Mixing it up for simple interpretation, using ecomix

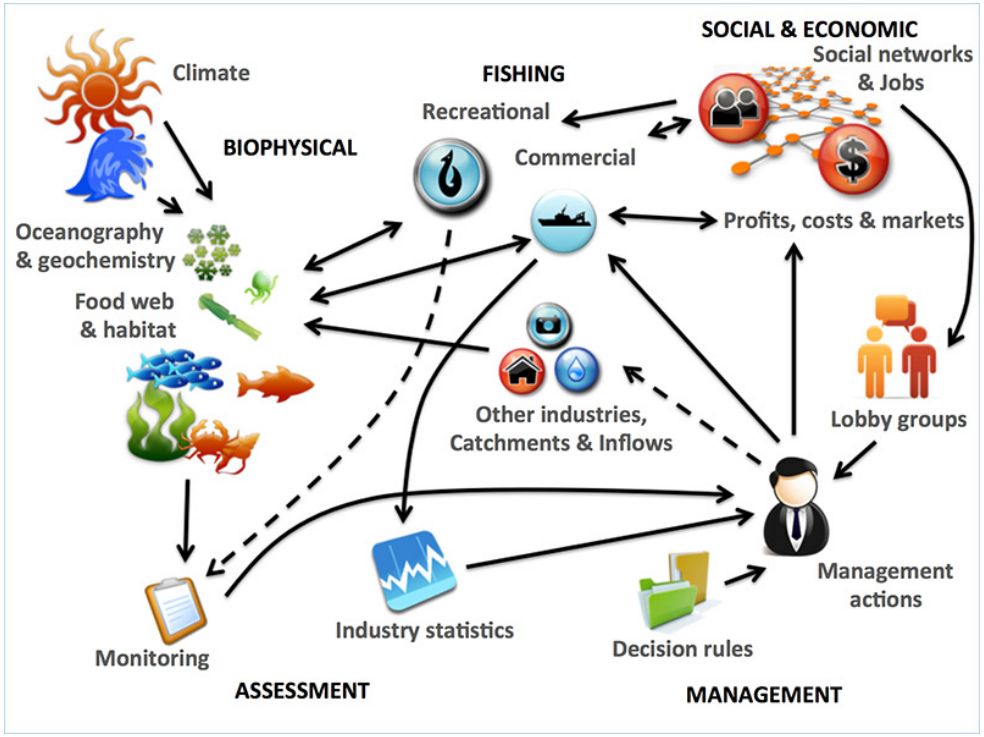
Skipton Woolley, Piers Dunstan, Nicole Hill & Scott Foster

ESA Statistical Methods Series

Ecology is Difficult

- Complex biology
- Complex geography
- Complex pressures
- Complex relationships
- Complex interactions

- It is just *complex*



Message Must be Simple

- Without a simple, clear message
  - Message may not be understood
  - Ecology considerations my not be considered in management decisions
  - Detrimental effects on environment
- But message must also be *real*
- Must be prediction-ready (but predicting what?)
- With statistical interpretation (uncertainty)

What is a Simple Message?

- Leverage off human nature
- Make it easy for humans to understand
  - Humans are naturally predisposed to categorisation
  - Colour, taxonomy, countries,
- Truthful to ecological reality whilst simplified
- Target the management/science question directly

Our Approach for Simple Messages

- Simply stated:
  - Analyse the data with the communication in mind
  - Often categorical (not continuous) delineation

- May not produce *best* fitting model
- Model will be fit-for-purpose though
  - No further analysis-of-analyses to obtain information

Model-based Methods

- We consider model-based methods only
- Provide a description of data (not a derivative of data)
- Formality requires thought about research question
- Repeatable
- Model checking

Research Questions – broad

- For managing the environment inferences about ecosystem are required
- Often questions are about unobserved properties
  - Assemblages
  - Ecoregions / bioregions
  - Functional groups
  - Species groups
  - Communities (gasp!)
  - Stocks
  - Genetic groups
- But none of these are **observed**
- Our solution is to pose statistical models containing these/related constructs
  - formally defines meaning (unlike some 'definitions')

