MARINE BIODIVERSITY RESEARCH
Prediction and Management of Australia's Marine Biodiversity
Australian Government
豪日交流基金
Australia-Japan FOUNDATION

Australia – Japan Workshop on Data Science

Keio University, 24 – 27 March 2009

Latent Class Models

Ross DARNELL Louise MARQUART

CSIRO Mathematical and Information Sciences

CSIRO

Outline

1. Background
2. Data set
3. Model
4. Example
5. Discussion

Australian Government
豪日交流基金 Australia-Japan FOUNDATION
MARINE BIODIVERSITY RESEARCH Prediction and Management of Australia's Marine Biodiversity
CSIRO
Australian Government Geoscience Australia

Backgrounds

CERF project:

The Commonwealth Environment Research Facilities (CERF) Marine Biodiversity Hub prediction project analyses patterns and dynamic of marine biodiversity to determine the appropriate units and models for effectively predicting Australia's marine biodiversity.

The project administered through the Australian Government Department of the Environment, Water, Heritage and the Arts

Major contributers are: University of Tasmania; CSIRO Wealth from Oceans Flagship; Geoscience Australia; Australian Institute of Marine Science; Museum Victoria.

Aims:

To construct predicting models of biodiversity (eg presence/absence, count, weight of each speceis/units) which

- show relationships with physical variables (eg);
- provide reasonable explanation to understand biology.

This presentation looks at clustering species according to their relationship with their physical environment using latent class models.

Data Set

Binary outcome:

$$Y_{ij} = \begin{matrix} 1 & \text{species } i \text{ present at site } j, \\ 0 & \text{species } i \textbf{ not} \text{ present at site } j \end{matrix}$$

and construct

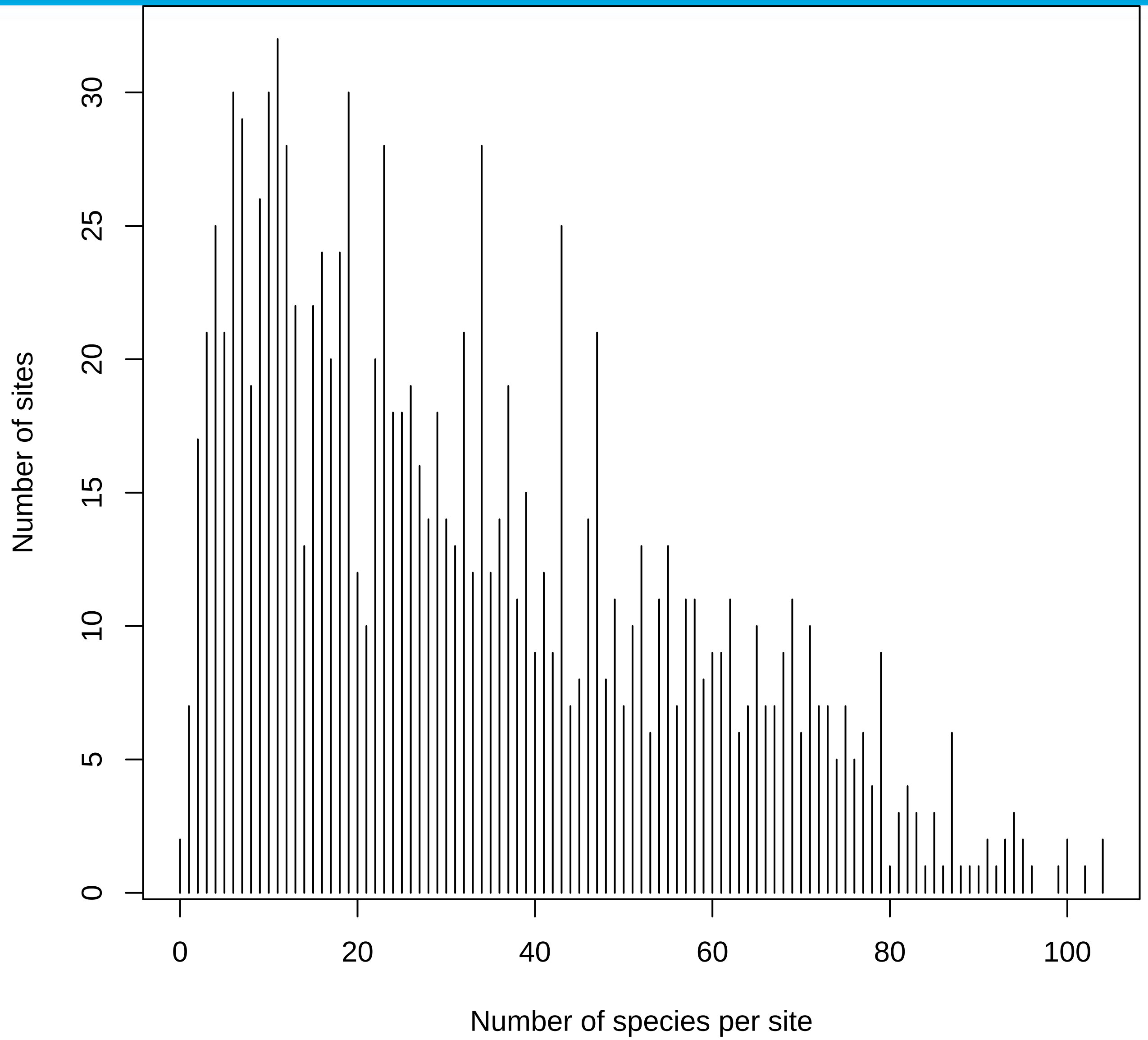
$$S_j = \sum_i Y_{ij} \quad \text{number of species observed at site } j$$

