**Modelling Capture Probability from Large Electrofishing Datasets.**

Colin P. Millar, Robert J. Fryer, Karen J. Millidine and Iain A. Malcolm.

# Abstract

* Fisheries management and conservation decisions are often based on estimates of juvenile salmonid abundance obtained from electrofishing data.
* Accurate and un-biased estimates of abundance require estimates of capture probability.
* Capture probability and abundance can both vary with habitat and over time.
* Capture probability can be estimated using multi-pass electrofishing data.
* Hierarchical Bayesian models have been developed to model capture probability (and jointly fish density) but these approaches do not scale well to larger datasets. Computing power (big n) time to fit therefore model selection and exploration is prohibitively time consuming [Have they been used in cases of quantitative fishing only or in cases of mixed 1 and 3 passes?].
* Consequently, the use of these models appears to be restricted to smaller case studies and recent development in capture probability modelling do not seem to have been taken up by the applied fisheries community. Also smaller case studies can be easier to model as there is consistent uniform data.
* In larger scale models (WFD, DFO), however, the most common approaches are to either assume constant capture probability or estimate it on a sample by sample basis using a Zippin estimator.
* This paper presents an approach for estimating capture probability based on classical conditional likelihood methods which allow capture probability to be modelled in terms of linear and non-linear (non-parametric) relationships with covariates that affect capture probability
* Capture probability models were fitted to multi-pass electrofishing data for Scotland over the years 1996 to 2014. GIS covariates were included as proxies for habitat thereby removing the requirement for site specific habitat data. I Information on data source (organisation) allowed for variability in equipment, procedures and personnel.
* Sampling effects were the most important, followed by fish life stage, species and sampling date. Habitat characteristics such as width and gradient affected capture probability to a lesser extent.
* The results emphasise the risks of spatial bias in density estimates with consequences for management decision making [when inappropriate assumptions are made for capture probability?].
* Furthermore, we show modelling capture probability improves density estimates over two commonly used approaches in applied management: constant capture probability and site-wise estimates of capture probability.

Keywords: Atlantic Salmon, Brown Trout, Lifestage, [modelling words], Capture Probability, Electrofishing.

# Introduction

Fish abundance and distribution are important indicators of population heath, and estimating density from electrofishing data is a common problem in fishery management and conservation. Accurate estimates of abundance from electrofishing data require robust estimates of capture probability, however, capture probability can be difficult to estimate on a site by site basis when abundance is low (ref). Capture probability varies with habitat characteristics (Bayley and Dowling, 1990) and sampling procedure (Deuwalter and Fischer, 2007), which are also known to affect abundance (ref), hence ignoring differences in capture probability can produce biased estimates of abundance (Price and Petterson, 2010). Thus, obtaining accurate estimates of electrofishing capture probability and how it varies is an important issue for fishery management and conservation, particularly so when dealing with stocks at low abundance.

Estimating and modelling capture probability for single sites has a long history (Seber, 1982; Huggins and Yip, 1997; Huggins and Hwang, 2011). Recently models extending to multiple sites have been developed (Wyatt, 2004, Rivot et al., 2008; and Conroy et al., 2008) which jointly model capture probability and abundance using hierarchical Bayesian models (HBMs). Although HBMs provide a powerful tool for ecological modelling (Cressie et al, 2009), they are computationally intensive, and hence do so at the expense of model fitting and setup time. So, *despite such advances there is still a place for computationally simpler approaches which can be used routinely, especially when a range of candidate models are under consideration (Diggle and Ribeiro, 2007)* and large dataset are under consideration. In this paper a conditional likelihood approach (Huggins and Hwang, 2011) is developed that is a simplification over HBMs but provides savings in fitting time and allows for a wide range of models to be applied relatively easily.

To illustrate the benefits of a conditional likelihood approach an analysis of capture probability is presented for a large scale national dataset covering 2749 discrete sites and 6049 site visits in Scotland over the years 1996 to 2013. Covariates that require minimal site specific data collection are used to allow the incorporation of data from a range of sources with different sampling and recording procedures. Among the factors known to affect capture probability are velocity (Bayley and Dowling, 1990; Price and Peterson 2010), cross sectional area (Price and Peterson, 2010), fish length (Price and Peterson, 2010; etc.), wood density, site width (Hedger et al, 2013), total fish captured (Hedger et al, 2013; Pregler 2015), temperature, conductivity, undercut bank (Pregler 2015, Rodtka, 2015). Some of these features can be derived directly from a GIS (e.g., width), while GIS derived gradient, for example, can be considered to approximate velocity; other examples are landuse for wood density, upstream catchment for discharge.

