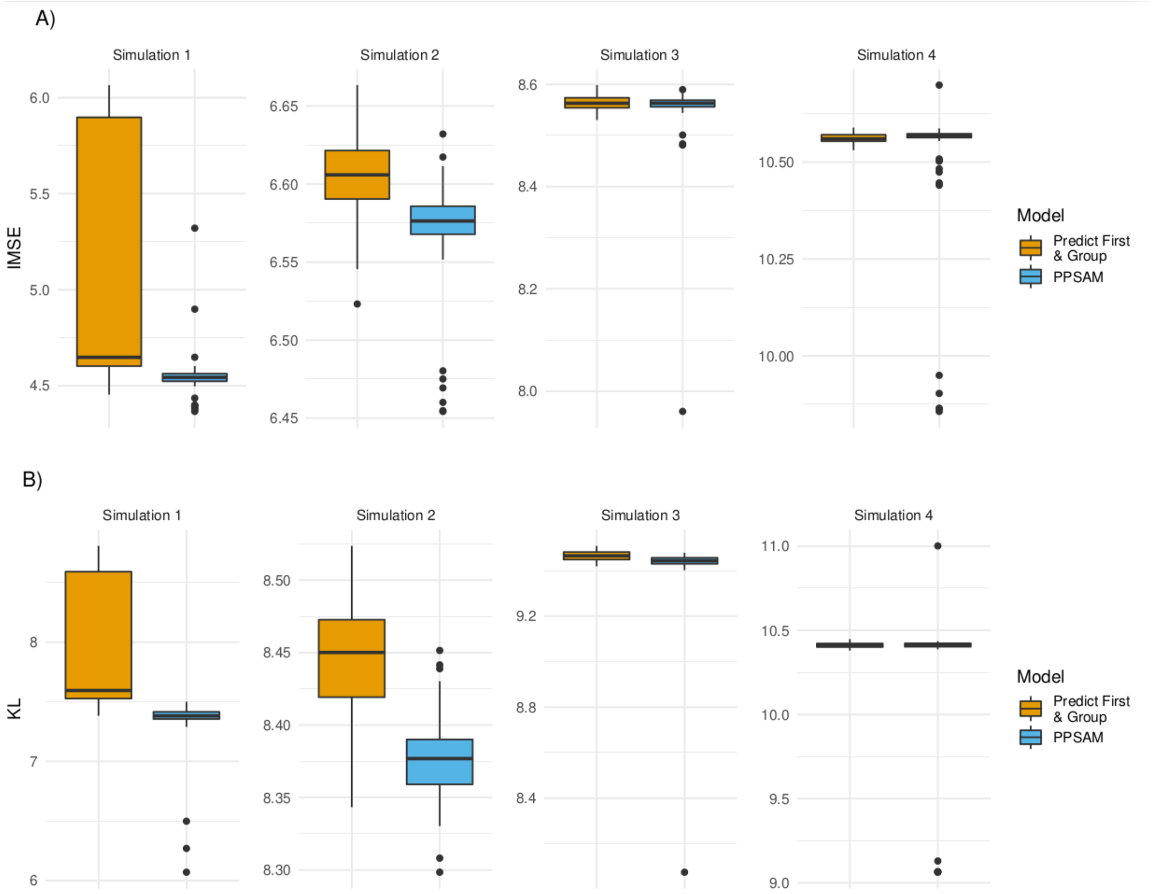
Estimated archetype level estimates `$\beta_k$` from the Predict First & Group approach, and Poisson Process Species Archetype Models.

Simulation study



Estimated bias covariates (`$\delta$`) we compared mean estimates (for each species) from Predict First & Group approach, and Poisson Process Species Archetype Models bias value (estimated across all species).