$$j = 1, \ldots, 1189$$

$$i = 1, \ldots, 278$$

One species was observed at 838 sites. 55 species were observed at 6 sites.

Frequency of S



Australian Government
豪日交流基金 Australia-Japan FOUNDATION
MARINE BIODIVERSITY RESEARCH
Prediction and Management of Australia's Marine Biodiversity
CSIRO
Australian Government
Geoscience Australia

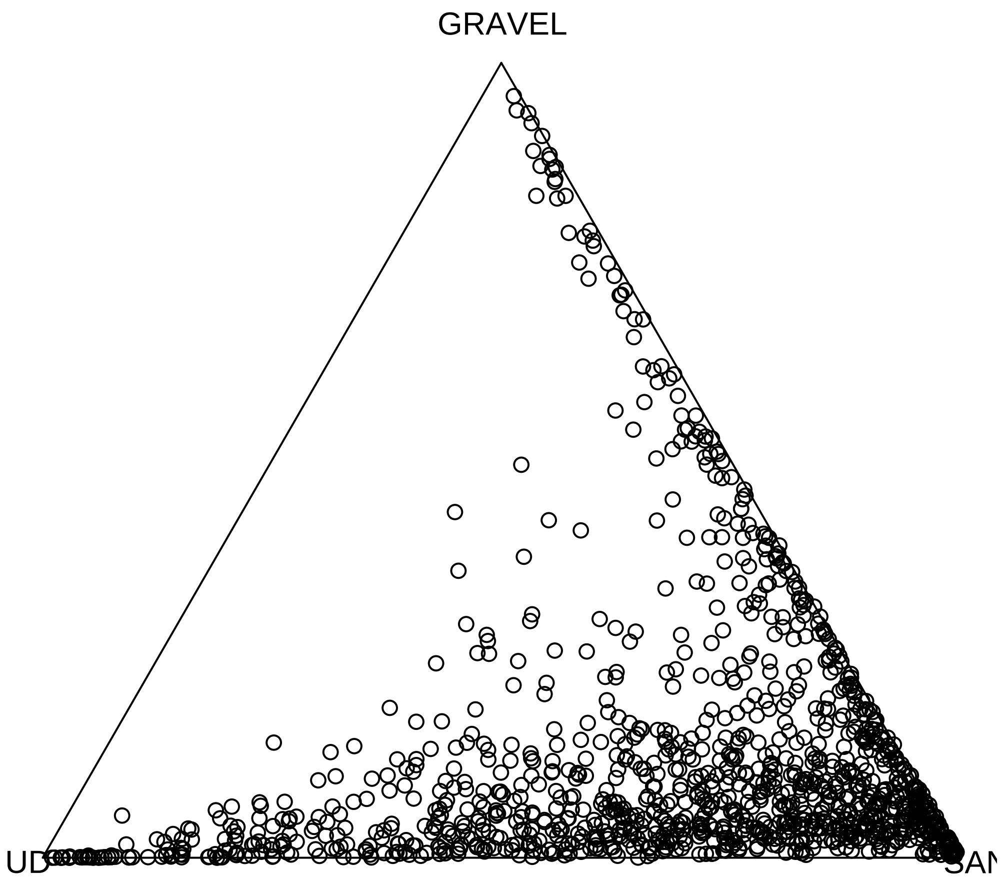
Physical predictors:

32 variables (X_l) measured at each site (j, X_{jl}):

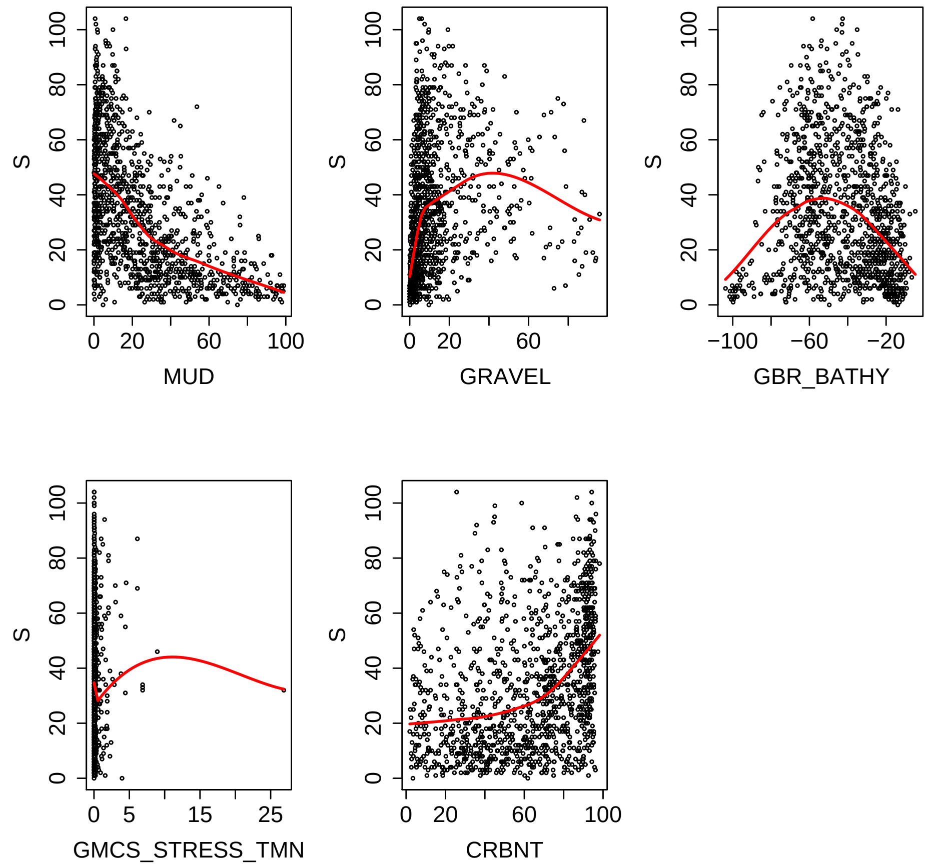
We would like to discuss some of these.

