


Australia – Japan Workshop on Data Science

Keio University, 24 – 27 March 2009

Investigating the Issues of Sampling in Marine Surveys



Hideyasu SHIMADZU^{1 2}

¹ Geoscience Australia

Ross DARNELL²

² CSIRO MIS



Australian Government

Geoscience Australia



CSIRO

1. Backgrounds
2. Data sets
3. System behind the data
4. Number of species
5. Presence/absence of taxonomic groups

Backgrounds

CERF project: (<http://www.marinehub.org/>)

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 contributors: University of Tasmania (Utas); CSIRO Wealth from Oceans Flagship; Geoscience Australia (GA); Australian Institute of Marine Science (AIMS); Museum Victoria (MV).

Members involved: Ross Darnell, Scott Foster, Hideyasu Shimadzu

Aims:

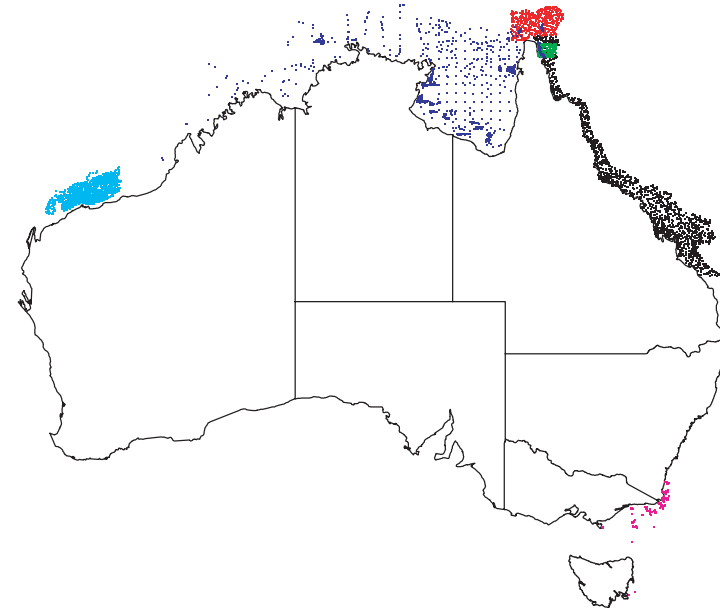
To construct predicting models of biodiversity (eg presence/absence, count, weight of each species/units, number of species etc) which

- show relationships with physical variables;
- provide reasonable explanation to understand.

Key outcome: (Species) distribution maps around Australia.

Data sets (biological data)

- Great Barrier Reef (GBR)
- Torres Strait (TS)
- Effect of Trawling (EoT)
- Gulf of Carpentaria (GoC)
- North West Shelf (NWS)
- South East Fishery (SEF)



	GBR	TS	EoT	GoC	NWS	SEF
Period	2003–06	2004–05	1992–95	1980–98	1982–97	1993–95
Sites	1252	197	383	1751	1544	277
Methods	PT, S	PT, S	FT, PT, S	FT, PT, S	FT	FT, S
Gear types	2	2	3	7	2	2
No species	2862	3639	1689	1999	805	434

FT: fish trawl; PT: prawn trawl; S: sled.

Physical data are also available for each site.

Original biological data

S_ID	GEAR	OTU_ID	NOS	WT	RATIO	START_TIME	AREA
60	S	MSAIMT193417	5	0.256	1	2003/10/4 14:06:56	0.032
60	S	SCQMSB-BRS192482	NA	0.031	1	2003/10/4 14:06:56	0.032
60	S	SCQMSB-BRS192735	NA	0.029	1	2003/10/4 14:06:56	0.032
60	S	CBMTQ-TVL192951	1	0.0012	0.33	2003/10/4 14:06:56	0.032
60	S	MSAIMT193417	14	0.042	0.33	2003/10/4 14:06:56	0.032
:							

S_ID: site id;

GEAR: survey gear type used;

NOS: number of count individual;

WT: observed weight (Kg);

RATIO: subsample ratio := (sample weight)/(total catch weight);