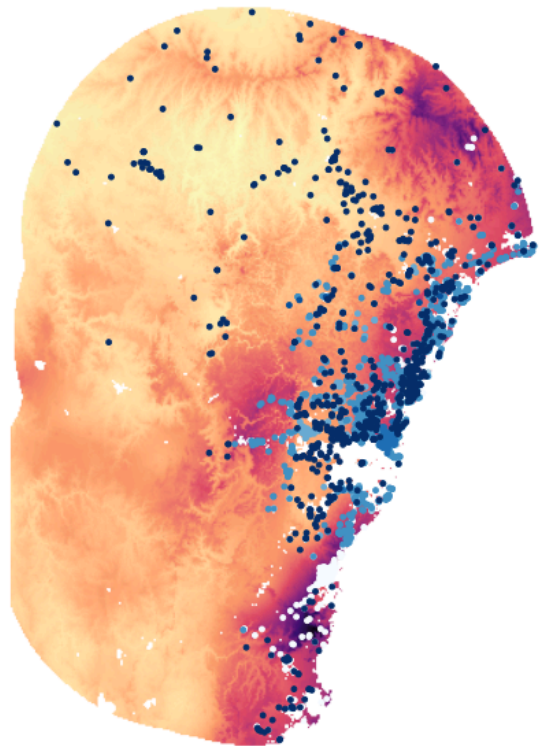
Simulation study



Predictive performance tests from the simulation study. A) Integrated Mean Square Error (IMSE) estimates for each of the four simulations. B) Kullback-Leibler (KL) divergence predictive scores summed across archetypes.

New South Wales Myrtacece Case Study

- Myrtacece dataset contains 41769 occurrences recorded for 296 species
- We took the 50 most common species with at least 100 presences
- We fitted either an PPSAM or 50 species-specific IPPMs
- The archetypes were fitted to fire frequency (FC), annual minimum temperature (MNT), annual maximum temperature (MXT), annual mean rainfall (Rain)
- Observation bias was fitted to distance from main roads (D.Main) & distance from main urban centres (D.Urb).



New South Wales Myrtacece study area

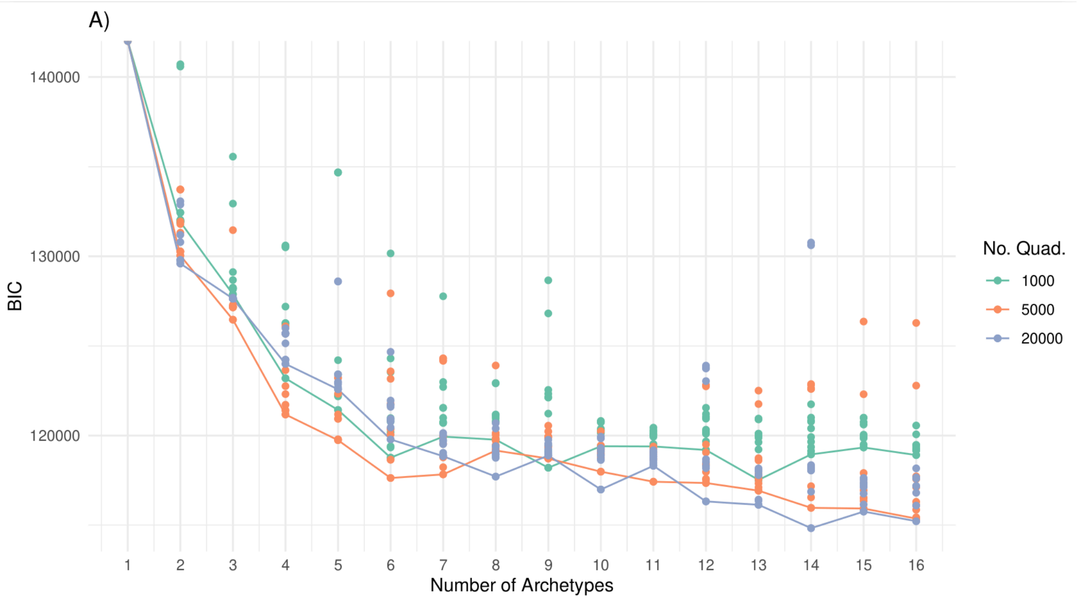
New South Wales Myrtacece Case Study

PPSAM Modelling steps.

