Mixture Models for Multi-Species and Environmental Data

[1,2]Scott Foster, [1,3]Piers Dunstan, [4]David Warton, [2,4]Francis Hui, [2]Ross Darnell, [5]Geof Givens, [6]Grant Dornan

[1] CSIRO Wealth from Oceans Flagship, [2]CSIRO Division of Computational Informatics, [3]CSIRO Division of Marine and Atmospheric Science,
[4]University of New South Wales, School of Mathematics and Statistics, [5]Colorado State University, Department of Statistics, [6]Steadman Philippon Research Institute

- Ecological inference and management decisions often depend on data from many species.
- A proper and useful statistical analysis quantifies the important patterns of variation, whilst reducing the complexity in multi-species data.
- Currently, analysis is frequently done by: 1) performing species-by-species analyses (e.g. univariate regression and extensions) and then combining results, or 2) by combining data (clustering) and then performing a group-by-group analysis.
- Neither of the standard approaches are entirely satisfactory as important aspects of the variance in the data can be lost when moving from step to step. Also, the propagation of uncertainty is difficult and is subsequently (often) ignored.
- We introduce two models, based on mixture models, that address these issues. One model type, *species archetype models (SAMs)* exploits similarities in individual species' responses to the environment. The second type, *regions of common profile (RCP) models,* exploits similarities in the assemblage patterns at each site.

Models

We use 2 variants of mixture models to represent variation in data from many species. Let the species' data be given by $\{y_{ij}\}$ and the environmental data be $\{\boldsymbol{w}_i, \boldsymbol{x}_i\}$, where $i = 1 \ldots N$ index sites and $j = 1 \ldots S$ index species. We have used these methods for $S \approx 300$ species and $N \approx 1200$ - limited by size of survey data at the moment.

Species Archetype Models (SAMs, for when inference on species is required)

Mixture of regressions model to group individual species' responses to environmental gradients into *archetypical* responses. Only $K << S$ archetypical responses are interpreted, instead of S models from a species-by-species analysis. The model for the expectation is:

$$\mathrm{E}\left[y_{ij}\right] = \sum_{k=1}^{K} \pi_k \mathrm{E}\left[y_{ij} | \text{archetype } k\right], \qquad \text{where}$$
$$g\left(\mathrm{E}\left[y_{ij} | \text{archetype } k\right]\right) = \alpha_i + \boldsymbol{w}_i^\top \boldsymbol{\tau}_j + \boldsymbol{x}_i^\top \boldsymbol{\beta}_k$$

and the distribution of each species' data may have species-specic parameters [1, 2, 3, 4, 5].

See Figure 1 for some archetypical environmental responses of fish off the south-eastern coast of mainland Australia. Analysis performed using physical environmental covariates, biomass data and a Tweedie model.

Regions of Common Proles (RCP, for when inference on sites is required)

Mixture of experts model to group sites based on its species prole in relation to environment. This allows interpretation of only $H << N$ RCPs to study assemblage-environment relationships. No species, nor site, is hard-clustered into a particular RCP. The model for expectation is:

$$\mathrm{E}\left[y_{ij}\right] = \sum_{h=1}^{H} \pi_h(\boldsymbol{w}_i, \boldsymbol{x}_i) \mathrm{E}\left[y_{\cdot j} | \text{RCP } h\right],$$

where $\pi_h(\boldsymbol{w}_i, \boldsymbol{x}_i)$ is a multinomial logit link function [6].

See Figure 2 for the predicted probability of each region-type, based on fish assemblages for the North West Shelf of Australia. Analysis performed using physical environmental covariates and fish presence/absence data.