Our Approach

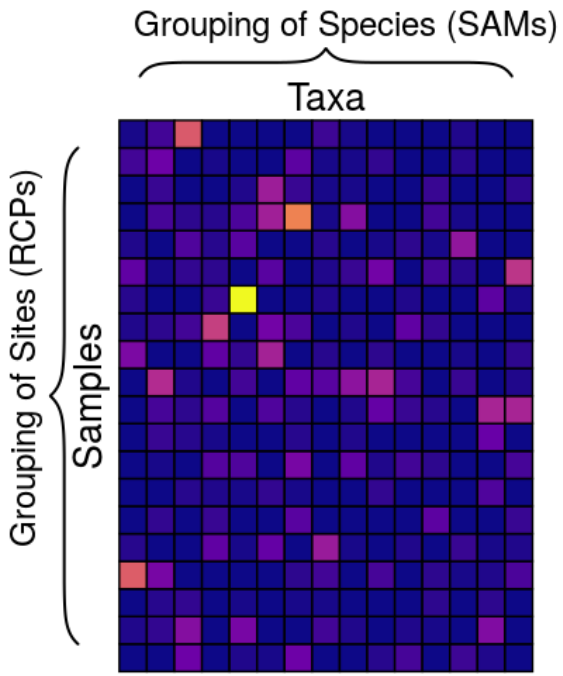
- Consists of two main thrusts:
  - questions about species
  - questions about sites / locations
- Assumes that there are a finite number of latent levels
- Allows for data-specific attributes (e.g. mean-var, sampling artefacts,...)
- Finite number of levels means that results are (more) easily interpreted
  - Grouping allows ingestion by the 'human machine'
- This leads us (of course) to finite mixture models of one kind or another

Data

- Ecological survey data can be rich in information *per observation*
  - abundance of multiple/many species
- Often not so many sites
- Gives a wide matrix of observations
- Often also have covariates at locations of samples
  - Position of sample, depth, temperature, ...

Our Mixture Models

- Multivariate response (conditionally independent)
- Mixture of regressions, aka Species Archetype Models (SAMs)
  - How do species group according to their responses to environmental gradients?
- Mixture of Expert Models aka Regions of Common Profiles (RCPs)
  - How do homogeneous groups of sites vary with environment?



Species Archetype Models (SAMs)

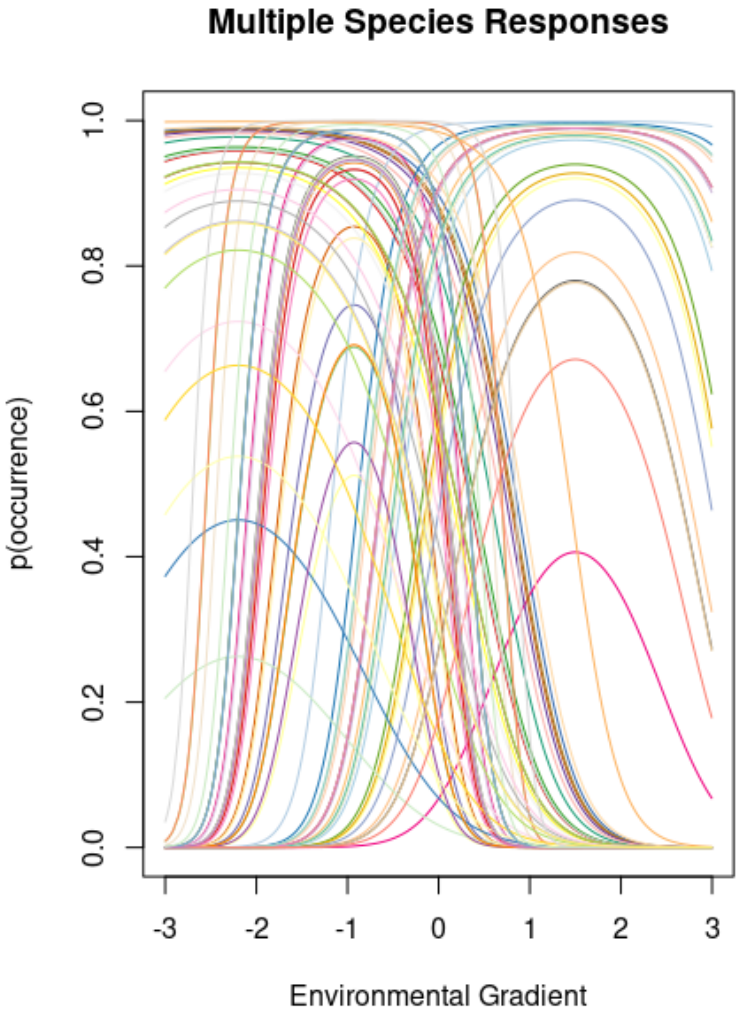
- Grouping *species* according to their responses to the environment
- Soft assignments
  - (to a hard theoretical group)
- Intuitively:
  - Perform a regression on each species, then
  - Cluster the regression coefficients
- Mixture models allow for a one-step process
  - Uncertainty propagation
  - Statistical efficiency