- Multiple starts across 1 to 16 species archetype groups (k)
- Select model k based on BIC.
- Check for groups with zero membership.
- Diagnostics via random quantile residuals.

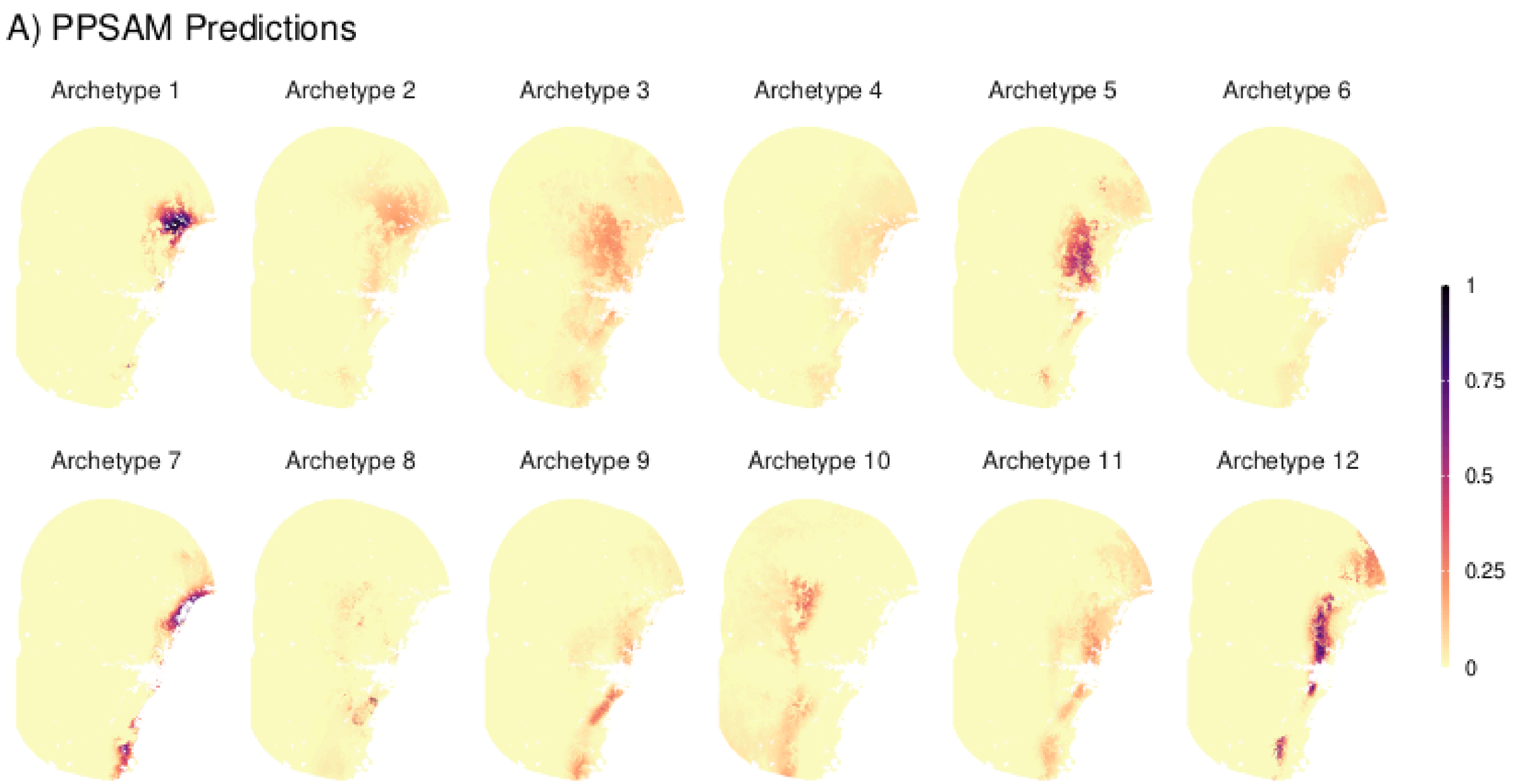
IPPM Modelling steps.

