Semiparametric Regression in Capture-Recapture Modeling

O Gimenez, C Crainiceanu, C Barbraud, S Jenouvrier, B J T Morgan

Semiparametric regression in capture-recapture modelling

O. Gimenez,[1,2] C. Crainiceanu,[3] C. Barbraud,[4] S. Jenouvrier[4] and B.J.T. Morgan[1,*]

[1]Institute of Mathematics, Statistics and Actuarial Science

University of Kent, Canterbury

Kent, CT2 7NF - UK

[2]Centre d'Ecologie Fonctionnelle et Evolutive - CNRS

1919 Route de Mende

34293 Montpellier Cedex 5 - France

[3]Johns Hopkins University

615 N. Wolfe St. E3037

Baltimore, MD 21205 - USA

[4]Centre d'Etudes Biologiques de Chizé

CNRS UPR 1934

79360 Villiers en Bois - France

4 October 2005

Summary. Capture-recapture models were developed to estimate survival using data arising from marking and monitoring wild animals over time. Variation in survival may be explained by incorporating relevant covariates. We propose nonparametric and semiparametric regression methods for es-

* *email:* B.J.T.Morgan@kent.ac.uk

timating survival in capture-recapture models. A fully Bayesian approach using MCMC simulations was employed to estimate the model parameters. The work is illustrated by a study of Snow petrels, in which survival probabilities are expressed as nonlinear functions of a climate covariate, using data from a 40-year study on marked individuals, nesting at Petrels Island, Terre Adélie.

KEY WORDS: auxiliary variables; Bayesian inference; demographic rates; environmental covariates; penalized splines; WinBUGS.

1. Introduction

Understanding population structure and changes in that structure for wild animals is essential for both species conservation and management. Because of human activities, it appears crucial to explain and forecast the effects of climatic and environmental perturbations on population dynamics. The analysis of data arising from observations of marked animals is an important tool for estimating demographic parameters that govern population change.

In the last forty years, a challenging research topic has been the estimation of wild animal survival, and when possible, to explain variations in survival using auxiliary variables such as time, age of animal or relevant covariates like temperature or rainfall. Many traditional models exhibit a product-multinomial likelihood structure, allowing inference in a unified context by classical maximum likelihood (Lebreton et al., 1992) through user-friendly software like MARK (White and Burnham, 1999) or M-SURGE (Choquet et al., 2005). The Bayesian approach has been proposed as an alternative – see Brooks et al., 2000 for a review.

To estimate survival probability, the modeling is usually embedded in the

Generalized Linear Model (GLM) framework (Lebreton et al., 1992). A logit link for survival probabilities is frequently used but other functions are possible (Williams et al., 2002); covariates may be readily incorporated, and here we will focus on environmental covariates that vary over sampling occasions but remain constant over individuals, as defined by Pollock (2002). Most frequently, covariates are related to survival by a linear or a quadratic function, on the logit scale. However, in general this may be unrealistic and we give three examples to motivate a nonlinear alternative. First, it has been shown that global indices such as the North Atlantic Oscillation (NAO) could relate to population dynamics in complex nonlinear ways (Mysterud et al., 2001; see also Stenseth and Mysterud, 2002 for a general discussion). Secondly, survival can be nonlinearly related to population density via a threshold effect (Lima, Merritt and Bozinovic, 2002). Thirdly, survival as a function of age may exhibit non-linear patterns, through senescence defined as a reduction in survival among old individuals (Loison et al., 1999; Catchpole et al., 2004). In these examples and in many others, a nonparametric alternative avoids strong parametric assumptions and is of interest in itself. It might also suggest a new, scientifically relevant, parametric model if one is needed.

In this paper we applied Generalized Additive Models (GAMs) ideas popularized by Hastie and Tibshirani (1990) that extend the traditional GLM framework. Rather than specifying a fixed link between survival and covariates in the model, the shape of the relationship is determined by the data, using penalized splines (Ruppert et al., 2003). Our choice has been guided by the equivalence between a penalized spline formulation of the nonparametric problem with Generalized Linear Mixed Models (GLMMs) that simplifies

further extensions.

The paper is organized as follows. In the next section, we give the likelihood for classical survival models, and the nonparametric regression of survival probabilities on covariates is established. In Section 3, we consider a natural extension to the nonparametric model, when a semiparametric regression model for survival is introduced. As well as including the nonparametric component, this allows us to model a parametric component at the same time. Section 4 gives the details of the Bayesian inference and its implementation through MCMC simulations. Section 5 gives the results of a simulation study which validate the ability of our approach to capture various nonlinearities in survival. Section 6 illustrates our method using data from a 40-year study of individually marked Snow petrels (*Pagodroma nivea*), in trying to relate their survival to a climate covariate. The last section gives general conclusions and discusses the potential of our approach.