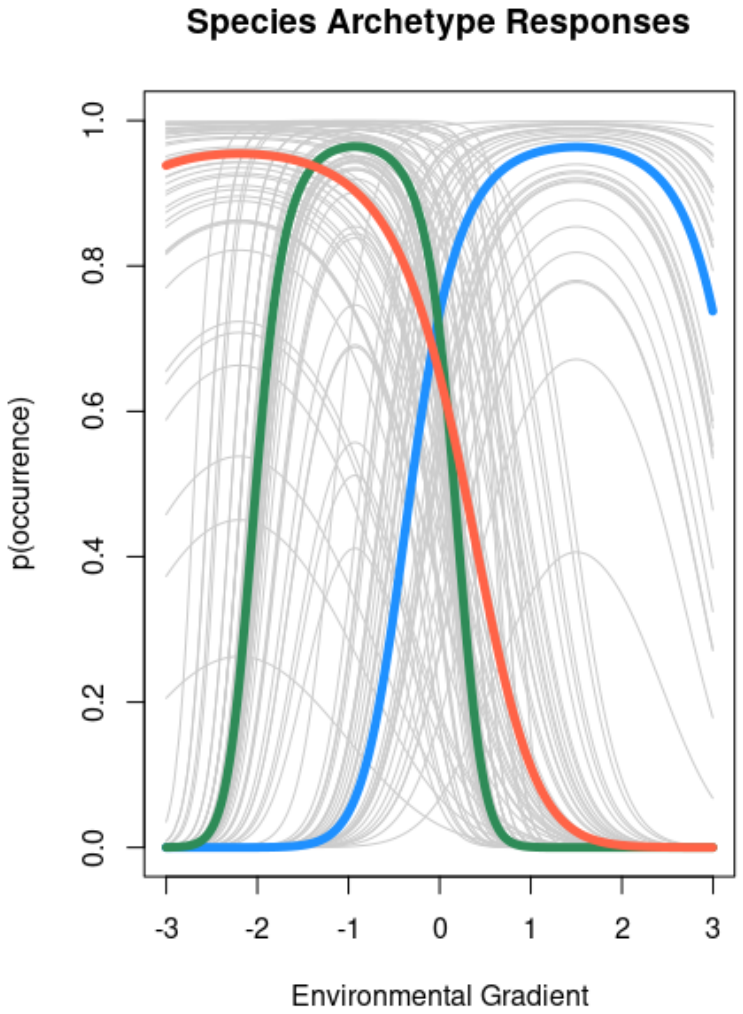
SAMs Modelling Strategy

$$h\left[\mathrm{E}(y_{ij}|z_k)\right] = \alpha_j + g_k(x_i)$$

- $i = 1 \ldots n$ sampling sites
  $j = 1 \ldots S$ species and
  $k = 1 \ldots K$ archetypes
- $g_k(\cdot)$ could be any functional form applicable to a glm, e.g. linear, quadratic or spline.
- $h[\cdot]$ is the appropriate link function

