Capture-recapture, finite mixtures, correspondence analysis and working for David Vere-Jones

Shirley Pledger

11 April 2012, DVJ Seminar Series

Thanks, outline

Thanks to SRA organisers, audience.

Topics:

- Career, move to statistics
- Teaching
- Research
- Consulting

Not talking about probability.

Begin by talking about (not chance but) luck.

Move to Statistics

Luck in

- 1943,
- state education system,
- JTC's Mathematics Department,
- mathematics lectureship,
- KEP,
- HO, Wellington Polytechnic,
- JH, teaching Mandarin
- DVJ, statistics lectureship, tenured 1/2 time from 1980.

Early years in Statistics

Teaching, various courses.

Contacts with biologists, Ben Bell, Murray Efford, others later.

December 1983, enrolled for PhD in stochastic modelling, population dynamics. BPD supervising.

January 1984, Lloyd.

Suspended PhD enrolment.

Continued teaching. Ken half time.

Started investigating capture-recapture methods (needed for population dynamics input parameters).

Enter, the possums

Hooked on Capture-Recapture (CR)

Worked with Ecology Division, DSIR. Possums.

Outstanding problems with CR, especially biased estimates of population size in presence of individual heterogeneity of capture. Trap-induced heterogeneity with possums.

Re-enrolled for PhD, BPD again, changed topic to CR.

Found useful models for closed populations, using finite mixtures for individual heterogeneity.

PhD 1999, *Biometrics* publication 2000.

Overseas contacts: Ken Pollock, Carl Schwarz, etc. Euring Conferences.

CR with two lists or samples

Graunt, J. (1662). Mortality rates in London.

Laplace, P. (1786). Population of France,

Petersen, C.G.J. (1896). Danish plaice fishery.

Lincoln, F.C. (1930). Banded waterfowl, US.

Working off two lists or samples, simple proportionality argument.

Wildlife Applications - Lincoln-Petersen Method

Two samples.

Closed population (no birth, death, migration between samples).

Sample 1, catch and mark n_1 animals.

Sample 2, catch n_2, of which m have marks.

Estimate population size N: assume $\frac{m}{n_2} = \frac{n_1}{N}$,

$$\hat{N} = \frac{n_1 n_2}{m}.$$

Some Problems

$$\hat{N} = \frac{n_1 n_2}{m}.$$

- Need $m \geq 10$.
- Trap-shyness, m too low, $\hat{N}$ too high.
- Trap-happiness, m too high, $\hat{N}$ too low.
- Individual heterogeneity of capture, high capture and low capture animals, m too high, $\hat{N}$ too low.

K samples, individual marks, n animals seen

		Sampling Occasion							
		1	2	3	4	.	.	.	K
Animal	1	1	1	0	0	.	.	.	0
	2	0	0	1	0	.	.	.	1
	3	1	0	0	1	.	.	.	0
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	n	0	1	0	0	.	.	.	1
	$n+1$	0	0	0	0	0	0	0	0
	.	0	0	0	0	0	0	0	0
	.	0	0	0	0	0	0	0	0
	N	0	0	0	0	0	0	0	0

Data: n **by** K **Capture Matrix** - Estimate N, total number

Closed populations

No births, deaths, migration.

X_{ij} = 1 if animal i caught in sample j, else 0. $X_{ij} \sim Bern(p_{ij})$.

Null model, M(0), all $p_{ij} = p$.

Otis *et al.* 1978: Three major influences on p_{ij}:

t = time effect, all $p_{ij} = p_j$.

b = behaviour effect, $p_{ij} = p$ until 1st capture, then $p_{ij} = r$.
($p > r$, trap-shy behaviour; $p < r$, trap-happy behaviour.)

h = (individual) heterogeneity, $p_{ij} = p_i$.

Multinomial models

Each distinct capture history (row of X) is one category.

$$L \propto \frac{N!}{(N-n)!} \prod_{i=1}^{N} \prod_{j=1}^{K} p_{ij}^{x_{ij}} (1-p_{ij})^{1-x_{ij}}.$$

Model M(0):

$$L = \frac{N!}{(N-n)!} p^S (1-p)^{NK-S}$$

where S = sum(X) = total no. captures. $\hat{p} = S/(NK)$.

Similarly M(t), M(b).
However for M(h), need model for p_i to reduce npar.
Hierarchical, random effects, p_i from some distribution.

Use finite mixtures!