- Fit species-specific IPPMs.
- K-means cluster coefs.
- Select k based on BIC.



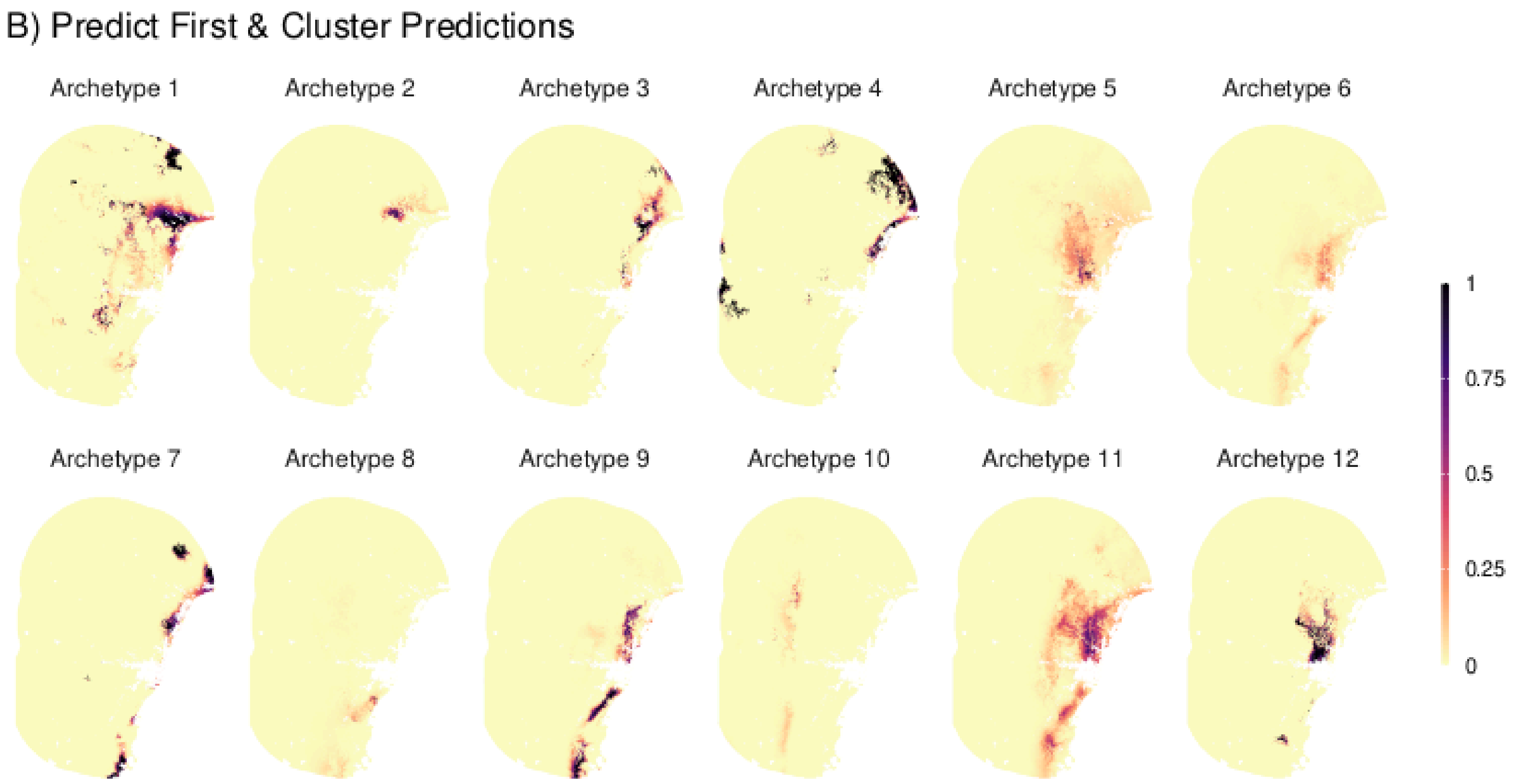
BIC from PPSAM fits

New South Wales Myrtacece Case Study



PPSAM Predictions

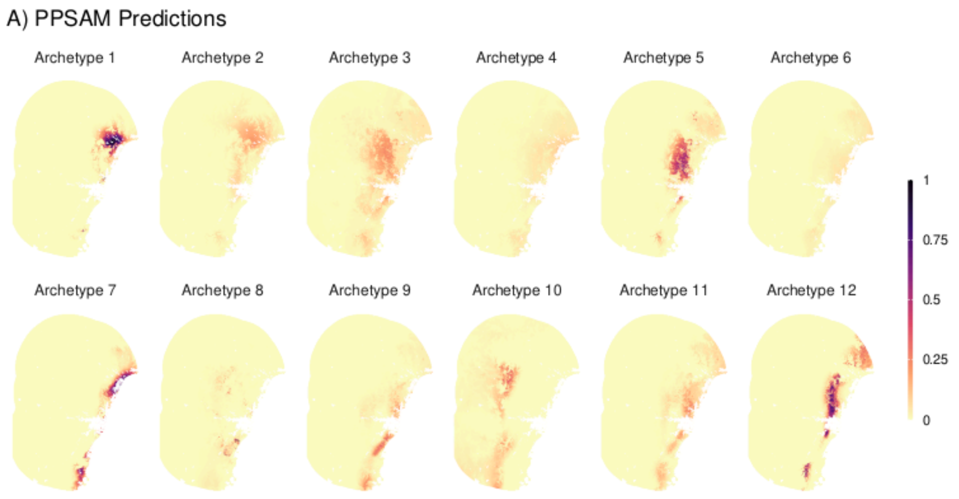
New South Wales Myrtacece Case Study



IPPM & Cluster Predictions

New South Wales Myrtacece Case Study

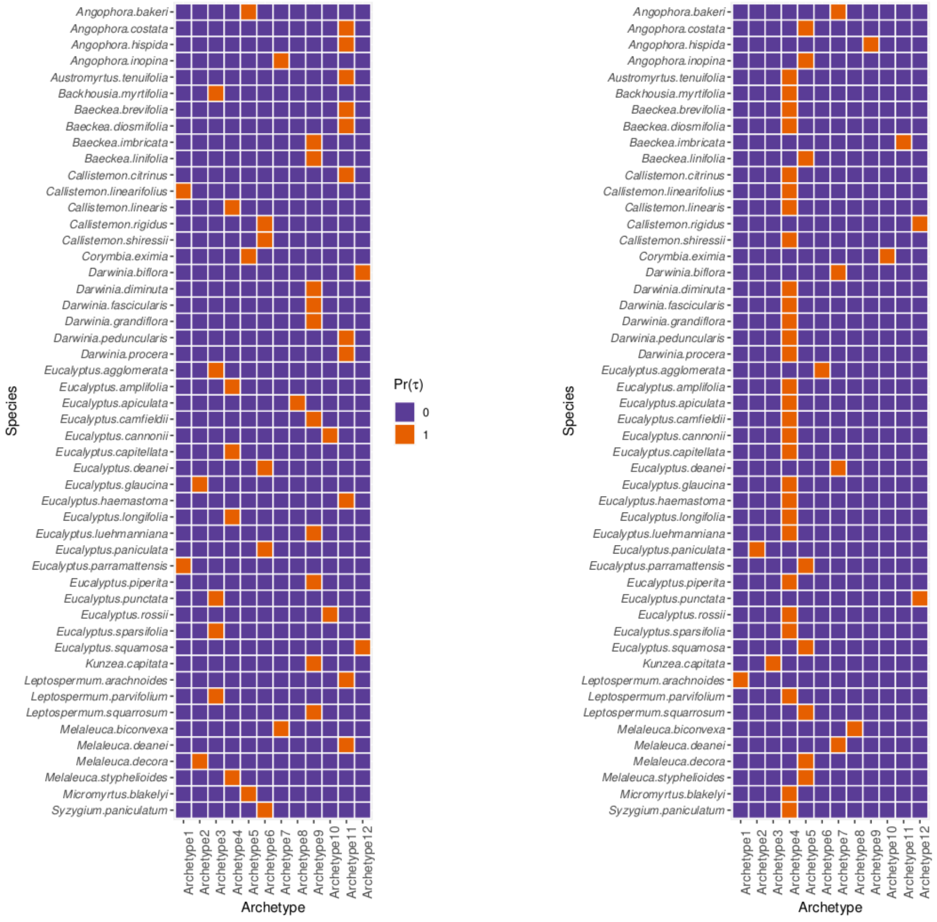