2. Theory

2.1 *CJS likelihood*

We assume here that our capture-recapture study includes $I+1$ sampling occasions at which animals are caught or observed, so that I recaptures or re-observations may actually be made. On each occasion, new unmarked animals are given unique marks and then released. Previously marked animals can also be sampled, and after their identity is recorded they are also released back into the studied population. This protocol gives rise to a set of animal encounter histories, made up of 1 and 0 depending respectively on whether an animal is detected or not. Cormack (1964), Jolly (1965) and Seber (1965) independently derived the likelihood for such capture-recapture

data, and this model is referred to as the CJS model. Schwarz and Seber (1999) and Williams et al. (2002) give reviews of the CJS model and its applications. Note that the model includes time-variation in parameters, but no age-variation. It may therefore be appropriate for describing the survival of adult animals. Data are frequently summarized in an upper triangular array, $\mathbf{m}$, called the m-array, where m_{ij}, $i = 1, \ldots, I$, $j = i+1, \ldots, I+1$, is the number of animals released at time t_i and subsequently recaptured for the first time at time t_j. Also the column vector $\mathbf{R}$ contains the R_i, $i = 1, \ldots, I$, which are the numbers of marked animals released into the population at times t_i; these comprise newly marked animals and those previously marked animals that are recaptured at time t_i. Under the assumption that animals are independent (see e.g. Williams et al., 2002 for a description of CJS model assumptions and consequences of possible violation), the likelihood is product-multinomial

$$[\mathbf{m}|\phi, \mathbf{p}, \mathbf{R}] \;\propto\; \prod_{i=1}^{I} \chi_i^{R_i - r_i} \prod_{j=i+1}^{I+1} \left\{ \phi_i p_j \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\}^{m_{ij}} \qquad (1)$$

where $[X]$ denotes the distribution of X, ϕ_i, $i = 1, \ldots, I$, is the probability that an animal survives to time t_{i+1} given that it is alive at time t_i and p_j, $j = 2, \ldots, I+1$ denotes the encounter probability of being detected at time t_j (see e.g. Brooks et al., 2000). We adopt the convention that a null sequence has product 1 so that for example $\prod_{k=i+1}^{j-1} \phi_k(1 - p_k) = 1$ for $j = i + 1$. Other terms involve $r_i = \sum_{j=i+1}^{I} m_{ij}$, the number of animals subsequently recaptured after release at time t_i and χ_i, the probability that an animal, alive at time t_i, is not subsequently encountered. This can be calculated recursively as $\chi_i = 1 - \phi_i \{1 - (1 - p_{i+1})\chi_{i+1}\}$, with $\chi_{I+1} = 1$ (e.g. Lebreton

et al., 1992).

2.2 *Nonparametric regression of survival*

We consider a nonparametric regression model for the probability that an animal survives from time t_i to time t_{i+1} of the form

$$\text{logit}(\phi_i) \quad = \quad f(x_i) + \varepsilon_i, \quad i = 1, \ldots, I \tag{2}$$

where x_i is the value of the covariate for the ith sampling occasion, ε_i are i.i.d $N(0, \sigma_\varepsilon)$, ε_i is independent of x_i and f is a smooth function. Here, the random effects $\{\varepsilon_i\}$ allow us to model the residual sampling-occasion-to-sampling-occasion variation not described by the covariates alone (Barry et al., 2003). Variations on the model of Equation (2) include:

- Semiparametric regression models in which some of the predictors enter linearly in the model, as illustrated in Section 3, and
- Models including interactions between covariates which are discussed in the last section.

Penalized splines using the truncated polynomial basis (Ruppert, 2002) were used to model the smooth function

$$f(x|\eta) \quad = \quad \beta_0 + \beta_1 x + \ldots + \beta_P x^P + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^P \tag{3}$$

where $P \geq 1$ is an integer, $\eta = (\beta_1, \ldots, \beta_P, b_1, \ldots, b_K)^T$ is a vector of regression coefficients, $(u)_+^p = u^p \mathbf{I}(u \geq 0)$ and $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ are fixed knots. The crucial problem in using relation (3) is the choice of the number and the position of the knots. A small number of knots may result in a smoothing function that is not flexible enough to capture variability in the data, whereas

a large number of knots may lead to overfitting. Similarly, the position of the knots will influence estimation. We used a penalized splines approach inspired by the smoothing splines of Green and Silverman (1994). First, the number of knots is chosen to ensure enough flexibility. Following Ruppert (2002), we considered $K = \min\{\frac{1}{4}I, 35\}$ and let κ_k be "equally-spaced sample quantiles" i.e. the sample quantiles of the x_i's corresponding to probabilities $k/(K+1)$. Other choices are possible, such as equally spaced knots within the domain of x, and Crainiceanu et al. (2004a) provide a simulation study comparing these two alternatives with a discussion. Then, following Ruppert et al. (2003) a quadratic penalty is placed on $\mathbf{b}$, which is here the set of jumps in the Pth derivative of $f(\bullet|\eta)$ so that with Equation (3) we associate the constraint

$$\mathbf{b}^T\mathbf{b} \;\leq\; \lambda \tag{4}$$

where λ is called the smoothing parameter. Equations (3) and (4) lead to the so-called P-splines approach (see e.g. Lang and Brezger, 2004). Because roughness is controlled by the penalty term (4), once a minimum number of knots is reached, the fit given by a P-spline is almost independent of the knot number and location (Ruppert, 2002).