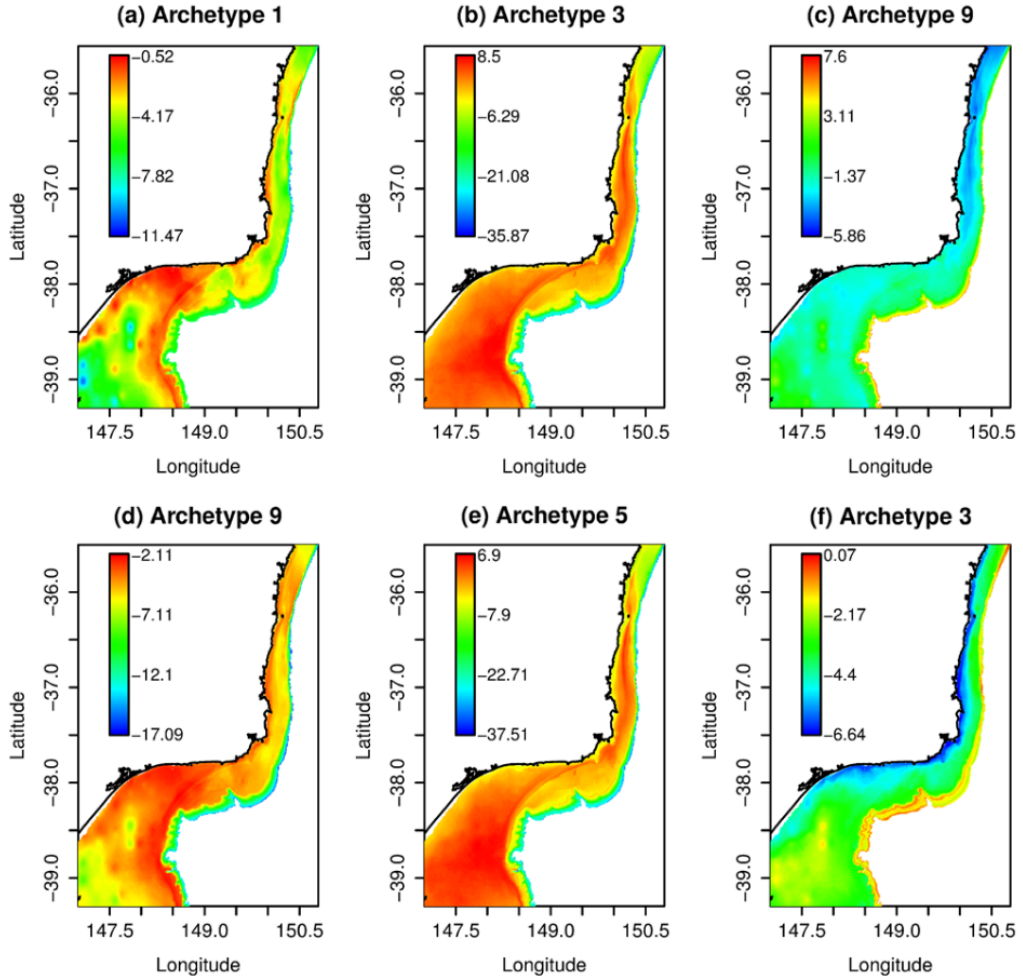
Species Archetype Models (SAMs)





Predicting Species Archetype

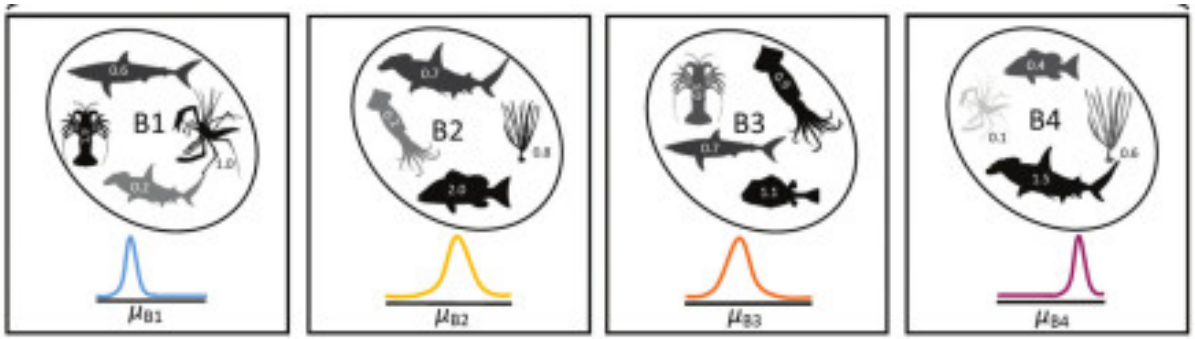
- Spatially predict fish archetypes

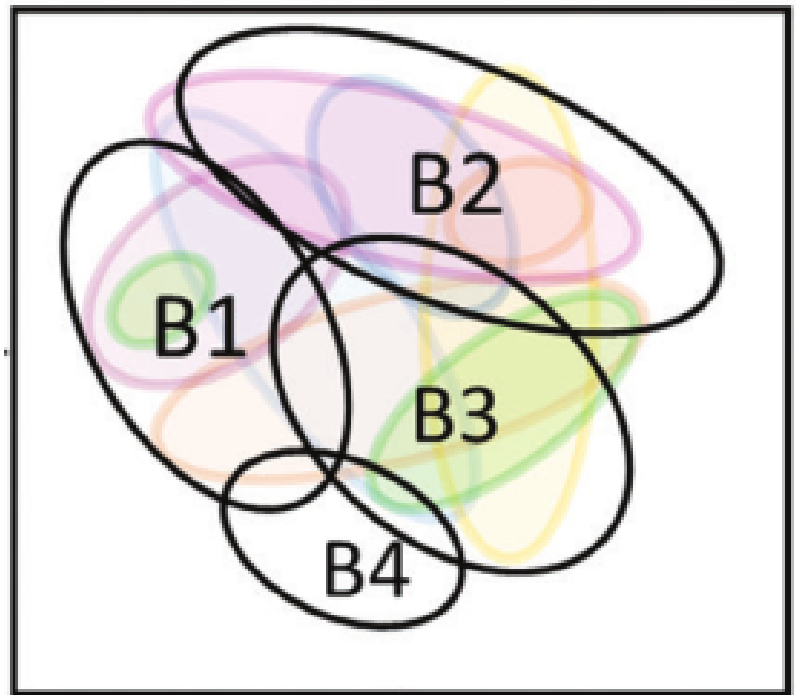


Regions of Common Profiles (RCPs)