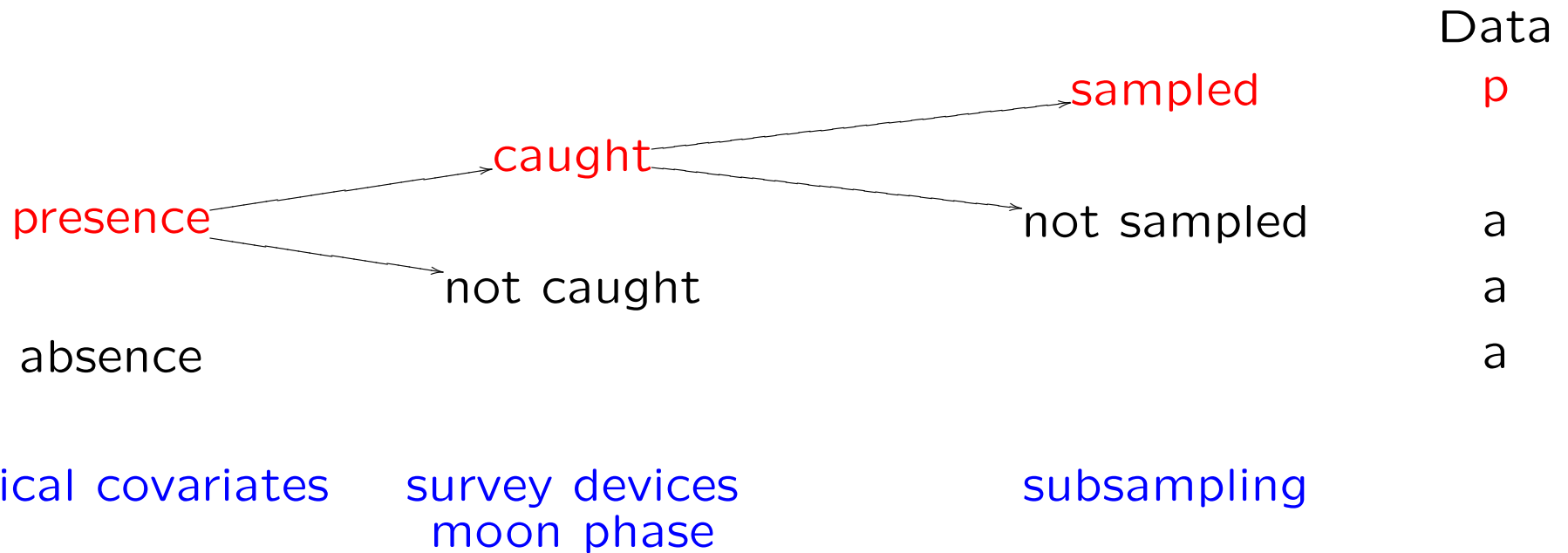
START_TIME: day & time survey conducted;

AREA: swept area (Ha).

- **Lack of absence records** (Zaniewski *et al.* 2002; Ward *et al.* 2008);
- Many statistical models have been proposed to cope with many zero observations (eg NB, ZIP, Hurdle, Tweedie model etc);
- It is also of importance to **investigate the reason** why zero counts are occurred **from data collection perspectives**;
- Observation is largely influenced by the sampling process.

System behind the data

Consider the simplest case of presence/absence of species:



The observed *absence* may include at least 3 reasons.

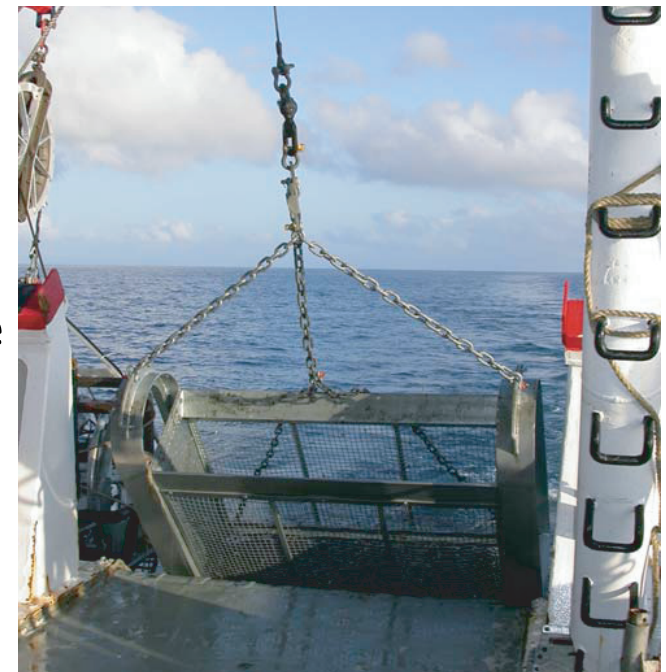
Sampling process

Pitcher *et al.* (2007)