P-spline models can be fruitfully expressed as GLMMs, which facilitates their implementation in standard software (Ngo and Wand, 2004; Crainiceanu et al., 2004b), and above all provides a unified framework for generalizations of the nonparametric model. Indeed, we first note that the P-splines approach is equivalent to minimizing

$$\sum_{i=1}^{I} \{\text{logit}(\phi_i) - f(x_i|\eta)\}^2 + \frac{1}{\lambda}\eta^t\mathbf{D}\eta \;, \tag{5}$$

where $\mathbf{D}$ is a known positive semi-definite penalty matrix. The truncated spline penalty matrix is

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{P\times P} & \mathbf{0}_{P\times K} \\ \mathbf{0}_{K\times P} & \mathbf{\Omega}_K \end{bmatrix},$$

where a standard choice for $\mathbf{\Omega}_K$ is $\mathbf{I}_K$. To avoid overfitting, the matrix $\mathbf{D}$ penalizes only the coefficients of the spline basis functions $(x-\kappa_k)_+^P$. Let $\phi = (\phi_1, \ldots, \phi_I)^T$, $\mathbf{X}$ be the matrix with the ith row $\mathbf{X}_i = (1, x_i, \ldots, x_i^P)^T$, and $\mathbf{Z}$ be the matrix with ith row $\mathbf{Z}_i = \{(x_i-\kappa_1)_+^P, \ldots, (x_i-\kappa_K)_+^P\}^T$. If we divide Equation (5) by the error variance σ_ε^2 we obtain

$$\frac{1}{\sigma_\epsilon^2}\left\|\text{logit}(\phi) - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}\right\|^2 + \frac{1}{\lambda\sigma_\epsilon^2}\mathbf{b}^T\mathbf{b}\,,$$

where $\beta = (\beta_0, \ldots, \beta_P)^T$ and $\mathbf{b} = (b_1, \ldots, b_K)^T$. Define $\sigma_b^2 = \lambda\sigma_\varepsilon^2$, consider the vector β as fixed parameters and the vector $\mathbf{b}$ as a set of random parameters with $E(\mathbf{b}) = 0$ and $cov(\mathbf{b}) = \sigma_b^2\mathbf{I}_K$. If $(\mathbf{b}^T, \varepsilon^T)^T$ is a normal random vector and $\mathbf{b}$ and ε are independent, then an equivalent model representation of the P-spline model in the form of a GLMM is

$$\text{logit}(\phi) \;=\; \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon, \quad \text{cov}\begin{pmatrix} \mathbf{b} \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2\mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2\mathbf{I}_I \end{pmatrix} \qquad (6)$$

for which $E(\text{logit}(\phi)) = \mathbf{X}\beta$ and $\text{cov}(\text{logit}(\phi)) = \sigma_\varepsilon^2\mathbf{V}$ where $\mathbf{V} = \mathbf{I}_I + \lambda^2\mathbf{Z}\mathbf{Z}^T$ (Brumback et al., 1999).

Note that the connection between the P-spline model and the mixed model of Equation (6) allows us to extend the nonparametric model to incorporate other nonparametric components as well (Ruppert et al., 2003).

3. Semiparametric regression of survival

In the preceding section, a regression model for survival over a continuous predictor modeled as a smooth function was considered. In this section, we

extend this model by including qualitative predictors assumed to enter the model linearly. Without loss of generality, we considered only one parametric categorical component s with one non-parametric component smoothing a continuous predictor x by linear P-splines. We want to let the relationship between $\text{logit}(\phi_i)$ and x_i vary differently but in parallel according to the variable s_i taking discrete values, i.e.

$$\begin{aligned} \text{logit}(\phi_i) &= \beta_0 + \gamma s_i + \beta_1 x_i + \\ &\quad \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \varepsilon_i, \quad i = 1, \ldots, I. \end{aligned} \tag{7}$$

The GLMM representation can also be used to handle the semiparametric model. Let us adjust the matrix $\mathbf{X}$ so that its ith row is $\mathbf{X}_i = (1, s_i, x_i)^T$ and $\beta = (\beta_0, \gamma, \beta_1)^T$, while the ith row of matrix $\mathbf{Z}$ is $\mathbf{Z}_i = \{(x_i - \kappa_1)_+, \ldots, (x_i - \kappa_K)_+\}^T$. Then the mixed model defined by Equation (6) can still be used to describe the semiparametric regression defined in Equation (7) (Ruppert et al., 2003).

4. Bayesian inference

In this section, we focus on the Bayesian analysis of the nonparametric model defined in Section 2.2. However, within the GLMM framework introduced before, the extension to additive and semiparametric models is straightforward (see Section 6).

The frequentist approach would require maximising the likelihood, which is obtained by integrating the distribution $[\mathbf{m}|\phi, \mathbf{p}, \mathbf{R}]$ over the random effects ε_i and b_k. This is therefore a problem involving a high dimensional integral that could be handled by using approximations like Laplace's method (Chavez-Demoulin, 1999; Wintrebert et al., 2005) or asymptotic arguments

(Burnham, 2002). For fitting our models, we opted for a Bayesian approach through Gibbs sampling. Invoking conditional independence properties, a first step is achieved by recursively factorizing the posterior distribution to give:

$$\begin{aligned}
&[\beta, \mathbf{b}, \varepsilon, \sigma_b^2, \sigma_\varepsilon^2, \mathbf{p}, \mathbf{R}|\mathbf{m}] \\
\propto \quad &[\mathbf{m}|\phi, \mathbf{p}, \mathbf{R}][\phi|\beta, \mathbf{b}, \varepsilon][\beta][\mathbf{b}|\sigma_b^2][\varepsilon|\sigma_\varepsilon^2][\sigma_b^2][\sigma_\varepsilon^2][\mathbf{p}]. \qquad (8)
\end{aligned}$$