- Grouping *sites* according to their species profiles
- Profiles based on expectations/means/prevalences
- Probability of observing that profile then mapped
- Intuitively (but **not** quite right):
  - Perform a clustering of sites
  - Regress clustering on environment
- Mixture models allow for a one-step process
  - Uncertainty propagation
  - Statistical efficiency (important for RCP)

Regions of Common Profiles (RCPs)





RCP Modelling Strategy

- Indices:
  - $i = 1 \ldots n$ (sites)
  - $j = 1 \ldots S$ (species)
  - $k = 1 \ldots K$ (assemblages / RCPs)
- Model *conditional* expectation (given site membership) for all species ($\mathrm{E}\left(y_{ij} | z_{ik} = 1\right)$)
- $h(\mathrm{E}\left(y_{ij} | z_{ik}\right)) = \alpha_j + \mathbf{z}_i^\top \boldsymbol{\tau}_j$

RCP Modelling Strategy

- Indices:
  - $i = 1 \ldots n$ (sites)
  - $j = 1 \ldots S$ (species)
  - $k = 1 \ldots K$ (assemblages / RCPs)
- Model *conditional* expectation (given site membership) for all species $(\mathrm{E}\left(y_{ij}|z_{ik} = 1\right))$
- Adjust the *profile* as the species-wise expectation (offset and survey artefacts)

- $h(\mathrm{E}\left(y_{ij}|z_{ik}\right)) = \alpha_j + \mathsf{z}_i^\top \boldsymbol{\tau}_j$
- $h(\mathrm{E}\left(y_{ij}|z_{ik}\right)) = \alpha_j + \mathsf{z}_i^\top \boldsymbol{\tau}_j + \mathsf{w}_i^\top \boldsymbol{\gamma}_j$