Label	Description
GBR_BATHY	Depth of water
GBR_TS_BSTRESS	Benthic stress
CRBNT	Carbonate
S_AV	Salinity
GRAVEL	Percent gravel
MUD	Percent mud
SAND	Percent sand (dropped)

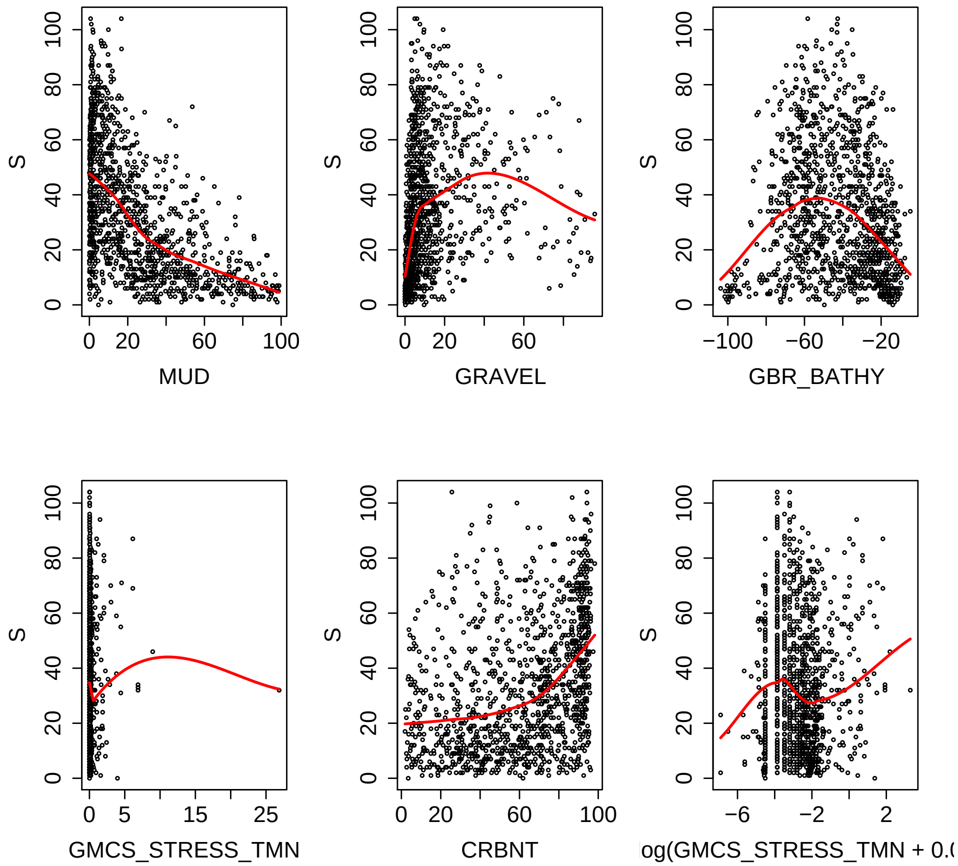
Mud , Sand and Gravel percentages for GBR Sled Sites