The organisation of the paper is as follows. A modelling framework allowing capture probability to vary from site to site is developed, where covariates enter the model through a linear predictor. The application to a large scale data set follows, in which a best approximating model for capture probability for Scotland is selected. Finally, because it is still common in applied management to either ignore the effects of variable capture probability, or to estimate capture probability separately for each sample, simple density estimates from these approaches applied to the Scottish data are compared to the modelled capture probabilities, demonstrating the potential impacts for bias and/or precision when using un-modelled capture probability.

* Capture probability estimation requires multipass fishing.
* The argument is simply about practicality of local scale data collection for large datasets and the lack of common local standards for Scotland (although have them for other countries e.g. Habscore) and the qualitative nature of local assessments.

# Materials and Methods

## Capture efficiency modelling

If a closed population is assumed where all *N* fish have equal capture probability *p* which does not change with fishing pass, then an electrofishing experiment can be thought of as a binomial experiment where T fish are captured with probability p and an unknown number evaded capture with probability 1-p. Because it is known on what pass a fish was caught, capture evasions can be split into X *observed* capture evasions and R *unobserved* capture evasions. The counts T, X and R can be written in terms of the number of fish *ni* caught in the *i* th fishing pass:

,  and 

Where S is the number of fish passes. The joint probability of the counts *T*, *X* and *R* gives the joint likelihood



which dates back to Moran (19xx). This likelihood includes information from observed and unobserved fish and the number of fish observed is binomial with probability 1 – (1-p)S. Hence, if interest lies only in *p*, then a likelihood for *p* conditional on the *T* observed fish is



* We can assume that the logit of the capture probabilities can be expressed as a linear function of the covariates.
* Although this sounds limiting this can include: factor level means, splines and spatial models which we illustrate in the application below.
* Models are fitted using maximum likelihood using a robust optimisation tool provided by the stan library for R.

## Modelling considerations

* Modelling was based on a forwards and backwards stepwise selection procedure starting from a model with no covariates.
* Continuous covariates were considered first as linear effects then as smoothers with 3 degrees of freedom. By fixing the degrees of freedom it was possible to fit these effects as linear effects (see appendix for details).
* Hydrometric area was fitted as a categorical variable were neighbouring regions are correlated. Achieved through a smoothed spatial effect with 12 degrees of freedom which provided a reasonable compromise between complexity and model fit based on preliminary investigation. Again this can be fitted as a linear effect by fixing the degrees of freedom (appendix 1).
* Model selection was based on BIC. BIC was used due to provide more parsimonious models give the large number of observations.
* Importance of covariates was assessed by the change in BIC by dropping one covariate at a time from the final model.
* Deviance test for overdispersion – chi square test
* Residuals for multinomial p: difference in samplewise likelihood
* If overdispersed then do a simple quasi likelihood type approach.

## Electrofishing data

* We modelled the capture efficiency from electrofishing in 208 catchments in Scotland between 1980 and 2014. **Figure 1** this consistent to 22000 site visits over xx distinct sites.
* Data was obtained from the fisheries trust (local level), MSS, SEPA (Scotland level) and Caithness DSFB. Again **Figure 1, and Figure 2** - table.
* Electrofishing samples varied between 2 and 6 passes, however the majority of samples had 3 (84%) or 2 (14%) passes.
* Counts of salmon and trout were collected at each pass with life stage (fry or parr) being estimated in the field.
* Common recording included location of sample site, date, and fished area.