1. Remove large animals;
2. Sort the rest into phylogenetic group (門) (eg sponges, crustaceans, algae, ascidians, seagrasses, fishes etc.);
3. Fully sort or subsample.

Subsampling is a widely used method to reduce the volume of catch.

$$\text{subsample ratio} = \frac{\text{sample weight}}{\text{total catch weight}}$$



Subsampling issues

Subsampling may easily reduce the estimates of

- number of species caught;
- probability of presence of each species;
- abundance of species, etc.

Note: No one knows how much biodiversity indices are influenced. It seems to have been few researches, see Heales *et al.* (2003) for example.



Types of sampling (Sled surveys)

10

1. fully sampled
2. Only a particular species subsampled
3. *Animalia* (動物界) full sampled; *Plantae* (植物界) subsampled
4. Fully subsampled

Note: Subsampling is not fully random. Samples are already influenced taxonomic classification at this sampling stage

Consider

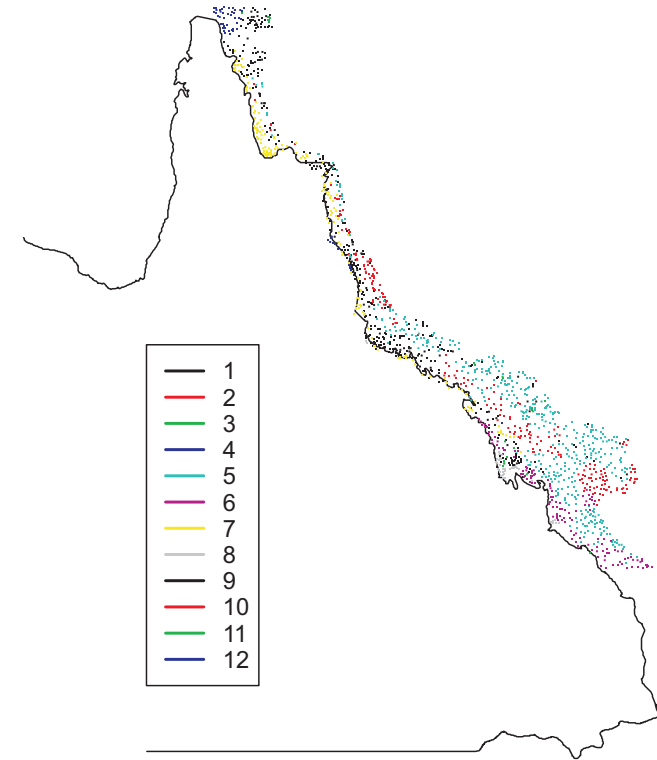
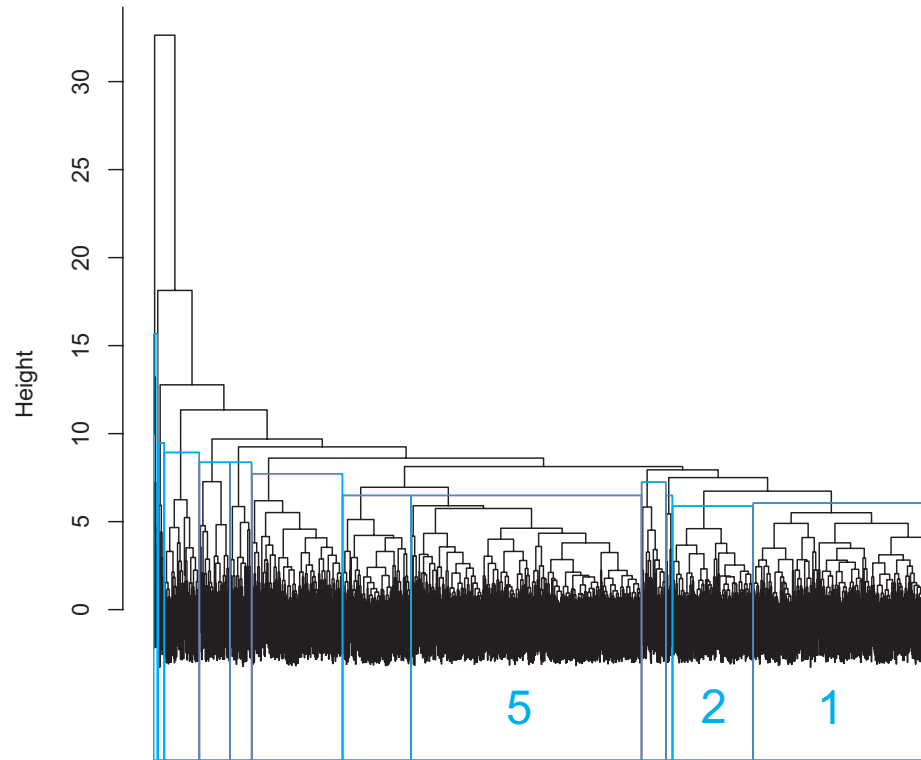
- number of species (richness);
- probability of presence of particular species.

