Markov-modulated nonhomogeneous Poisson processes for dealing with availability bias in surveys of marine mammal abundance

Roland Langrock (St Andrews)
David Borchers (St Andrews)
Hans Skaug (Bergen)

Conventional distance sampling

In short:

1.) record sightings of animals within predetermined strips

2.) quantify number of animals missed within strips

3.) scale corresponding estimate up to the area of interest

Details/assumptions involved in 2.):

- model detection probability $g(x)$ as a function of perpendicular distance x to trackline
- crucial assumption: $g(0) = 1$ (certain detection on the line)
- $g(0) \neq 1 \Rightarrow$ abundance estimates (negatively) biased
- for marine mammals typically $g(0) \neq 1$ ("availability bias")

Existing methods for dealing with availability bias

1.) correction factors
   - → very simplistic, can easily lead to substantially biased abundance estimates
2.) mark-recapture distance sampling (Borchers et al., 1998)
   - → logistically more challenging
   - → may suffer from bias due to unmodelled heterogeneity
   - → can't deal with animals who regularly dive longer than they are in view
3.) explicitly model availability via Poisson process (Skaug and Schweder, 1999)
   - → can give good estimates provided that availability events roughly follow a Poisson process

Overview of our approach

Two main reasons for missing whales in surveys:

1.) animal is underwater

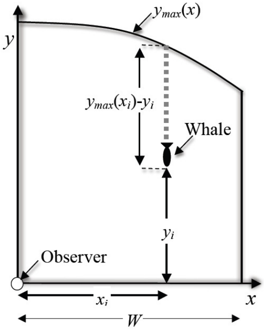
2.) animal is at surface but not detected

Our strategy:

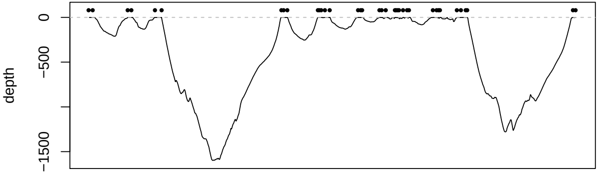
- quantify availability via suitable stochastic process
- model conditional detection probability, given the animal is available, as a function of distance from animal to observer
- formulate a corresponding model for detections
- use fitted model to obtain abundance estimates

Model illustration (for a single whale)

- observer at $(x = 0, y = 0)$
- assume animal movement is negligible
  $\rightarrow$ perpendicular distance fixed
- availability events occur according to some stochastic process
  (operating on the forward distance scale)
- detection process = thinned surfacing process (thinned according to some detection function)



MMPPs for surfacing events



- surfacings are continuous-time events that tend to occur in clusters (due to behavioural switches)
- this motivates the use of Cox processes:
  - Poisson process with rate $\lambda(I)$
  - $\lambda(I)$: unobserved non-negative stochastic process
- Markov-modulated Poisson process (MMPP):
  - Cox process where $\lambda(I)$ is a finite state Markov process
  - extends the approach of Skaug and Schweder (1999)

Embedding an MMPP in a DS survey

- MMPP can be used to describe surfacings
- but not all surfacings are associated with a detection...
- assume prob. of detections of surfacings to depend on x and y
- e.g., hazard probability:

$$h(x, y) = \mu \exp\left(-\frac{x^\gamma + y^\gamma}{\sigma^\gamma}\right)$$

- corresponds to a thinning of the MMPP
- detections are then realizations of a *nonhomogeneous* MMPP, with rates $\lambda_i h(x, y)$, $i = 1, \ldots, N$

→ given a fitted nonhomogeneous MMPP, standard techniques can be used to obtain abundance estimates

Likelihood of nonhomogeneous MMPP for detections

- bad news: in general, for nonhomogeneous MMPPs there is no closed-form likelihood expression available
- good news: we found such an expression for MMPPs with *piecewise constant* rates (i.e., $h(x, y)$ a step function)
- this can be exploited to approximate, arbitrarily accurately, the likelihood for any given form of $h(x, y)$
  (by approximating $h(x, y)$ by a step function,
  using a high number of intervals)

Proposition (single animal likelihood)

Assume that h is piecewise constant on the intervals $[i_0, i_1), \ldots, [i_{p-1}, i_p]$. Let $h^{(r)}$ denote the constant value of $h(l)$ on the rth interval and let $\boldsymbol{\Lambda}^{(r)} = diag(\lambda_1 \cdot h^{(r)}, \ldots, \lambda_N \cdot h^{(r)})$. Then the joint density of detections occurring at distances $l_1 < l_2 < \ldots < l_d$, with $i_{R_j} \leq l_j \leq i_{R_j+1}$, is given by

$$f(l_1, l_2, \ldots, l_d) = \boldsymbol{\pi} \prod_{j=1}^{d} \left(\mathcal{P}(l_{j-1}, l_j) \boldsymbol{\Lambda}^{(R_j+1)} \right) \mathcal{P}(l_d, y_{max}) \mathbf{1}^t \ ,$$