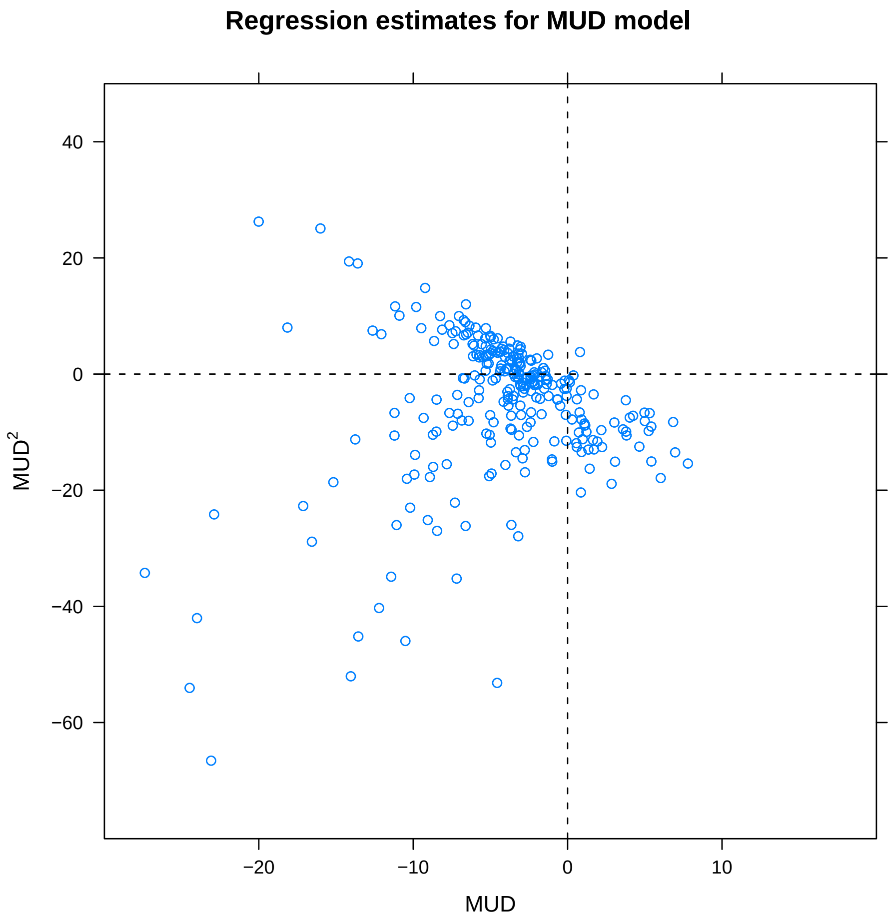
Australian Government
豪日交流基金 Australia-Japan FOUNDATION
MARINE BIODIVERSITY RESEARCH
Prediction and Management of Australia's Marine Biodiversity
CSIRO
Australian Government
Geoscience Australia



Australian Government
豪日交流基金 Australia-Japan FOUNDATION
MARINE BIODIVERSITY RESEARCH
Prediction and Management of Australia's Marine Biodiversity
CSIRO
Australian Government
Geoscience Australia


S
MUD
GRAVEL
GBR_BATHY
GMCS_STRESS_TMN
CRBNT
og(GMCS_STRESS_TMN + 0.(

Example — Mud


Regression estimates for MUD model
MUD
MUD^2
40
20
0
−20
−40
−60
−20
−10
0
10

Model

Latent class model.

$$Y_{ij} \sim \texttt{binomial}(1, g^{-1}(\eta_{ij}))$$

where g represents the logit link and

$$\eta_{ij} = \beta_{1i} \times \texttt{MUD}_j + \beta_{2i} \times \texttt{MUD}_j^2 + Z_i + U_i \times \texttt{MUD}_j + V_i \times \texttt{MUD}_j^2$$

Here $\beta_{1i} = \beta_1 + U_i$, where U_i represents variation about a "mean" β_1 and similarly Z_i for β_0 and V_i for β_2. Marginally Z_i, U_i and V_i have an unknown joint distribution $g(z, u, v)$. The likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_i \left\{ \int \left[\prod_j f(y_{ij} \mid z_i, u_i, v_i) \right] g(z_i, u_i, v_i) \mathrm{d}z_i \mathrm{d}u_i \mathrm{d}v_i \right\}.$$

Estimation

- LATENT GOLD package — uses a hybrid of the EM and Newton Raphson algorithm.

- Uses multiple starting points to avoid local minima

- Tend to avoid using asymptotic p value of $-2\times$log-likelihood (ℓ) difference. Bootstrap samples based on model probability distribution and define p_{boot} as the proportion of bootstrap samples with a larger -2ℓ difference than original sample. Unfortunately this is very slow for datasets of this size.

Latent Gold developed by Jeroen K Vermunt and Jay Magidson

Criteria.