- A bias towards smaller range restricted groups in two stage approach.
- Two stage approach seems to select more clusters based on information criteria
  - Typically need to know the number of groups/clusters for this approach to work (e.g. Hill et al., 2020)
- Some ecologically relevant group such as:
  - Wet sclerophyll forest (Archetypes 5-7)
  - Dry sclerophyll woodland (Archetype 2-4)
  - Heathland/sandplain (Archetypes 8-10)



PPSAM Predictions

New South Wales Myrtacece Case Study

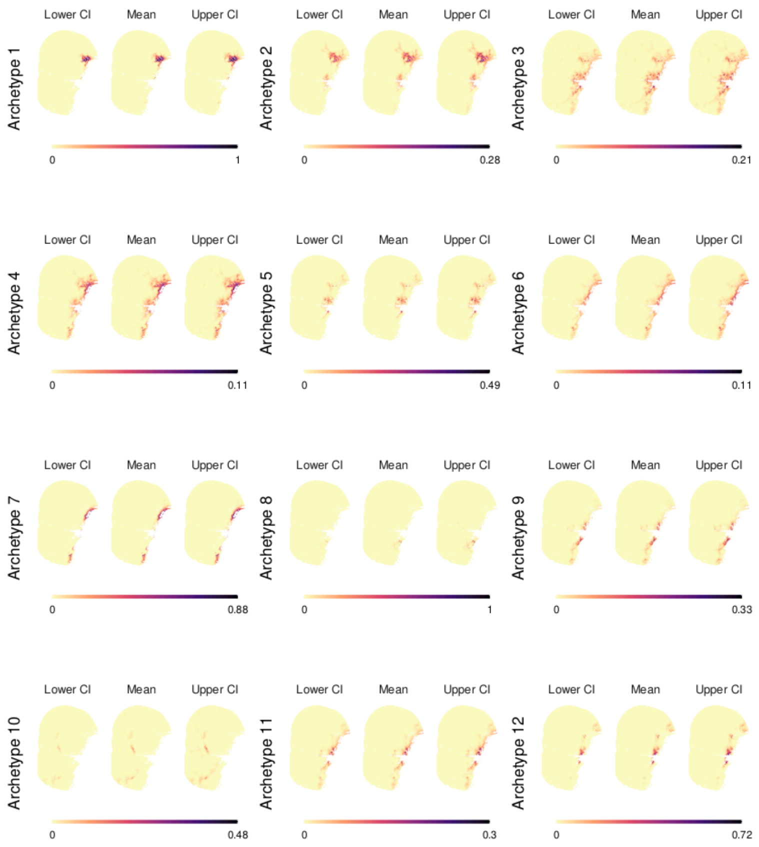
- Posterior probabilties of a species belong to an archetype/group is more well mixed under PPSAMs
- The two stage approach tends to lump most species into the same group and pull out extremes.