Norris and Pollock (1996). Models M(h) and M(b+h).

Pledger (2000), all feasible models with t, b, h.

$$L = \frac{N!}{(N-n)!} \prod_{i=1}^{N} \sum_{c=1}^{C} \pi_c \prod_{j=1}^{K} p_{ij}^{x_{ij}} (1-p_{ij})^{1-x_{ij}}$$

where π_c = prior prob. animal in class c ($c = 1, \ldots, C$).
If $i \in c$, p_{ij} becomes p_{cj}.
Use linear predictor $\log(p_{cj}/(1-p_{cj}))$ =

$$\mu + \tau_j + \mathscr{I}.\beta + \eta_c + \mathscr{I}.(\tau\beta)_j + (\tau\eta)_{cj} + \mathscr{I}.(\beta\eta)_c + \mathscr{I}.(\tau\beta\eta)_{cj},$$

where $\mathscr{I} = \mathscr{I}_{ibj}$ indicates if animal i is caught before j.

Finite vs. infinite mixture

More realistic to have infinite mixture?
Use beta distribution for p_i in M(h)?

2005: Simulation study evaluating $\hat{N}$. Several generating distributions, analysis by beta and two-group mixture.

- Finite vs infinite irrelevant. (Mediated through Bin(K, p_i).)
- Skewness of p_i distribution matters - low p_i gives most information about numbers not seen. Hence npar important: three pars (2-group mix) better than two (beta). Analysis using beta best if beta generating distribution, otherwise two-group mixture better.
- Beta doesn't generalise easily to time, behavioural effects.
- $\hat{N}$ good with finite mixture, but can't trust θ_1, θ_2, π_1 estimates.

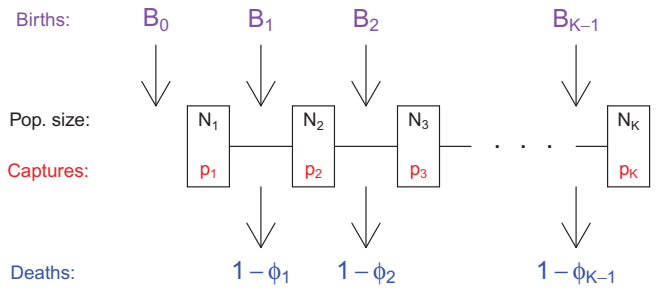
Open Populations

Models allowing for births and deaths.
Two papers using finite mixtures for survival rates:
- Bird banding, follow banded cohorts. (Pledger and Schwarz, 2002.)
- Extended a survival rate model which conditions on first capture, doesn't try to estimate N. (Pledger, Pollock and Norris, 2003.)

Since Lebreton *et al.* 1992, big developments in survival models, no estimation of abundance.

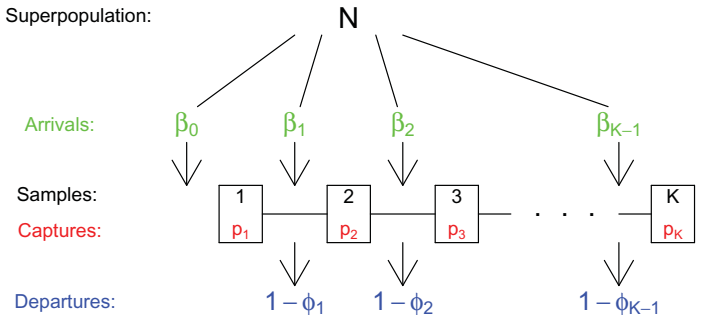
But when I talk to biologists, they always want to know population size.

Jolly-Seber Model



ϕ_j = Prob(survive to $j+1$ given alive at j).
Jolly (1965), Seber (1965). Model partly likelihood-based.

Schwarz-Arnason Model



N = number available for capture at least once during study.
β_j = proportion of N born between j and $j+1$.
Schwarz and Arnason (1996). Fully likelihood-based.