## Explanatory data / Covariates

* Several types of covariates considered: habitat, sampling, spatial and temporal. [Note consider salmon and trout and life stage as separate models first then think about combining them]
* These were all generated retrospectively in a consisted way. In the case of habitat variables these were generated in a GIS.
* The habitat covariates considered are: altitude, upstream catchment area, distance to sea, gradient, landuse and channel width. These are proxies for processes known to affect capture probability, for example altitude affects river temperature, gradient influences channel velocity. More explanation of land use.
* All effects of sampling procedures are explained through the use of organisation, a categorical variable with xx levels.
* To account for spatial variability not incorporated in habitat variables we use SEPA hydrometric area which represents spatially coherent catchment groupings of roughly similar areas, a categorical variable with x levels (Figure 1). These contained between 1 (e.g. Tay) and x (e.g. west coast) catchments reflecting regional differences in catchment size.
* Temporal effects were year which is treated as a factor, and day of the year continuous.
* A scatterplot covariates of continuous covariates is shown in **fig 2**

## Performance of different estimates of capture probability

* In order to assess the relative performance of capture probability approaches, site-wise estimate of abundance and their variance was compared.
* 1) Common capture probability (bias); 2) different capture probability for each site (precision); 3) modelled p.
* Put the equation for ML estimates of abundance.

# Results

## Capture efficiency modelling

* The best approximating model was: [and state model]
* List effects in order of importance
* Plot effects and describe them
* Plot residuals? [temporal variability by organisation]
* Brief understanding: this is consistent with understanding of… [this can include the interpretation to focus the discussion on the modelling]

## Performance of different estimates of capture probability

* Plot maps for 1) – describe spatial trends…
* Plots comparing errors in 2)

# Discussion

## General discussion/recommendations

* We present an approach for estimating and modelling electrofishing capture probability for large datasets that provides a balance of complexity and computational efficiency.

## Application

* What was found and how it relates to literature. Is it consistent with previous studies?
* Discuss how GIS was at generating proxies for the reported influencers on catchability.
* Issues with spatial confounding. Spatial terms can be thought of as capture un-modelled variation – advisable to attempt to find covariates that can describe the likely cause of the variation possibilities include. Other spatial model to be used as covariates: SRTMN.
* equipment use:
  + Organisation is a mixture of effects. Something like: A variety of sampling methods were used across the organisations who supplied data: with and without stop nets, backpack electrofishing, bank based equipment and generators. Unfortunately information on sampling equipment was not routinely or reliably recorded across data sources and as such could not be formally included in the analysis.
  + In a more detailed study it might be possible to model these effects separately (see refs )
* Talk about correlation between covariates – talk about differences between distance to sea, altitude gradient. It is a potential issue (compute correlations and quote maximum correlation).
* Discuss residual diagnostics?

## Methods

* A wide variety of models can be fitted as linear terms when the degrees of freedom are selected prior to analysis, for example, interactions smoothers and … refer to package.
* Work is underway to … fitting penalised splines then AIC or GCV can be used to estimate the smoothing parameters. It is important when using AIC to estimate the appropriate degrees of freedom of the smoothing terms which reduce as penalisation increases.
* Although HBMs are a very useful tool in ecological modelling (Cressie), they can be difficult to extend to large datasets such as the Scotland wide data being analysed here. Previous HBM developments include: Conroy et al. (2008), Wyatt, Rivot. However all of these approaches also consider density either as a random effect or in terms of covariates and in order to model capture probability end up with a hierarchical Bayesian model. Why is this approach better – model selection and exploration. Although people don’t report time taken fit models. SNIFFER reported it took 7 days for model fitting, whereas the approach reported here takes 15mins for the entire model selection procedure.

## Management consequences

* An issue identified is the lack of uptake of recent developments for modelling capture probability into the applied fisheries community. This paper highlights Management consequences when assuming constant capture probability / site by site capture prob.
* Although single pass electrofishing allows for greater spatial coverage, it comes with implication of bias and precision.
* The use of a model for capture probability would allow the incorporation of single pass fishing data (in conjunction with sufficiently representative multipass data) into quantitative advice.
* Although the model can be used to estimate p for new sites, it is still essential to maintain a level of multipass fishing to permit valid use of single pass sampling. Critical given year and organisation effects, especially where it is unknown what causes the effects of organisation! Also important to record information on sampling approaches and equipment and personnel. New databases allow this but not possible for older data. Effects of various components should be investigated.

# References

# Tables

# Figures

# Appendix 1

Include the derivation of reduced rank spatial effects.

NOTES:

Include reference to software.