where $l_0 = 0$, $\mathbf{1} \in \mathbb{R}^N$ is a row vector of ones, and

$$\mathcal{P}(l_{j-1}, l_j) = \begin{cases} \exp\big((\mathbf{Q} - \boldsymbol{\Lambda}^{(R_j+1)})(l_j - l_{j-1})\big) & \textit{if } l_{j-1} \textit{ and } l_j \textit{ are in the same interval} \\ \exp\big((\mathbf{Q} - \boldsymbol{\Lambda}^{(R_{j-1}+1)})(i_{R_{j-1}+1} - l_{j-1})\big) \\ \quad \cdot \prod_{r=R_{j-1}+1}^{R_j-1} \exp\big((\mathbf{Q} - \boldsymbol{\Lambda}^{(r+1)})(i_{r+1} - i_r)\big) \\ \quad \cdot \exp\big((\mathbf{Q} - \boldsymbol{\Lambda}^{(R_j+1)})(l_j - i_{R_j})\big) & \textit{otherwise} \ . \end{cases}$$

Further remarks on statistical inference

- Data required:
  - survey data (perp. and forward distances from observer to sighted animals)
  - auxiliary data (e.g., tag data) on availability (otherwise identifiability issues)

- Inference:
  1.) estimate parameters of availability process from auxiliary data
  2.) fixing those parameters, use survey data to estimate parameters determining the conditional detection probability
  3.) based on fitted model, compute effective strip width, abundance estimates, etc.

Minke whale application

- Data:
  - seven sequences of minke whale surfacing events
  - 870 sightings of whales in the Northeastern Atlantic, made during shipboard surveys
- Analysis:
  - MMPPs were fitted to the series of surfacings
  - fixing the MMPP parameters, we estimated the detection probability (given an availability event):

$$h(y,x) = \frac{\theta_1 \theta_2^{\theta_3}}{(\theta_2^2 + x^2 + y^2)^{\theta_3/2}} ,$$

$$\text{logit}(\theta_2) = \beta_0 + \beta_1 I\big(B \in \{0,1\}\big) + \beta_2 I\big(B \in \{3,4\}\big) ,$$

with B giving the sea state in Beaufort (categorical variable)

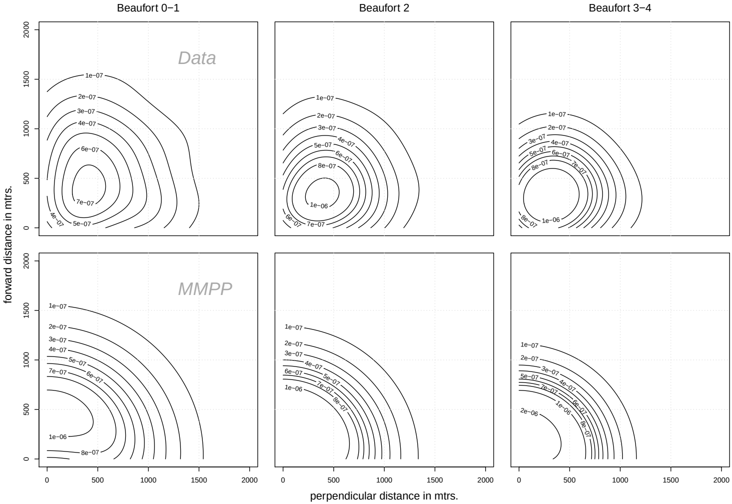


Figure: Bivariate nonparametric density estimates of observed sighting locations (top plots), and bivariate densities of sighting locations predicted by the fitted model (bottom plots).

Table: Estimated probability of detecting an animal that is on the trackline, $g(0)$, for the three different covariate levels (and 95% confidence intervals obtained through bootstrapping).

	estimate	95% *C.I.*
$g(0)$ for $B \in \{0, 1\}$	0.75	*(0.49,0.80)*
$g(0)$ for $B = 2$	0.68	*(0.42,0.73)*
$g(0)$ for $B \in \{3, 4\}$	0.61	*(0.35,0.66)*

And minke whales aren't even extreme divers!!

For beaked whales there can easily be a 60% chance of an individual not surfacing at all while within detectable range...

Table: Estimated probability of detecting an animal that is on the trackline, $g(0)$, for the three different covariate levels (and 95% confidence intervals obtained through bootstrapping).

	estimate	95% *C.I.*
$g(0)$ for $B \in \{0,1\}$	0.75	*(0.49,0.80)*
$g(0)$ for $B = 2$	0.68	*(0.42,0.73)*
$g(0)$ for $B \in \{3,4\}$	0.61	*(0.35,0.66)*

And minke whales aren't even extreme divers!!

For beaked whales there can easily be a 60% chance of an individual not surfacing at all while within detectable range...

- Borchers, D., Zucchini, W., Fewster, R. (1998). Mark-Recapture Models for Line Transect Surveys. *Biometrics*.
- Skaug, H., Schweder, T. (1999). Hazard Models for Line Transect Surveys with Independent Observers. *Biometrics*.
- Langrock, R., Borchers, D., Skaug, H. (in press). Markov-modulated nonhomogeneous Poisson processes for modeling detections in surveys of marine mammal abundance. *Journal of the American Statistical Association*.
- Borchers, D., Zucchini, W., Heide-Jørgensen, M., Cañadas, A., Langrock, R. (in press). Using hidden Markov models to deal with availability bias on line transect surveys. *Biometrics*.

(The slides of this talk can be found on my web page)