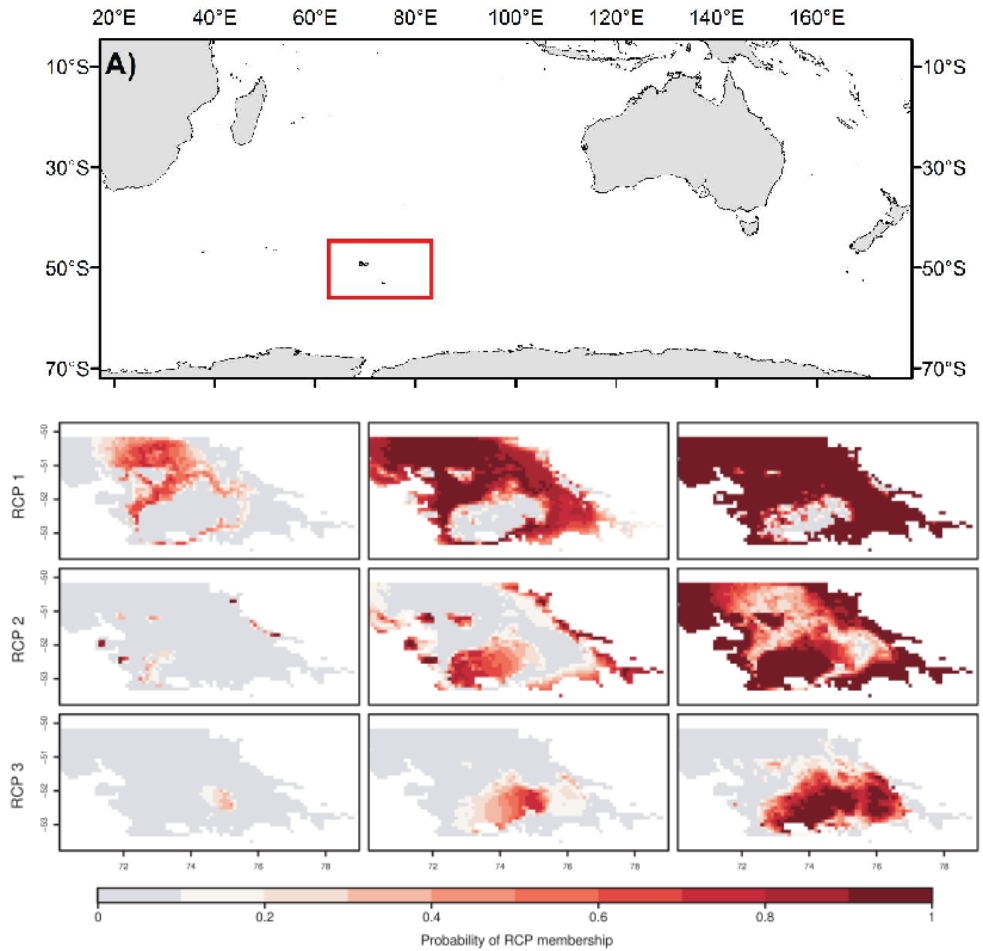
RCP Modelling Strategy

- Allow the probability of observing each RCP (π_i) to vary with environment
- Multinomial regression model (but observations are latent)

- $\pi_{ik} \triangleq$
$$\begin{cases} \frac{\exp(\mathrm{x}_i^\top \beta_k)}{1+\sum_{k'=1}^{K-1} \exp(\mathrm{x}_i^\top \beta_{k'})}, & \text{if} 1 \leq k \leq K \\ 1 - \sum_{k'=1}^{K-1} \pi_{ik'}, & \text{if } k = K \end{cases}$$

Predicting Regions of Common Profile

- Spatial prediction of Kerguelen Fish RCPs.

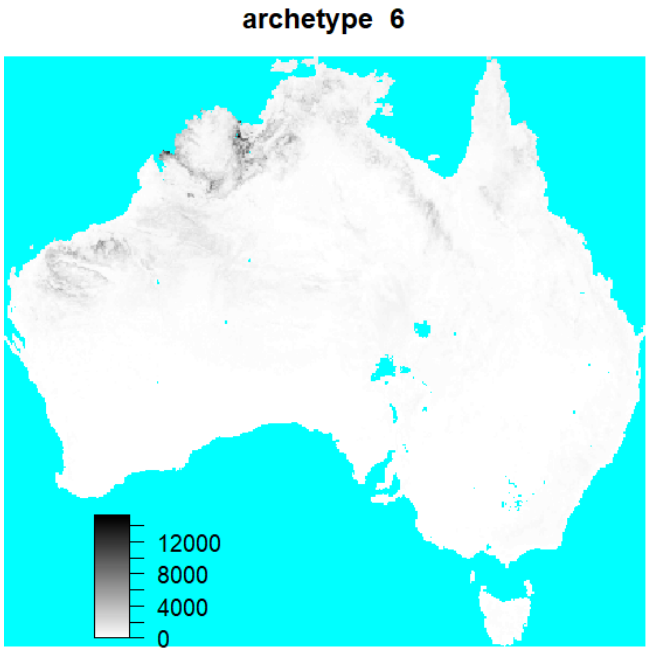


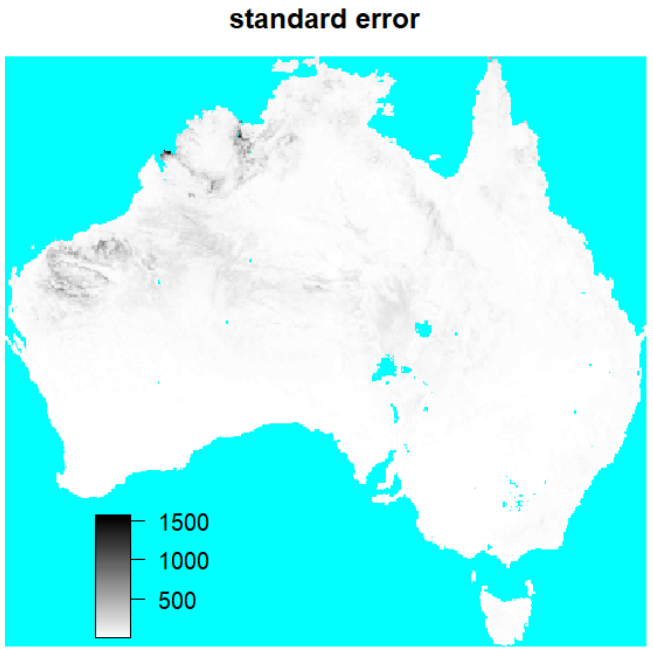
Summary

- Model-based analysis are now a viable and useful alternative to 'traditional' community methods
- Model-based grouping is a viable option over two-step approaches (e.g. predict & group).
- Plenty of attractions: prediction, quantification of uncertainty, diagnostics & formality
- Mixture models are useful for these purposes
  - Allow for appealing discrete interpretation
  - Simple models
- Downsides are in computation and potential simplification of problem (compared to say a JSDM)