Even if one is only interested in the marginal posterior distribution of a subset of parameters, high-dimensional integrations have to be carried out. In general, such complex integrals are intractable analytically and we made use of MCMC methods which provide powerful computer-intensive methods for making approximations (e.g. Brooks, 1998). We employed Gibbs sampling (e.g. Casella and George, 1992), however, in the context of capture-recapture model parameter estimation, generally full conditional distributions are non-standard (Brooks et al., 2000; Barry et al., 2003; Johnson and Hoeting, 2003), so that usual random variate generation algorithms cannot be used. Instead, more elaborate algorithms are needed such as adaptive rejection sampling or Metropolis-within-Gibbs sampling (see Gilks, 1996 for a review). We therefore used software WinBUGS (Spiegelhalter et al., 2003), which performs the latter.

5. Simulation study

Before turning to the real example, we conducted a simulation study to provide empirical support for our approach. We considered two scenarios with different forms for the underlying nonlinear regression function f of Equation (2). Study 1 used the regression function $f(x) = 2.2$ if $x \leq -0.06$

and $f(x) = 2.08 - 2\ x$ otherwise. This function is a broken line which mimics a threshold effect, for instance the covariate might represent an environmental constraint on resources, which negatively affects survival only above a given level. The xs were equally spaced on $[-1.5, 1.5]$, and the error variance σ^2_ε was equal to 0.1. Study 2 used the regression function $f(x) = 1.5\ g((x - 0.35)/0.15) - g((x - 0.6)/0.1)$ where $g(x) = \exp(-x^2/2)/\sqrt{2\pi}$. This function exhibits non-trivial non-linear patterns, which could correspond to complex relationships between climatic conditions and survival. The xs were equally spaced on $[0, 1]$, and the error variance σ^2_ε was equal to 0.02. For both studies, we simulated 50 capture-recapture data sets covering 26 sampling occasions, so that 25 survival probabilities had to be estimated, with 100 newly marked individuals per occasion. The capture probability was set constant and equal to 0.7.

For 5 randomly chosen data sets, we first ran two overdispersed parallel MCMC chains to check if convergence was reached. As a result, we decided to use 100,000 iterations with 50,000 burned iterations for posterior summarization. Details on the priors used and the convergence assessment can be found in Section 6. We then applied our non-parametric approach on each data set, using linear P-splines with 6 knots. For each x value, we computed the median along with a 95% confidence interval for the posterior medians of f and then back-transformed in order to compare the estimated survival curve to its true counterpart.

The results are shown in Figure 1. For each of the two examples, our approach was successful in capturing the nonlinearities in the survival function. Note that in Study 1, a relatively simple regression function was specified,

resulting, for the same number of knots and sample size, in better precision than for Study 2.

[Figure 1 about here.]

6. Application to Snow petrels data

We illustrate the approach of the paper with data from a 40-year study on individually marked Snow petrels, nesting at Petrels Island, Terre Adélie, from 1963-2002. Two previous studies have showed that a large part of the variation in annual survival was explained by climatic covariates such as the extent of sea-ice and air temperature (Barbraud et al., 2000; Jenouvrier, Barbraud and Weimerskirch, *unpublished results*). Here, for illustration, we used only a subset of the whole dataset, from 1973-2002 ($I = 29$, 630 males and 640 females), and considered the Southern Oscillation Index (a covariate denoted by SOI) as a summary of the overall climate condition, with positive (respectively negative) values of the SOI corresponding to cold (respectively warmer) climatic conditions. While the NAO is a useful synthesis of climatic variables that might affect ecology in the Northern hemisphere (see Section 1), the SOI provides its counterpart for the Southern hemisphere (see Stenseth et al., 2003 for a general discussion). The SOI is available from the Climatic Research Unit (http://www.cru.uea.ac.uk/cru/data/soi.htm).

Preliminary analysis of goodness-of-fit of the CJS model identified lack of fit due to the presence of transients (146 males and 169 females were seen only once) (Pradel, Hines, Lebreton and Nichols, 1997) and trap-dependence (Pradel, 1993). The transients were removed, and trap-dependence was handled by considering different capture probabilities depending on whether a capture occurred or not at the previous sampling occasion.

We modeled the survival probability nonparametrically as a function of the SOI using P-splines. The effect of this covariate was additively differentiated according to the sex of individuals. We used linear splines ($P = 1$) but quadratic or even cubic splines could have been used instead, resulting mainly in a smoother estimated survival curve (Ruppert et al., 2003). We used $K = 6$ knots chosen so that the kth knot is the sample quantile corresponding to probability $k/(K + 1)$. Note that the covariate SOI was first standardized in order to avoid numerical instabilities and to improve MCMC mixing (Gilks and Roberts, 1996). We therefore considered the following model

$$\text{logit}(\phi_i^l) \quad = \quad \beta_0 + \gamma \text{SEX} + \beta_1 \text{SOI}_i + \sum_{k=1}^{6} b_k \left(\text{SOI}_i - \kappa_k\right)_+ + \varepsilon_i \qquad (9)$$

where ϕ_i^l is the survival probability over the interval $[t_i, t_{i+1}]$ for l = male (SEX = 0) or l = female (SEX = 1) and SOI_i denotes the SOI in year i, $i = 1, \ldots, I$. The random effects $\{b_k\}$ are independent as well as the $\{\varepsilon_i\}$.