Number of species

11

Strategy: Select a subset of survey sites where are homogeneity in terms of physical covariates, survey conditions and subsampling.

- To avoid the effect of physical covariates:
 - Make clusters of sites based on **11 physical covariates** (BATHY, STRESS, CRBNT, GRAVEL, SAND, MUD, NO3, S, T, SI, CHLA);
- To avoid the effect of survey conditions:
 - Select sites where **sled** was used;
- To avoid the effect of subsampling:
 - Select sites where all catches are sorted (**subsample ratio = 1**).



Number of sites in each cluster:

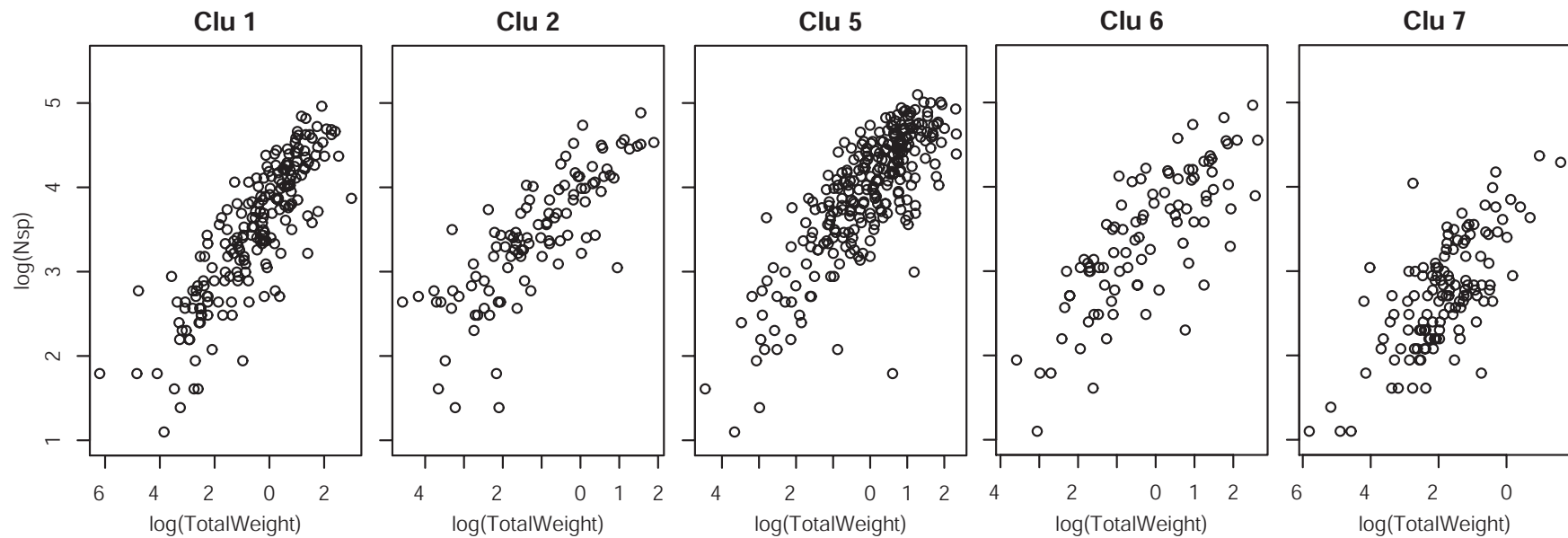
Cluster	1	2	3	4	5	6	7	8	9	10	11	12
Sites	351	159	13	48	454	135	179	43	60	69	12	7