Species membership to each group

New South Wales Myrtacece Case Study

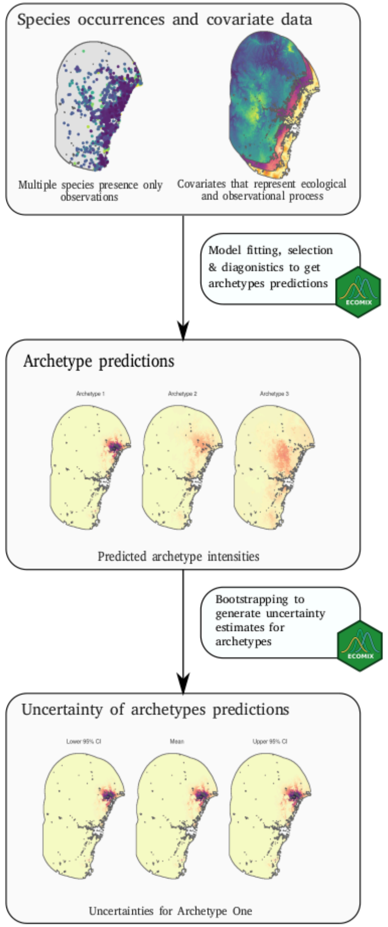
- Uncertainty quantification can be done via bootstrapping (Cowling et al., 1996).
- Two step approach fails to transfer data variance through to prediction at the clustering step.
- Bayesian approaches could deal with this, but you need to correct for label switching when clustering posterior predictions.



Uncertainty in PPSAM predictions

New South Wales Myrtacece Case Study

- We show that point process Species Archetype Models allow for the propagation of variance and uncertainty from the data through to predictions.
- Improving inference made on multiple species presence-only occurrence data.
- Or at least making it simpler to understand biodiversity patterns for a large number of species.
- However, it does not account for inter-species correlation in occurrence as you would see in a JSDM/GLLVM.