Let us denote $\phi = (\phi_1^{female}, \ldots, \phi_{28}^{female}, \phi_1^{male}, \ldots, \phi_{28}^{male})^T$. Then, in matrix notation, Equation (9) can be expressed in the form of Equation (6) using

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \gamma & \beta_1 \end{pmatrix}^T$$

$$X = \begin{pmatrix} 1 & 1 & \text{SOI}_1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & \text{SOI}_{28} \\ 1 & 0 & \text{SOI}_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & \text{SOI}_{28} \end{pmatrix}$$

for the fixed effects and

$$\boldsymbol{b} = \begin{pmatrix} b_1 & \ldots & b_6 \end{pmatrix}^T$$

$$Z = \begin{pmatrix} (\mathrm{SOI}_1 - \kappa_1)_+ & \dots & (\mathrm{SOI}_1 - \kappa_6)_+ \\ \vdots & \vdots & \vdots \\ (\mathrm{SOI}_{28} - \kappa_1)_+ & \dots & (\mathrm{SOI}_{28} - \kappa_6)_+ \end{pmatrix}$$

for the random effects.

The model proposed here differs from the semiparametric approach presented before in that the sex parametric component acts at the individual level rather than on sampling occasions. The likelihood is therefore slightly modified consisting of the product of two sub-components, one for each sex, based on the product-multinomial structure of the m-array (e.g. Lebreton et al., 1992).

To completely specify the Bayesian nonparametric model, we need to provide prior distributions for all parameters. Specifically, we chose

$$\begin{aligned} [p_{i+1}] = Beta(A_p, B_p), \qquad & [\varepsilon_i] = N(0, \sigma^2_\varepsilon), \quad i = 1, \dots, I \\ [\beta_0], [\beta_1], [\gamma] &= N(0, \sigma^2_\beta), \\ [b_k] &= N(0, \sigma^2_b), \quad k = 1, \dots, K, \end{aligned}$$

where the parameter σ_b controls the degree of smoothing for the covariate. Following Brooks et al. (2000), we chose $A_p = B_p = 1$ which leads to a uniform distribution, while following Ruppert et al. (2003), σ^2_β was set to 10^6, and priors for hyperparameters were chosen as

$$\left[\sigma^2_b\right], \ \left[\sigma^2_\varepsilon\right] = \Gamma^{-1}(0.001, 0.001).$$

All priors were selected as sufficiently vague in order to induce little prior knowledge, but can be easily refined if required. We generated two chains of length 100,000, discarding the first 50,000 as burn-in. These simulations took approximatively 25 hours on a PC (512Mo RAM, 2.6GHz CPU). Convergence was assessed using the Gelman and Rubin statistic, also called the

potential scale reduction, which compares the within to the between variability of chains started at different and dispersed initial values (Gelman, 1996). We found that the Markov chains exhibit moderate autocorrelation but poor mixing regarding the parameters b_ks and βs. We thus tried low–rank thin–plate splines because in that case the posterior correlation of the parameters is generally smaller than for other bases. However, in our example, this only improved the mixing slightly, so that we decided to retain the truncated polynomial basis throughout, coupled with chains of adequate length to achieve convergence. According to our experience, inference based on P-splines within the Bayes framework may be sensitive to the choice of priors, especially regarding σ_b (see Crainiceanu et al., 2004a for a discussion of prior distributions for nonparametric P-spline regression). In order to check for the robustness of our results, we ran our model using different priors and in all cases there were only minimal changes.

We used the software WinBUGS (downloadable freely from http://www.mrc-bsu.cam.ac.uk/bugs/) by calling it from software R through the package R2WinBUGS (see R web site at http://r-project.org/ and Crainiceanu et al, 2004b for implementation examples of nonparametric Bayesian P-splines in WinBUGS). Priors and likelihood are specified with WinBUGS, while it appears more useful in practice to process data, set initial values, check for convergence and draw inference after the model is fitted using R. The codes used for fitting the model are available from the first author on request.

Posterior medians, standard deviations, and 95% credible intervals are given in Table 1.

[Table 1 about here.]

Because it does not contain 0, the posterior credible interval for parameter γ suggests that the sex of individuals affects the survival probability. As demonstrated by other studies (Jenouvrier, Barbraud and Weimerskirch, *unpublished results*), male petrels survive better than females, whatever the climatic conditions (see Figure 2).

Of particular interest, it appears that survival is nonlinearly related to the SOI covariate (Figure 2). When the SOI increases, survival first decreases and then stabilizes. From a biological point of view, lower values of the SOI may favor access to prey, whereas higher values may improve prey abundance (Loeb et al., 1997), resulting in the non-linearity found.

[Figure 2 about here.]

In order to know if the nonparametric part our model was needed, we compared the nonparametric model with the simple standard approach in which the SOI is just entered linearly on the logistic scale. From Figure 2, the linear curve (dotted line) differs clearly from the nonparametric curve (solid line), but the 95% credible interval (dashed lines) for the latter partly contains the former, which means that this difference is only marginal. This conclusion was supported by the DIC values (Spiegelhalter et al., 2002) and the credible intervals for the b_ks, which include zero. However the non-linearity has a biological explanation, and as we can see from Figure 2, in this example we require more years of data corresponding to large values of SOI in order to discriminate better between the two models.