Simple model base analysis:

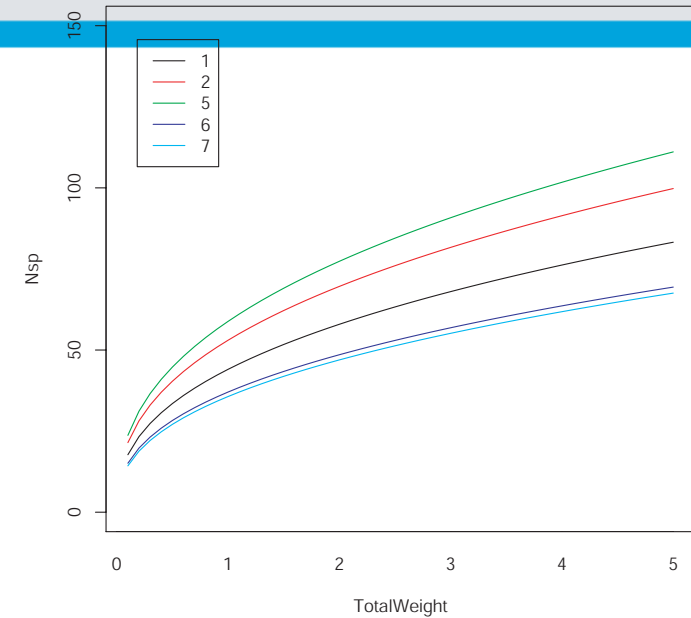
Number of species (S) is proportional to the power of total catch weight (W):

$$E[S] = aW^b$$

$$\log(E[S]) = \log(a) = \log(a) + b \log(W)$$



Clu No	$\log(\hat{\lambda})$	$\hat{\lambda}$	$\hat{\alpha}$
1	3.785	44.036	0.395
2	3.971	53.038	0.393
5	4.075	58.850	0.395
6	3.612	37.040	0.390
7	3.574	35.659	0.397