LATENT GOLD reports other statistics...

$$
\begin{aligned}
BIC &= -2\log\mathcal{L} + \log N npar, \\
AIC &= -2\log\mathcal{L} + 2npar, \\
AIC3 &= -2\log\mathcal{L} + 3npar \\
CAIC &= -2\log\mathcal{L} + [\log(N) + 1]npar,
\end{aligned}
$$

Posterior class membership probabilities

For response class k, these are

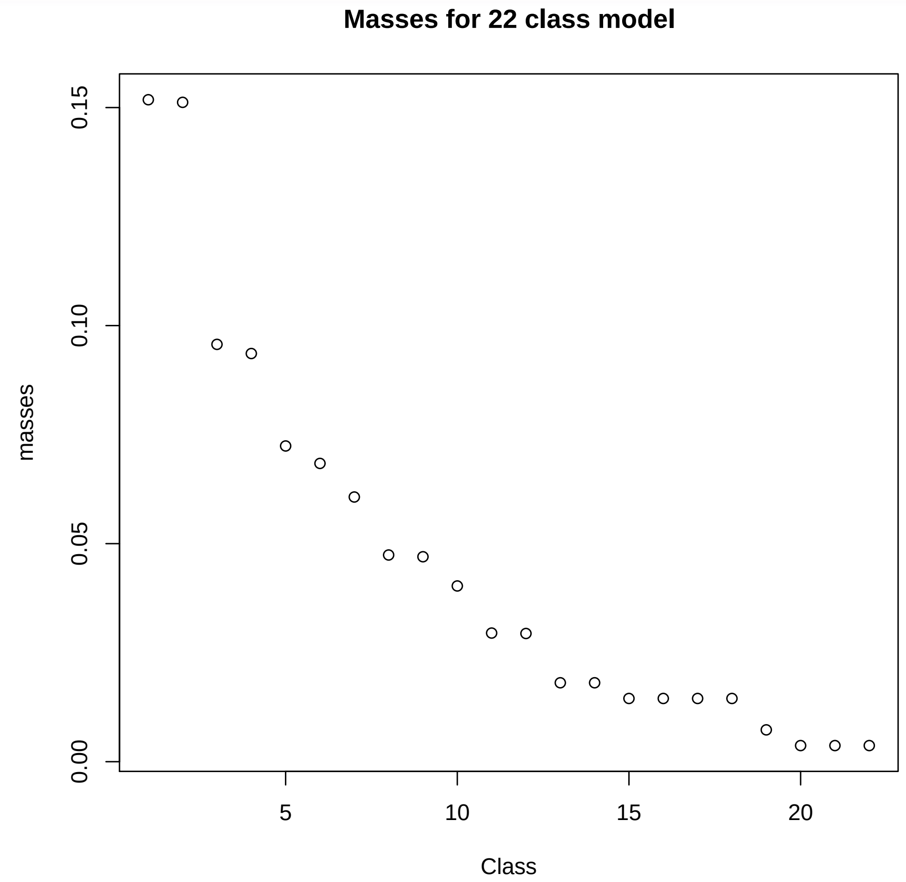
$$P(\text{Species } i \text{ belonging to class } k) = \frac{\pi_k \prod_j f_{ijk}}{\sum_l \pi_l \prod_l f_{ijl}}$$

in our case f_{ijk} is the binomial density for species i at site j for class k.