Note that the mean encounter probabilities were 67% for males and 61% for females if a capture occurred at the previous occasion, and 62% for males

and 58% for females if not. This sex-dependent positive trap-effect is in agreement with a recent study on Snow petrels (Jenouvrier, Barbraud and Weimerskirch, *unpublished results*).

7. Discussion

This paper presents a Bayesian approach for nonparametric modeling of survival estimated using capture-recapture data, where smooth functions were modeled as penalized splines. Extensions such as additive and semiparametric models are straightforward within the unified framework based on the mixed model representation. In addition, due to the hierarchical structure of our Bayesian approach, the degree of smoothness is data-driven and controlled by the smoothing parameter estimated jointly with the unknown regression parameters.

The modelling of this paper does not include interactions between covariates. For example, an interaction between sex and a climatic covariate would involve considering different smooth functions for males and females (Coull, Ruppert and Wand, 2001). Following the suggestion of a referee, we considered this interaction for the real data but it did not appear to improve the fit. An interaction between two continuous covariates can be achieved by using bivariate smoothing (Ruppert et al., 2003). For example, it would be interesting to include an interaction between population density and climate in a model (Coulson et al., 2001), requiring an extension of the power truncated function basis to a tensor product basis (Green and Silverman, 1994).

In this paper we dealt with goodness-of-fit by first of all applying standard procedures to the CJS model, which identified transients, which were

excluded, and the presence of trap-dependence, which was included in the semiparametric model. Any further lack of fit was accommodated in part through the inclusion of the random effect terms in Equation (2), which are seen to be needed from the estimate of σ_ε in Table 1.

The work of this paper has wider application than just to the CJS model, e.g. in models with age-dependence of survival, including modelling senescence (e.g. Catchpole et al., 2004).

Acknowledgements

The authors would like to thank V. Grosbois and R. Pradel for very stimulating and helpful discussions. O. Gimenez's research was supported by a Marie-Curie Intra-European Fellowship within the Sixth European Community Framework Programme. B.J.T. Morgan was supported by a Leverhulme Fellowship.

Résumé

Les modèles de capture-recapture servent à estimer la survie d'une population sauvage, grâce à des données issues du marquage et du suivi dans le temps d'individus. Il est d'une importance toute particulière de pouvoir expliquer les variations de survie en fonction de variables judicieuses. Nous développons des modèles de régression nonparamétriques et semiparamétriques pour la probabilité de survie des modèles de capture-recapture. Nous nous plaçons dans un cadre Bayésien, et l'estimation des paramètres s'effectue grâce à des méthodes MCMC. Nous illustrons notre travail par l'étude de la survie de Pétrels des neiges comme une fonction non-linéaire d'une variable climatique,

en utilisant des données d'un suivi de 40 années concernant des individus nichant sur l'ile des Pétrels, en Terre Adélie.

References

Barbraud, C., Weimerskirch, H., Guinet, C. and Jouventin, P. (2000). Effect of sea-ice extent on adult survival of an antarctic top predator: the snow petrel pagodroma nivea. *Oecologia* **125**, 483–488.

Barry, S. C., Brooks, S. P., Catchpole, E. A. and Morgan, B. J. T. (2003). The analysis of ring-recovery data using random effects. *Biometrics* **59**, 54–65.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician* **47**, 69–100.

Brooks, S. P., Catchpole, E. A. and Morgan, B. J. T. (2000). Bayesian animal survival estimation. *Statistical Science* **15**, 357–376.

Brumback, B., Ruppert, D. and Wand, M. P. (1999). Comment on Variable selection and function estimation in additive nonparametric regression using data-based prior, by Shively, Kohn, and Wood. *Journal of the American Statistical Association* **94**, 794–797.

Burnham, K. (2002). Evaluation of some random effects methodology applicable to bird ringing data. *Journal of Applied Statistics* **29**, 245–264.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *American Statistician* **46**, 167–174.

Catchpole, E. A., Fan, Y., Morgan, B. J. T., Clutton-Brock, T. and Coulson, T. (2004). Sexual dimorphism, survival and dispersal in red deer. *Journal of Agricultural, Biological and Environmental Statistics* **9**, 1–26.

Chavez-Demoulin, V. (1999). Bayesian inference for small-sample capture-recapture data. *Biometrics* **55**, 727–731.

Choquet, R., Reboulet, A. M., Pradel, R., Gimenez, O. and Lebreton, J.-D. (2005). M-SURGE : new software specially designed for multistate capture-recapture models. *Animal Biodiversity and Conservation* **27**, 207–215.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51**, 429–438.

Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics* **57**, 539–545.

Coulson, T., Catchpole, E. A., Albon, S. D., Morgan, B. J. T., Pemberton, J. M., Clutton-Brock, T. H., Crawley, M. J. and Grenfell, B. T. (2001). Age, sex, density, winter weather, and population crashes in Soay sheep. *Science* **292**, 1528–1531.

Crainiceanu, C. M., Ruppert, D. and Carroll, R. (2004a). Spatially Adaptive Bayesian P-Splines with Heteroscedastic Errors. *submitted to Journal of Computational and Graphical Statistics* .

Crainiceanu, C. M., Ruppert, D. and Wand, M. (2004b). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *submitted to Journal of Statistical Software* .

Gelman, A. (1996). Inference and monitoring convergence. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice.*, pages 131–143. Chapman and Hall.