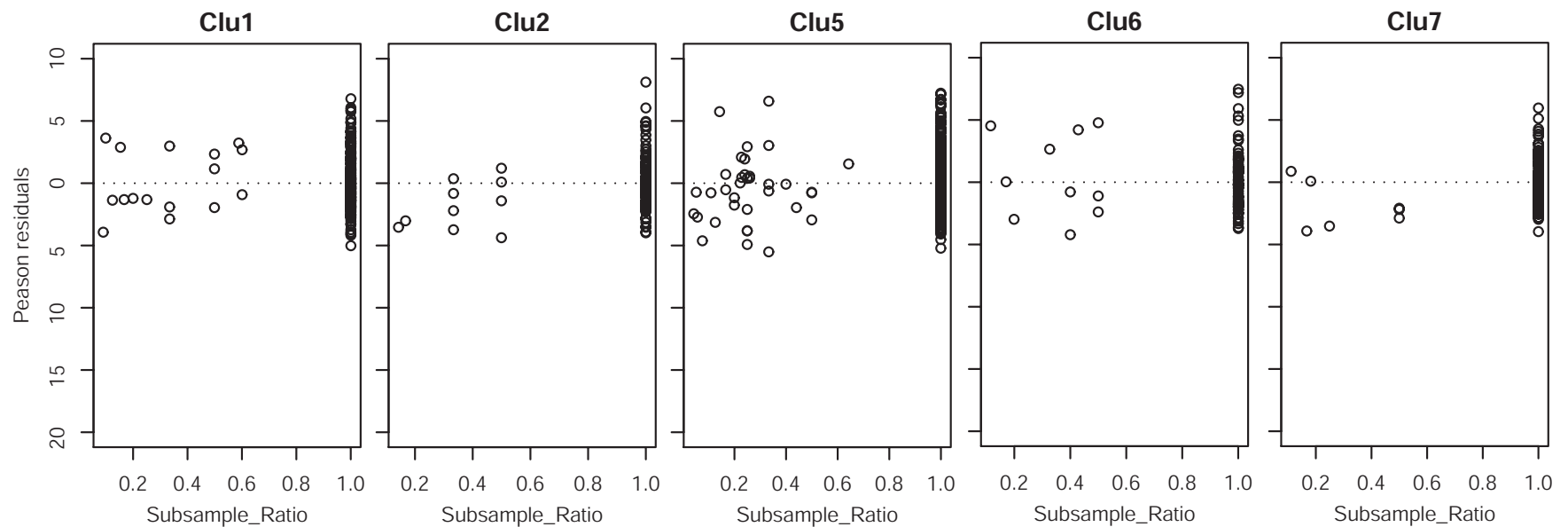


$$\log(E[\lambda]) = \log(\hat{\lambda}) = \log(\hat{\lambda}) + \log(\hat{\alpha})$$

Note if λ is proportional to α , $\hat{\alpha}$ should be 1.

Apply the model for the the data of which subsampling ratio $\neq 1$ except the sites where a particular huge volume of species were observed.

Residual plots:



Presence/absence of taxonomic group

Simple model base analysis:

$$(s | s) = \begin{cases} 0 & \text{absence} \\ 1 & \text{presence} \end{cases}$$

$$:= \sum_{s \in \mathcal{R}_j} (s | s)$$

$$\mathcal{R} = \bigcup_{=1}^5 \mathcal{R} = \{(0 \ 0 \ 2] \ (0 \ 2 \ 0 \ 4] \ (0 \ 4 \ 0 \ 6] \ (0 \ 6 \ 0 \ 8] \ (0 \ 8 \ 1]\}$$

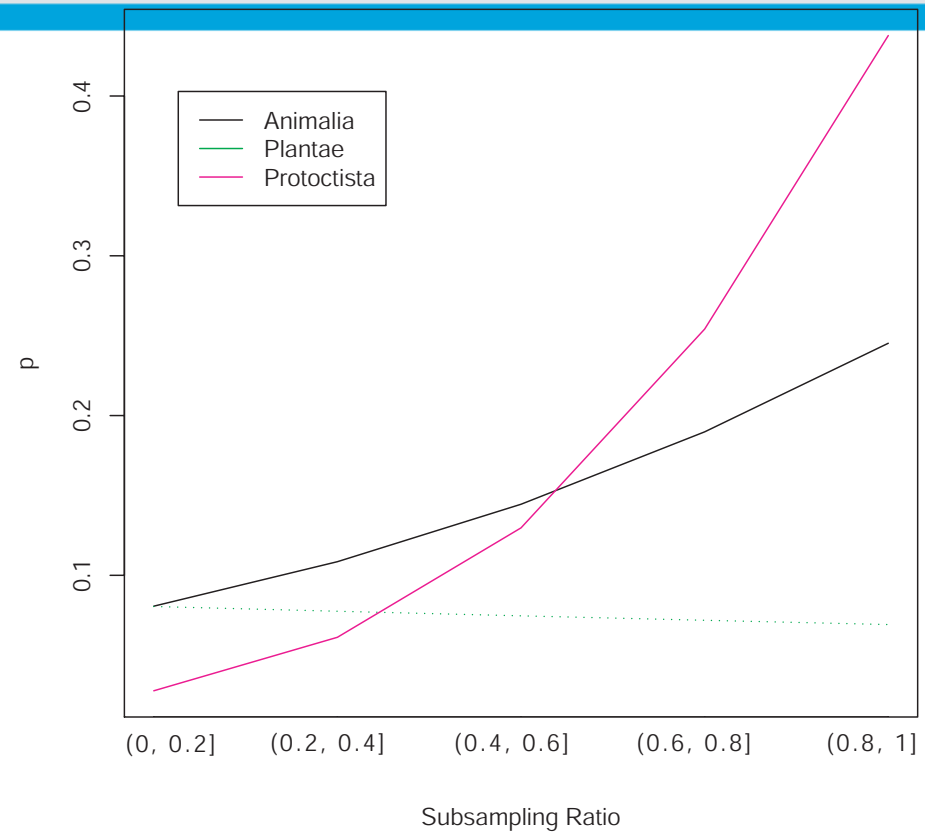
$$:= \Pr(= 1)$$

$$\Pr(=) = \binom{ij}{j} (1 -)^{j - ij}$$

$$\text{logit}() = 0 + 1 ()$$

Kingdom (界)