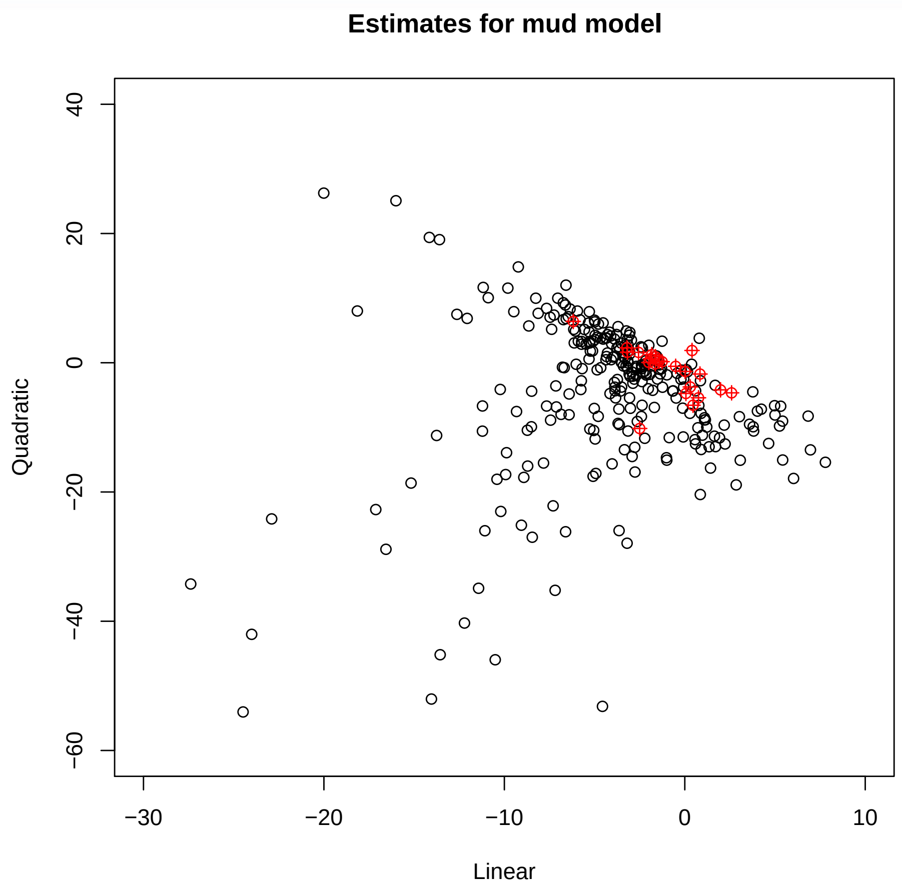
Example — Mud criteria

K	AIC
1	232900
2	216554
3	212285
⋮	⋮
15	204401
16	204187
17	204122
18	204028
19	203924
20	203870
21	203766
22	203668
23	203742
24	203643
25	203534

Example — Masses



Example — Estimates



Example — Class Membership

species	Modal	Class1	Class2	Class3	Class4
CBMTQ.TVL192602	1	0.8817	0.0243	0.0053	0
CBMTQ.TVL192643	1	0.9949	0.0051	0	0
CBMTQ.TVL192670	1	0.9996	0.0003	0	0.0001
CBMTQ.TVL192683	1	0.9982	0.0013	0	0.0004
CBMTQ.TVL192830	1	1	0	0	0
CBMTQ.TVL192833	1	0.9851	0	0	0.0001
CBMTQ.TVL192848	1	0.9997	0	0.0001	0
CBMTQ.TVL192987	1	0.5548	0	0	0.4451
CBMTQ.TVL193079	1	0.9998	0	0	0.0002
⋮					
CBMTQ.TVL193617	2	0	0.9996	0	0
CBMTQ.TVL194248	2	0	1	0	0
⋮					
CBMTQ.TVL192932	20	...	1	0	0
SCQMSB.BRS194716	20	...	1	0	0
MSAIMT192631	21	...	0	1	0
MSAIMT193417	22	...	0	0	1

Number of species in each class

Class	# species	Class	# species
1	43	12	10
2	42	13	10
3	26	14	6
4	19	15	6
5	16	16	5
6	17	17	5
7	15	18	5
8	13	19	4
9	11	20	2
10	11	21	1
11	10	22	1

Discussion

- "Unsupervised learning" approach to clustering species according to their relationship with physical environment.
- Asymptotic results questionable
- Bootstrap methods requires large amounts of computing resources.
- Have I answered the marine biologist's question ? Maybe not.
- Has the approach been useful? (data support, computer power)

Thank you for your kind attentions.
Comments and suggestions are welcomed!

Ross DARNELL
ross.darnell@csiro.au

Louise Marquart