Gilks, W. R. (1996). Full conditional distributions. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in*

practice., pages 75–86. Chapman and Hall.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving mcmc. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice.*, pages 89–114. Chapman and Hall.

Green, P. and Silverman, B. (1994). *Nonparametric regression and Generalized Linear Models.* Chapman and Hall, New York. USA.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Johnson, D. S. and Hoeting, J. A. (2003). Autoregressive models for capture-recapture data: A Bayesian approach. *Biometrics* **59**, 341–350.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52**, 225–247.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

Lebreton, J.-D., Burnham, K. P., Clobert, J. and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monographs* **62**, 67–118.

Lima, M., Merritt, J. and Bozinovic, F. (2002). Numerical fluctuations in the northern short-tailed shrew: evidence of non-linear feedback signatures on population dynamics and demography. *Journal of Animal Ecology* **71**, 159–172.

Loeb, V., Siegel, V., Holm-Hansen, O., Hewitt, R., Fraser, W., Trivelpiece, W. and Trivelpiece, S. (1997). Effects of sea-ice extent and krill or salp dominance on the Antarctic food web. *Nature* **387**, 897–900.

Loison, A., Festa-Bianchet, M., Gaillard, J. M., Jorgenson, J. T. and Jul-

lien, J. M. (1999). Age-specific survival in five populations of ungulates: Evidence of senescence. *Ecology* **80**, 2539–2554.

Mysterud, A., C., S. N., Yoccoz, N. G., Langvatn, R. and Steinheim, G. (2001). Nonlinear effects of large-scale climatic variability on wild and domestic herbivores. *Nature* **410**, 1096–1099.

Ngo, L. and Wand, M. (2004). Smoothing with mixed model software. *Journal of Statistical Software* **9**.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: an overview. *Journal of Applied Statistics* **29**, 85–102.

Pradel, R. (1993). Flexibility in survival analysis from recapture data: Handling trap-dependence. In Lebreton, J.-D. and North, P., editors, *Marked Individuals in the Study of Bird Population*, pages 11–36. Basel: Birkhauser.

Pradel, R., Hines, J. E., Lebreton, J.-D. and Nichols, J. D. (1997). Capture-recapture survival models taking account of transients. *Biometrics* **53**, 60–72.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D., Wand, M. P. and Carroll, R. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

Schwarz, C. J. and Seber, G. A. (1999). Estimating animal abundance: review III. *Statistical Science* **14**, 427–56.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika* **52**, 249–259.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). WinBUGS User

Manual. Version 1.4 (http://www.mrc-bsu.cam.ac.uk/bugs.). Technical report, Medical Research Council Biostatistics Unit. Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Lind, A. (2002). Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

Stenseth, N. C. and Mysterud, A. (2002). Climate, changing phenology, and other life history traits: nonlinearity and match-mismatch to the environment. *PNAS* **99**, 13379–13381.

Stenseth, N. C., Ottersen, G., Hurrell, J. W., Mysterud, A., Lima, M., Chan, K.-S., Yoccoz, N. G. and Ådlandsvik, B. (2003). Studying climate effects on ecology through the use of climate indices: the North Atlantic Oscillation, El Niño Southern Oscillation and beyond. *Proceedings of the Royal Society of London Series B - Biological Sciences* **270**, 2087–2096.

White, G. C. and Burnham, K. P. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**, 120–39.

Williams, B. K., Nichols, J. D. and Conroy, M. J. (2002). *Analysis and Management of Animal Populations.* Academic Press, San Diego, California.

Wintrebert, C., Zwinderman, A. H., Cam, E., Pradel, R. and Van Houwelingen, J. C. (2005). Joint modeling of breeding and survival of Rissa tridactyla using frailty models. *Ecological Modelling* **181**, 203–213.