JSSA Model

$$L \propto \frac{N!}{(N-n)!}\prod_{i=1}^{N} L_i = \frac{N!}{(N-n)!}\left(\prod_{i=1}^{n} L_i\right) L_0^{N-n}$$

If animal i first seen at f_i, last at ℓ_i,

$$L_i = \sum_{b=1}^{f_i}\sum_{d=\ell_i}^{K} \beta_{b-1}\left\{\prod_{j=b}^{d-1}\phi_j\right\}(1-\phi_d)\left\{\prod_{j=b}^{d} p_j^{x_{ij}}(1-p_j)^{1-x_{ij}}\right\}$$

For an uncaught animal, all $x_{ij}=0$, change limits of summation:

$$L_0 = \sum_{b=1}^{K}\sum_{d=b}^{K} \beta_{b-1}\left\{\prod_{j=b}^{d-1}\phi_j\right\}(1-\phi_d)\left\{\prod_{j=b}^{d}(1-p_j)\right\}$$

Heterogeneous JSSA Model

$$L \propto \frac{N!}{(N-n)!}\left(\prod_{i=1}^{n}\sum_{c=1}^{C}\pi_c L_{ic}\right)\left(\sum_{c=1}^{C}\pi_c L_{0c}\right)^{N-n}$$

If $i \in c$ (prior prob π_c)

$$L_{ic} = \sum_{b=1}^{f_i}\sum_{d=\ell_i}^{K} \beta_{b-1}\left\{\prod_{j=b}^{d-1}\phi_{cj}\right\}(1-\phi_{cd})\left\{\prod_{j=b}^{d} p_{cj}^{x_{ij}}(1-p_{cj})^{1-x_{ij}}\right\}$$

$$\text{Uncaught}: \quad L_{0c} = \sum_{b=1}^{K}\sum_{d=b}^{K} \beta_{b-1}\left\{\prod_{j=b}^{d-1}\phi_{cj}\right\}(1-\phi_{cd})\left\{\prod_{j=b}^{d}(1-p_{cj})\right\}$$

Pledger, Pollock and Norris (2010).

Age-structured Models

Age-related survival? Senescence?
Previously people used first capture as birth time (or recruitment-to-adult time). Bias, only have lower bound for age.

With our formulation, can make ϕ depend on time since birth, ϕ_{aj} where a = age = no. years since birth, $a = j-b+1$.

$$L_i = \sum_{b=1}^{f_i}\sum_{d=\ell_i}^{K} \beta_{b-1}\left\{\prod_{j=b}^{d-1}\phi_{aj}\right\}(1-\phi_{ad})\left\{\prod_{j=b}^{d} p_{cj}^{x_{ij}}(1-p_{cj})^{1-x_{ij}}\right\}$$

Stopover duration analysis, Pledger *et al.* 2009.

Combine with heterogeneity. In prep. with Eleni Matechou.

Current work - sturgeon

Pledger and Baker, in prep.
Sturgeon live in Black Lake (near L. Michigan). Return to spawn in Black River, site fidelity. Eleven years of CR data.

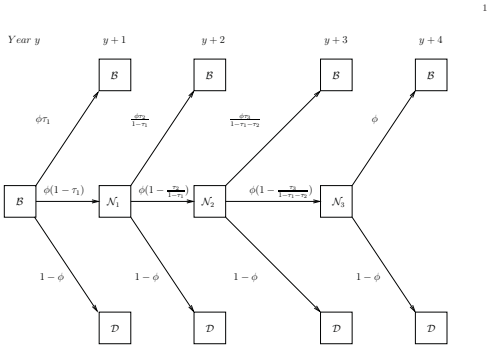
• Can only catch them once per year, in river.

• Long lived, don't breed every year. If not seen, either not breeding, or breeding but not caught.

• Want to estimate return time distribution.
τ_r = prob. next return to breed after r years, $\sum_{r=1}^{\max} \tau_r = 1$.

• Want abundance estimates - needed for monitoring population and setting fishing quotas.

Branching process for breeding states



Return time model

Modify JSSA model. "Birth" = first breeding during study.
Subdivide capture history into segments:

0 0 1 | 1 | 0 0 1 | 0 0 0 0

- Initial segment sums over possible times of "birth".
- Final segment sums over possible death times.
- All segments sum over all possible combinations of breeding and not breeding when alive but not seen.

Likelihoods, therefore can
• estimate return time distribution, including `max`,
• compare return time model with JSSA (`max`=1, $\tau_1 = 1$).

Have found **overestimation of population size** if analysed by JSSA when this type of temporary emigration is present.

Ecological communities - not CR!

Pledger and Arnold, in prep.

n by *p* matrix of binary (presence/absence) or count (abundance) data, *n* species, *p* samples.

Near redundancy - some species occur in very similar patterns, some samples have very similar species composition.

Biclustering - simultaneous clustering of rows and columns.