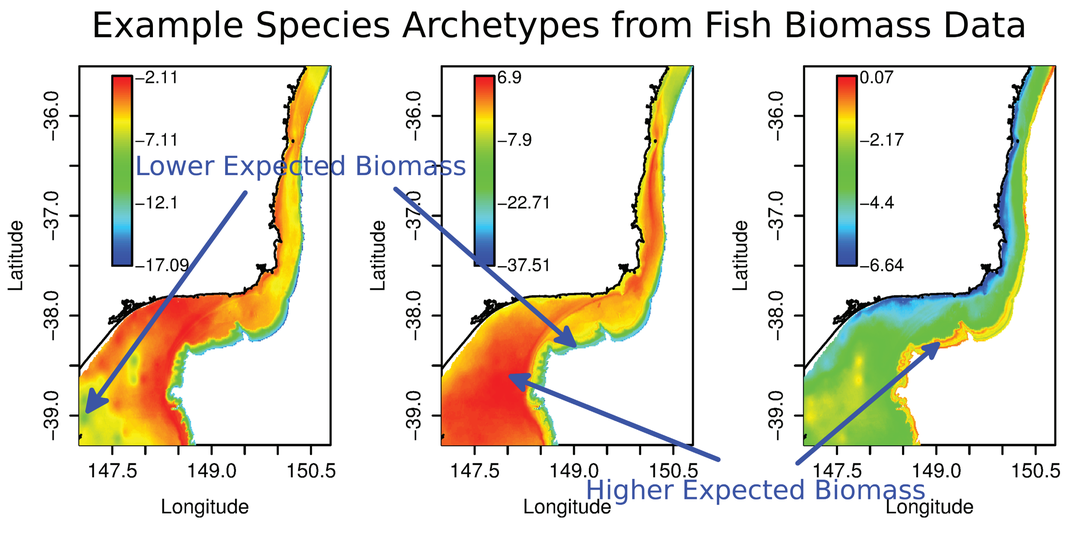


Figure 1: Example results from a SAM analysis. Maps are for expected biomass, on the log scale.

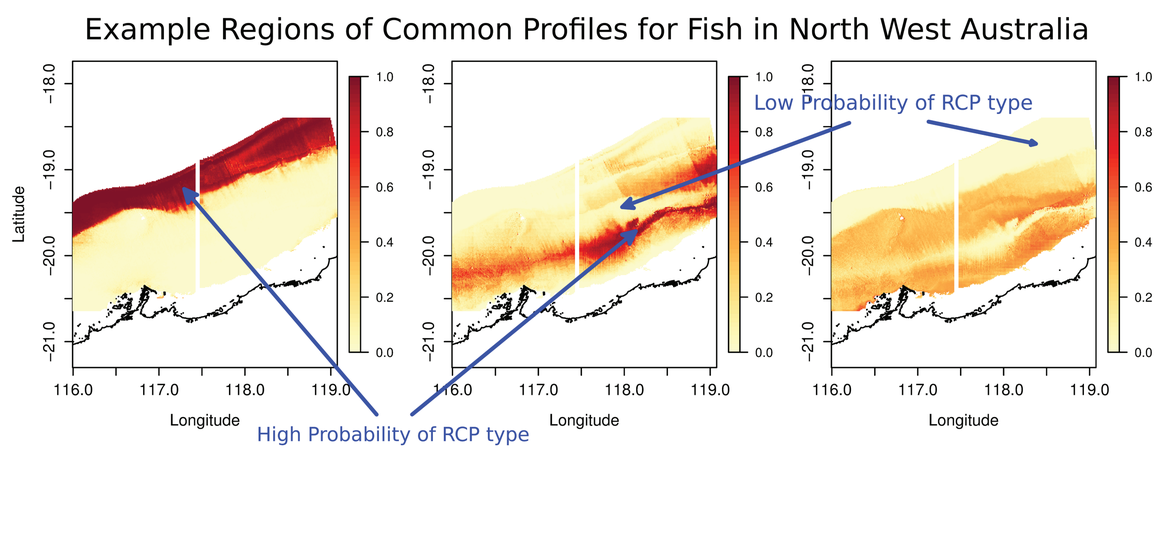
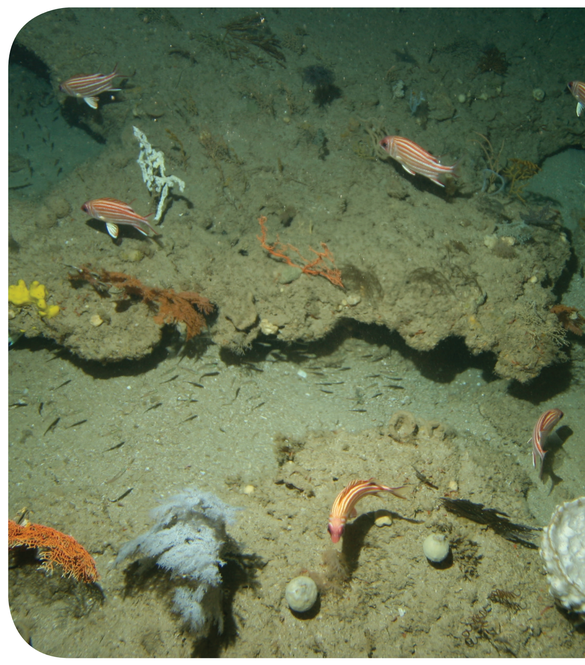