Extensions

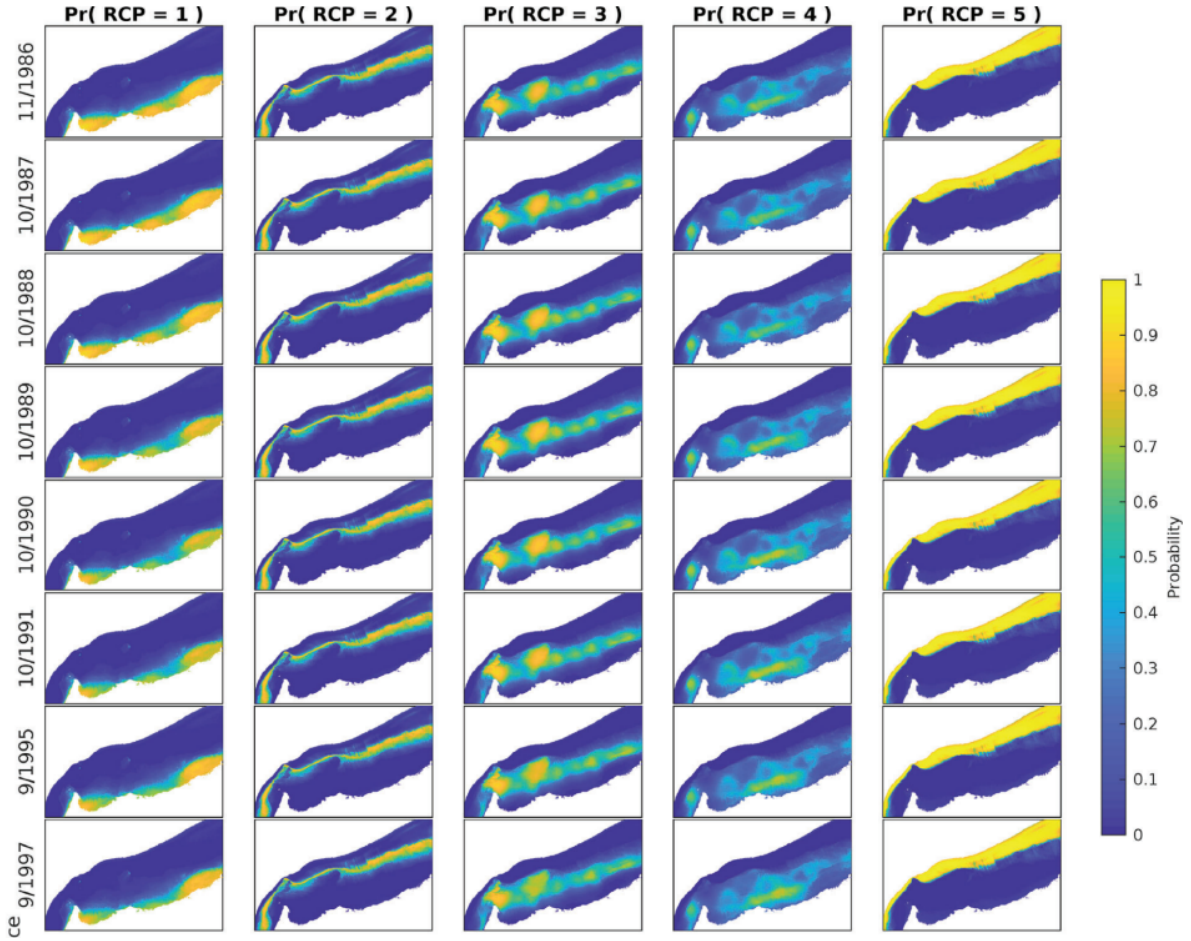
- Poisson Point Process extensions of SAM & RCP





Extensions

- Spatio-temporal extensions of RCP using Gaussian Processes (Vanhatalo et al., 2021).



Useful but Not The Only Model

- Before embarking on any analysis though:
  - Make sure that you define the research question
  - Analyse the data so that the question can be answered

- This may **not** be ecomix.
  - Mixture models are great for exploration and communication
  - Not designed for (e.g.) hypothesis testing (mvabund) or predicting species (hmsc)

Selected References – #1

ter Braak, C.; Hoijtink, H.; Akkermans, W. & Verdonschot, P. Bayesian model-based cluster analysis for predicting macrofaunal communities *Ecological Modelling*, 2003, 160, 235 - 248

Dunn, P. K. & Smyth, G. K. Randomized Quantile Residuals *Journal of Computational and Graphical Statistics*, 1996, 5, 236-244

Dunstan, P.K., Foster, S.D. and 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 and Warton, D.I. (2013) Finite Mixture of Regression Modelling for high-dimensional count and biomass data in Ecology. *Journal of Agricultural, Biological and Environmental Statistics* 18, 357-375.