Finding structure in ecological data

Data from Whittaker (1956)

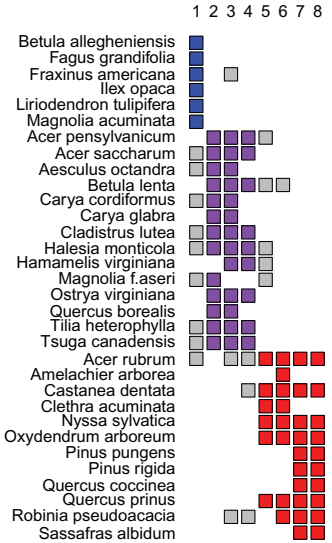
Presence/Absence of tree species in a deep valley forest, and at 25m intervals along a transect from moist to drier conditions.

Species	1	2	3	4	5	6	7	8
Acer pensylvanicum	0	1	1	1	1	0	0	0
Acer rubrum	1	0	1	1	1	1	1	1
Acer saccharum	1	1	1	1	0	0	0	0
Aesculus octandra	1	1	1	0	0	0	0	0
Amelachier arborea	0	0	0	0	0	1	0	0
Betula allegheniensis	1	0	0	0	0	0	0	0
Betula lenta	0	1	1	1	1	1	0	0
Carya cordiformus	1	1	1	0	0	0	0	0
Carya glabra	0	1	1	0	0	0	0	0
Castanea dentata	0	0	0	1	1	1	1	1
Cladistrus lutea	1	1	1	1	0	0	0	0
Clethra acuminata	0	0	0	0	1	1	0	0
Fagus grandifolia	1	0	0	0	0	0	0	0
Fraxinus americana	1	0	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tilia heterophylla	1	1	1	1	0	0	0	0
Tsuga canadensis	1	1	1	1	0	0	0	0

Sorted using 3 RG and 3 CG

	1	2	3	4	5	6	7	8
Betula allegheniensis	1							
Fagus grandifolia	1							
Fraxinus americana	1		1					
Ilex opaca	1							
Liriodendron tulipifera	1							
Magnolia acuminata	1							
Acer pensylvanicum		1	1	1	1			
Acer saccharum	1	1	1	1				
Aesculus octandra	1	1	1					
Betula lenta		1	1	1	1	1		
Carya cordiformus	1	1	1					
Carya glabra		1	1					
Cladistrus lutea	1	1	1	1				
Halesia monticola	1	1	1	1	1			
Hamamelis virginiana			1	1	1			
Magnolia f.aseri	1	1			1			
Ostrya virginiana		1	1	1				
Quercus borealis		1	1					
Tilia heterophylla	1	1	1	1				
Tsuga canadensis	1	1	1	1				
Acer rubrum	1		1	1	1	1	1	1
Amelachier arborea						1		
Castanea dentata				1	1	1	1	1
Clethra acuminata					1	1		
Nyssa sylvatica					1	1	1	1
Oxydendrum arboreum					1	1	1	1
Pinus pungens							1	1
Pinus rigida							1	1
Quercus coccinea							1	1
Quercus prinus					1	1	1	1
Robinia pseudoacacia			1	1		1	1	1
Sassafras albidum							1	1

Allocation to RGs and CGs



Count data (abundance)

Poisson building blocks. $E(X_{ij}) = \lambda_{ij}$

Paxillus involutus in Liphook Forest

Funghi in Liphook Pine Forest

	86	87	88	89	90
Bf	0	7	18	17	7
Cs	0	0	12	45	151
Gr	0	99	430	896	222
Il	0	0	16	1	6
Lp	567	2759	3868	182	266
Lr	0	0	0	2	47
Pi	117	131	20	21	40
Sb	0	219	2982	5823	2427
Sl	0	0	11	12	23
Sv	0	2	151	534	154

Ten species of toadstools, over five years.
Thanks to Peter Shaw, University of Roehampton.