- *Animalia* (動物界)
- *Plantae* (植物界)
- *Protoctista* (原生生物界 ;
コンブ, アマクサ, ノリ)

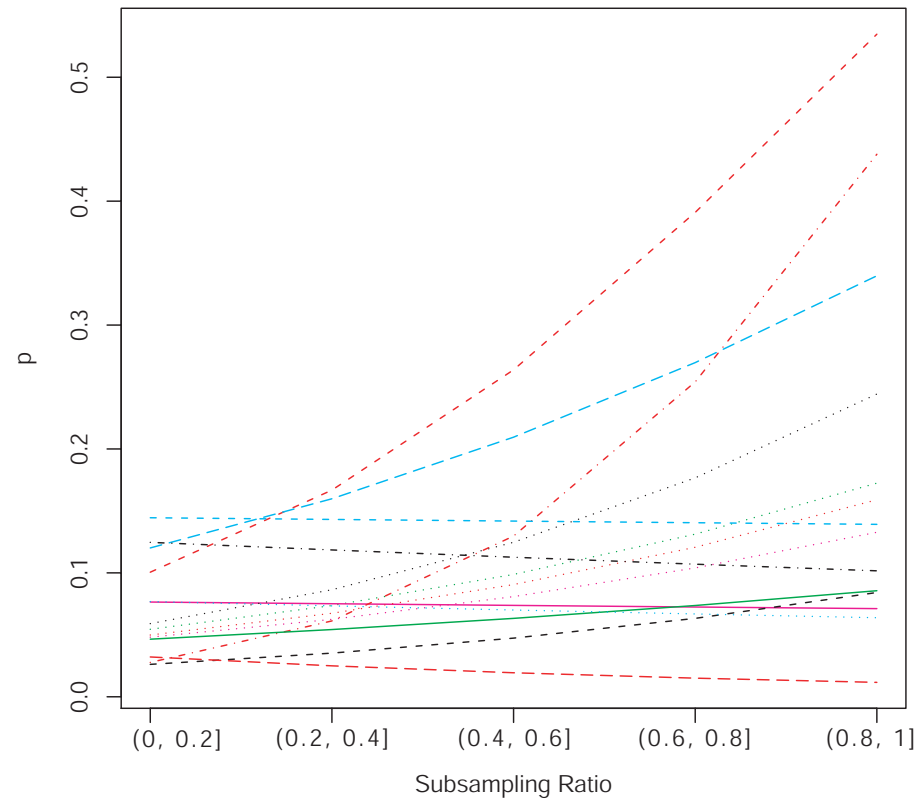


Phylum (門)

- *Annelida* (ゴカイ)
- *Foraminifera* (有孔虫)



- *Bryozoa* (コケムシ)



Summary

- Issues of sampling in marine surveys are addressed;
- Number of species (especially *Animalia*) seems not to be much influenced by subsampling;
- Presence/absence of taxonomic group seem to be influenced by subsampling depending on group.

References

1. Heales *et al.* (2003). Does the size of subsamples taken from multispecies trawl catches affect estimates of catch composition and abundance? *Fishery Bulletin*, **101**:790–799.
2. Pitcher *et al.* (2007). Seabed Biodiversity on the Continental Shelf of the Great Barrier Reef World Heritage Area. AIMS/CSIRO/QM/QDPI CRC Reef Research Task Final Report.
3. Ward *et al.* (2008). Presence-Only Data and the EM Algorithm. *Biometrics*, doi: 10.1111/j.1541-0420.2008.01116.x
4. Zaniwski *et al.* (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**: 261–280.

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

Ross DARNELL
Ross.Darnell@csiro.au

Hideyasu SHIMADZU
Hideyasu.Shimadzu@ga.gov.au