Foster, S.D., Dunstan, P.K., Althaus, F., and Williams, A. (2015) The cumulative effect of fishing on a multi-species assemblage in South Eastern Australia. *Journal of Applied Ecology* 52, 129-139.

Foster, S.D., Givens, G.H., Dornan, G.J., Dunstan, P.K., and Darnell, R. (2013) Regions of common species profiles: grouping sites by biological and environmental data. *Environmetrics* 24, 489-499.

Foster, S.D., Hill, N.A. and Lyons, M. (2017) Ecological Grouping of Survey Sites when Sampling Artifacts are Present. *The Journal of the Royal Statistical Society – Series C* To appear.

Selected References – #2

Hill, N.A., Foster, S.D, Duhamel, G., Welsford, D., Koubbi, P., and Johnson, C. (*Accepted*) Model-based mapping of assemblages for ecology and conservation management: A case study of demersal fish on the Kerguelen Plateau. *Diversity and Distributions* To appear.

Hill, N., Foster, S.D., Woolley, S.N.C., Dunstan, P.K., McKinlay, J., Ovaskainen, O., and Johnson, C. Determining marine bioregions: A comparison of quantitative approaches *in review*

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. and Warton, D.I. (2015) Model-Based Approaches to Unconstrained Ordination. *Methods in Ecology and Evolution* 6, 399-411.

Hui, F.K.C., Warton, D.I. and Foster, S.D. (2015–a) Multi-Species Distribution Modeling Using Penalized Mixture of Regressions. *Annals of Applied Statistics* 2, 866-882.

Hui, F.K.C., Warton, D.I. and Foster, S.D. (2015–b) Order selection in finite mixture models: complete or observed likelihood information criteria?. *Biometrika* 102, 724-730.

Hui, F.K.C., Warton, D.I., Foster, S.D., and Dunstan, P.K. (2013) To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology* 94, 1913-1919.

Johnson, D. S. & Sinclair, E. H. Modeling joint abundance of multiple species using Dirichlet process mixtures *Environmetrics*, 2017, e2440-n/a

Selected References – #3

Jupke, J. F. & Schafer, R. B. Should ecologists prefer model- over distance-based multivariate methods? *Ecology and Evolution*, 2020

Leaper, R.; Dunstan, P.; Foster, S.; Barrett N.S. & Edgar, G. Do communities exist? Complex patterns of overlapping marine species distributions *Ecology*, 2014, 95, 2016-2025

Lyons, M., Foster, S.D., Keith, D. (*Accepted*) Simultaneous vegetation classification and mapping at large spatial scales. *Journal of Biogeography*

Norberg, A.; Abrego, N.; Blanchet, F. G.; Adler, F. R.; Anderson, B. J.; et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels *Ecological Monographs*, 2019, 89, e01370

Ovaskainen, O. & Soininen, J. Making more out of sparse data: hierarchical modeling of species communities *Ecology*, 2011, 92, 289-295 Pledger, S. & Arnold, R. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection *Computational Statistics & Data Analysis*, 2014, 71, 241 - 261

Valle, D.; Albuquerque, P.; Zhao, Q.; Barberan, A. & Fletcher, R. J. Extending the Latent Dirichlet Allocation model to presence/absence data: a case study on North American breeding birds and biogeographic shifts expected from climate change *Global Change Biology*, 2018

Valle, D.; Baiser, B.; Woodall, C. W. & Chazdon, R. Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method *Ecology Letters*, 2014, 17, 1591-1601

Selected References – #3

Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. and Dunstan, P.K. (2015) Model-Based Thinking for Community Ecology. *Plant Ecology* 216, 669-682
Woolley, S. N. C.; Foster, S. D.; Bax, N. J.; Currie, J. C.; Dunn, D. C.; Hansen, C.; Hill, N.; O'Hara, T. D.; Ovaskainen, O.; Sayre, R.; Vanhatalo, J. P. and Dunstan, P. K. Bioregions in Marine Environments: Combining Biological and Environmental Data for Management and Scientific Understanding *BioScience*, 2019, 70, 48-59
Vanhatalo, J., Foster, S.D. and Hosack, G.R., (2021). Spatiotemporal clustering using Gaussian processes embedded in a mixture model. *Environmetrics*.