Pattern detection models with count data

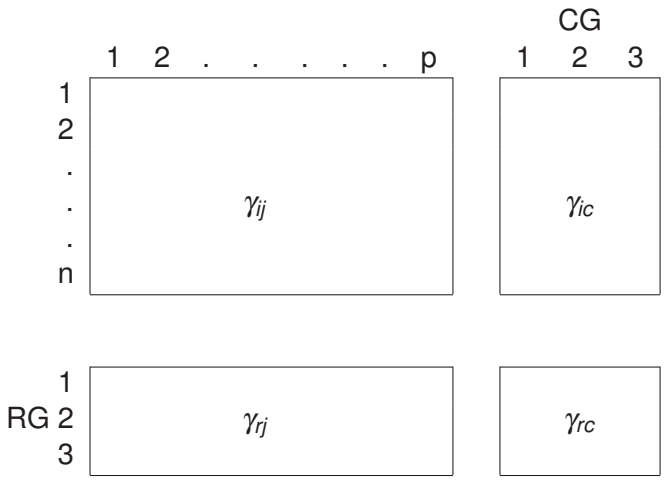
Null Model:	$\log \lambda_{ij} =$	$\mu + \alpha_i + \beta_j$
Cluster rows only:	$\log \lambda_{ijr} =$	$\mu + \alpha_i + \beta_j + \gamma_{rj}$
Cluster columns only:	$\log \lambda_{ijc} =$	$\mu + \alpha_i + \beta_j + \gamma_{ic}$
Cluster both rows and columns:	$\log \lambda_{ijrc} =$	$\mu + \alpha_i + \beta_j + \gamma_{rc}$
Saturated model:	$\log \lambda_{ij} =$	$\mu + \alpha_i + \beta_j + \gamma_{ij}$

Null and saturated: log-linear models, two-way contingency table.

Terms α_i and β_j allow for differing row sums, column sums. Clustering is driven by association patterns.

How much association is explained by the interpolated mixture models?

Cluster rows and/or columns

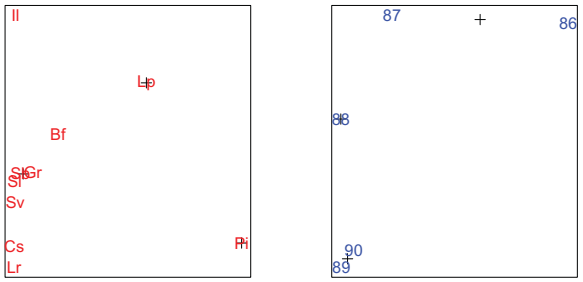


Plotting the patterns using γ

- Cluster rows, get profile plot of row groups
- Cluster columns, get profile plot of column groups
- Row clustering gives low-D scatterplot of columns (MDS)
- Column clustering gives low-D scatterplot of rows (MDS)

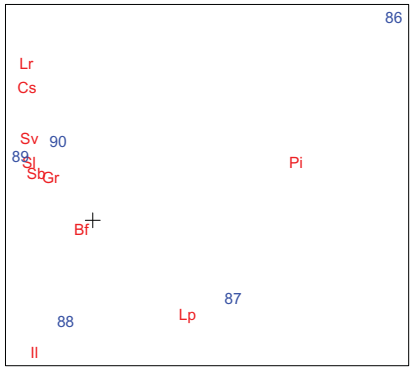
Three CGs gives 2-D plot of rows, on a stretched simplex in 3-D.

Two low-D plots for Liphook Forest data

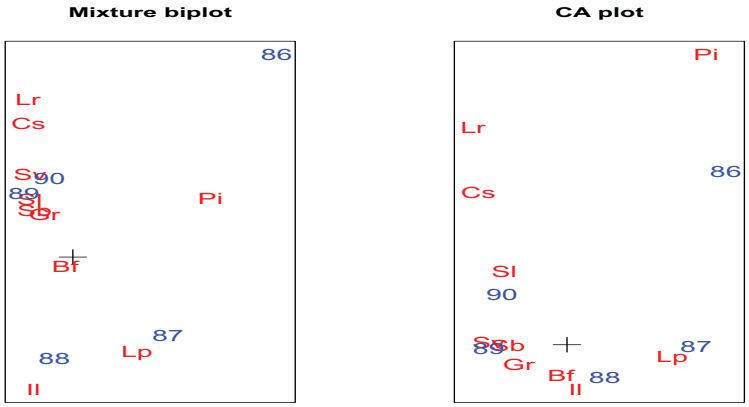


Mixture-based biplot for Liphook Forest data

Superimpose two low-D scatterplots using standard biplot methodology.



Compare with mixture biplot with CA plot



Likelihood-based analogue of correspondence analysis.

Advantages and Disadvantages

Biologists - grateful, generous with shared publications, field trips, BUT too many of them.

DVJ - ISOR, wonderfully supportive and stimulating working environment, my half-time tenured job, BUT handwriting, “secretary”, my desk ...