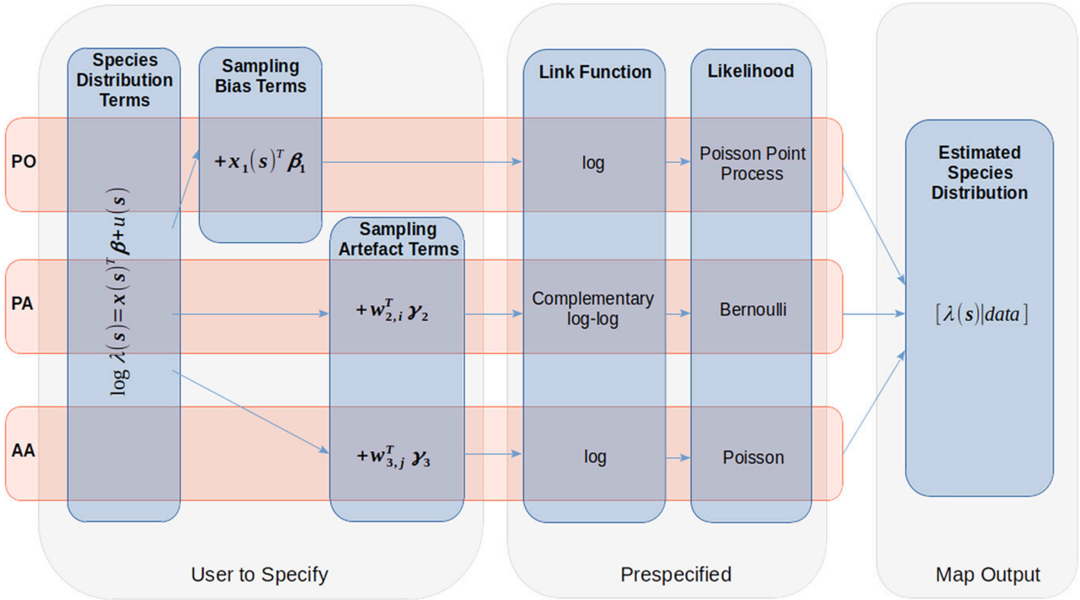
PPSAM framework

Some considerations and extensions

- Currently, PPSAM does not scale well due as number of species, archetypes and sites grow.
- Group selection/regularization is hard and is often important for presence-only data
- Luckily approximate SAMs (asSAM) are actively being developed to deal with these problem (Hui et al., In prep - see his talk tomorrow)

Some considerations and extensions

- Integrated with other data sources to help better correct observation bias (e.g Fithian et al., 2015)
- Point patterns naturally extends to spatial models
    - e.g. Log-Gaussian Cox Process
        - Latent GRF(s):
        - $Z(s) \sim \mathcal{GP}(0, C(s, s'))$ is a Gaussian Process (GP) with mean zero and covariance function C, capturing spatial (and temporal if in scope) dependencies.
    - GRF on what? On bias? on species? on archetypes? on multiple?
    - Identifiability might be a problem with many GRFs.
    - Approximation will be important, e.g Vecchia approximation/basis functions.



Example of an integrated single species model using the RISDM package; Foster et al., 2024

Some considerations and extensions

- How interpretable are species archetype models?
  - We tend to think of distribution in terms of composition, especially for characterisation
  - But, based on my experience experts and managers tend to think in terms of process, or at least can conceptualise this more easily.
  - Do we need to cluster or bicluster over composition and functional groups which allow us to lean heavier on ecological theory (e.g. ecosystem models/state-and-transition models) when trying to under stand how groups of species will respond to impacts/management?

References

- Aitkin, M. & Aitkin, I. (1996) A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. Statistics and Computing, 6, 127–130.
- Berman, M. & Turner, T.R. (1992) Approximating point process likelihoods with GLIM. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41, 31–38.
- Cowling, A., Hall, P. & Phillips, M.J. (1996) Bootstrap confidence regions for the intensity of a Poisson point process. Journal of the American Statistical Association, 91, 1516–1524.418
- Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradients. Ecological Modelling, 222, 955–963.
- Dunstan, P.K., Foster, S.D., Hui, F.K.C. & Warton, D.I. (2013) Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. Journal of Agricultural, Biological, and Environmental Statistics, 18, 357–375.
- Fithian, W., Elith, J., Hastie, T. and Keith, D.A., 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in ecology and evolution, 6(4), pp.424-438.
- Foster, S.D., Peel, D., Hosack, G.R., Hoskins, A., Mitchell, D.J., Proft, K., Yang, W.H., Uribe-Rivera, D.E. and Froese, J.G., 2024. ‘RISDM ‘: species distribution modelling from multiple data sources in R. Ecography, 2024(6), p.e06964.
- Hill, N., Woolley, S.N., Foster, S., Dunstan, P.K., McKinlay, J., Ovaskainen, O. & Johnson, C. (2020) Determining marine bioregions: A comparison of quantitative approaches. Methods in Ecology and Evolution, 11, 1258–1272. Publisher: Wiley Online Library.
- Hui, F.K., Warton, D.I., Foster, S.D. & Dunstan, P.K. (2013) To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. Ecology, 94, 1913–1919
- Hui, F.K., Menendez, P., Foster, S.D. & Woolley S.N.C (In prep.) Scalable Finite Mixture of Regression Models for Clustering Species Responses in Ecology
- Warton, D. & Shepherd, L. (2010) Poisson point process models solve the ”pseudo-absence problem” for presence-only data in ecology. Annals of Applied Statistics, 4, 1383–1402.