Figure 2: Example results from an RCP analysis. Maps are for the probability of each region type occurring at each location.

Benets of Using SAMs and RCPs

Interpretability - complexity of multi-species data reduced by clustering species (SAMs) or sites (RCPs) [1, 6, 4].
Prediction Performance - rarer species `borrow strength' from common species [3].
Diagnostics - checking model adequacy, e.g. residual plots [2, 5].
Model Selection - using variants of common methods [7, 8].
Flexibility - many choices to match model to data [2, 5].
Transparency - model is formally specied.
Efficiency - proper statistical inference [1, 7].



References

[1] P.K. Dunstan, S.D. Foster, and R. Darnell. Model based grouping of species across environmental gradients. Ecological Modelling, 222:955 - 963, 2011.
[2] P.K. Dunstan, S.D. Foster, F.K.C. Hui, and D.I. Warton. Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. Journal of Agricultural, Biological, and Environmental Statistics, 18:357 - 375, 2013.
[3] F.K.C. Hui, D.I. Warton, S.D. Foster, and P.K. Dunstan. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. Ecology, 94:1913 - 1919, 2013.
[4] R. Leaper, P.K. Dunstan, S.D. Foster, N.S. Barrett, and G.J. Edgar. Do communities exist? complex patterns of overlapping marine species distributions. Ecology, in press.
[5] S.D. Foster, P.K. Dunstan, F. Althaus, and A. Williams. The cumulative effect of trawling on a multi-species fish assemblage off south eastern Australia. In review.
[6] S.D. Foster, G.H. Givens, G.J. Dornan, P.K. Dunstan, and R. Darnell. Modelling biological regions from multi-species and environmental data. Environmetrics, 24(7):489 - 499, 2013.
[7] F.K.C. Hui, D.I. Warton, and S.D. Foster. An akaike information criterion for mixture models. In review.
[8] F.K.C. Hui, D.I. Warton, and S.D. Foster. The group lasso for finite mixture of regression models, with applications to species distribution modelling. In review.

Australian Government
Department of the Environment
Australian Government
Geoscience Australia
Australian Government
Australian Institute of Marine Science
CSIRO
Charles Darwin University
MUSEUMVICTORIA
UTAS
The University of Western Australia
Achieving International Excellence

The NERP Marine Biodiversity Hub is supported through funding from the Australian Government's National Environmental Research Program, administered by the Department of Sustainability, Environment, Water, Population and Communities (DSEWPaC). Our goal is to support marine stakeholders in evidence-based decision making for marine biodiversity management. Stakeholders include DSEWPaC, the Australian Fisheries Management Authority (AFMA), the Australian Petroleum Production and Exploration Association (APPEA) and the Integrated Marine Observing System (IMOS).

National Environmental Research Program
MARINE BIODIVERSITY hub

Further information:
Scott Foster
p +61 3 6232 5178 **e** scott.foster@csiro.au