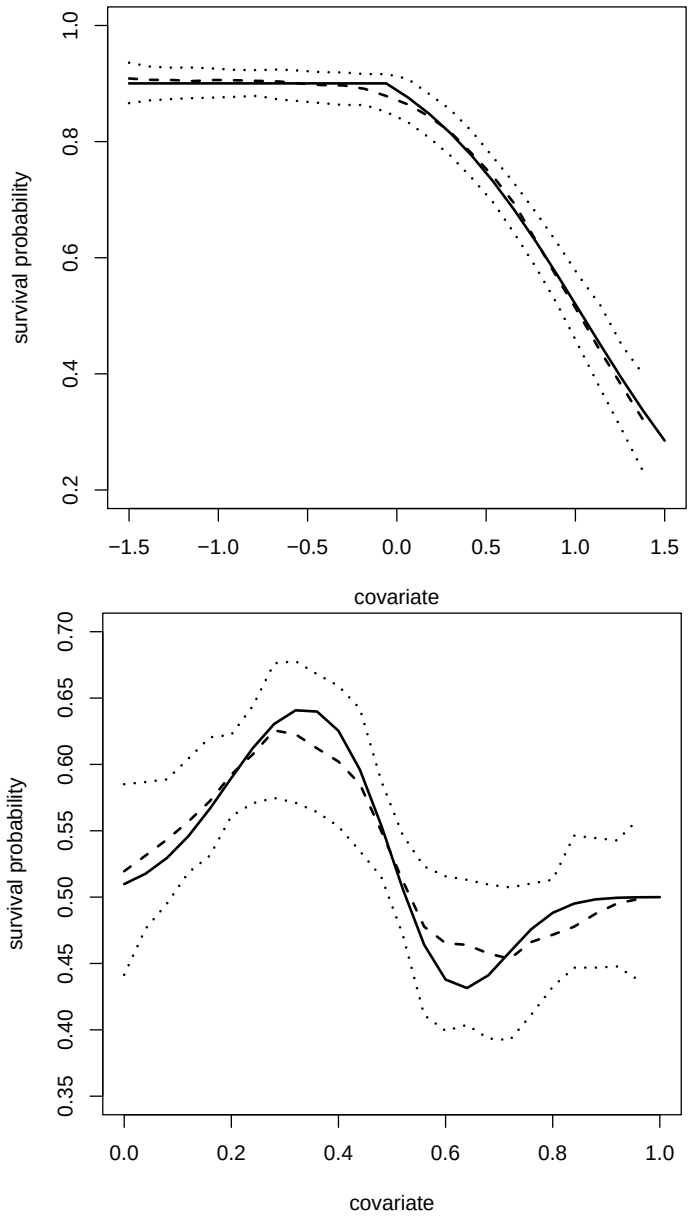

Figure 1. Performance of the non-parametric approach for estimating non-linearities in the survival probability (top: Study 1, and bottom: Study 2; see text for details). For both scenarios, 50 simulated capture-recapture data sets were used. The solid line is the true regression function, the dashed line is the median of the 50 estimated posterior medians and the dotted lines indicate the associated 95% confidence interval.

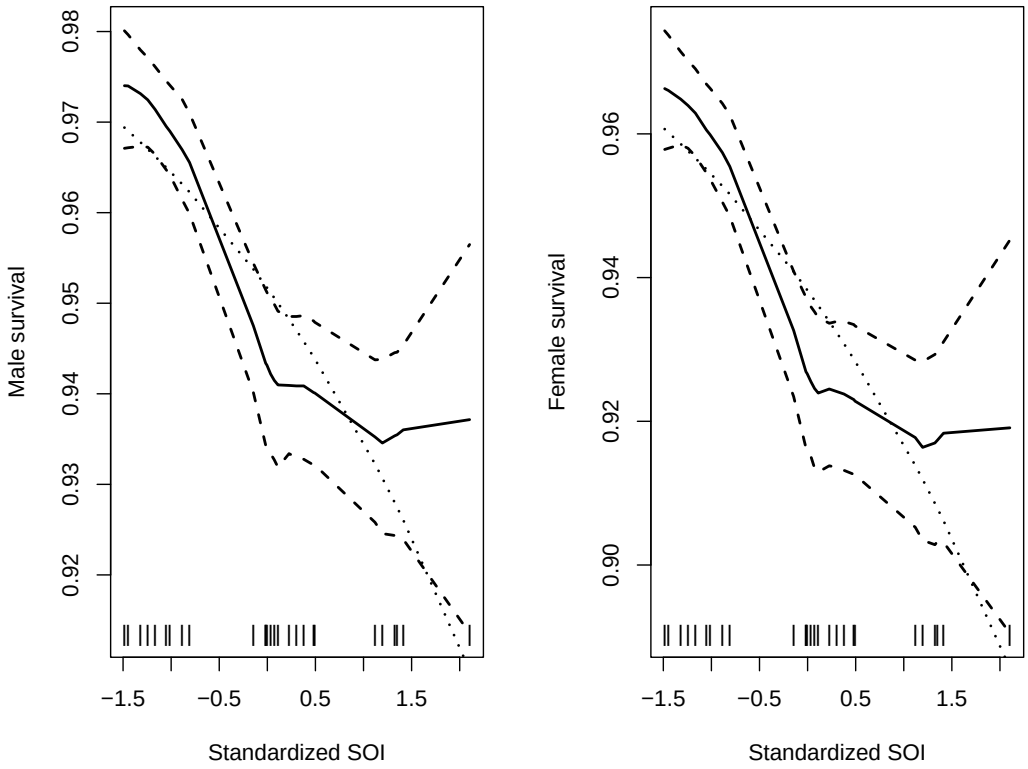

Figure 2. Annual variations in survival of male (left) and female (right) Snow petrels, as a function of the standardized Southern Oscillation Index (SOI) using the semiparametric model (Equation (9)). Note that the two vertical scales are different. Medians (solid line) with 95% pointwise credible intervals (dashed lines) are shown, along with the estimated linear effect (dotted line) on the logistic scale and the standardized covariate values (vertical lines).

Table 1

Posterior medians, standard deviations, and 95% credible intervals for the semiparametric model applied to the Snow petrels data set (see Equation (9)).

Parameter	Median	St. Dev.	95% Cred. Int.
β_0	2.93	0.40	[1.93;3.55]
γ	-0.26	0.10	[-0.45;-0.06]
β_1	-0.47	0.38	[-1.39;0.07]
σ_b	0.23	0.36	[0.03;1.15]
σ_ε	0.56	0.14	[0.35;0.91]
b_1	0.01	0.23	[-0.52;0.51]
b_2	0.00	0.33	[-0.83;0.62]
b_3	0.08	0.35	[-0.29;1.01]
b_4	0.08	0.42	[-0.39;1.38]
b_5	0.02	0.45	[-1.43;0.75]
b_6	0.03	0.44	[-0.50;1.23]