12-8-2004

Semiparametric Regression in Capture-Recapture Modelling

O. Gimenez
Institute of Mathematics, Statistics & Actuarial Science, University of Kent, United Kingdom & Centre d'Ecologie Fonctionnelle et Evolutive-CNRS, France, olivier@mcs.st-and.ac.uk

C. Barbraud
Centre d'Etudes Biologiques de Chize, France, barbraud@cebc.cnrs.fr

Ciprian M. Crainiceanu
Johns Hokins Bloomberg School of Public Health, Department of Biostatistics, ccrainic@jhsph.edu

S. Jenouvrier
Centre d'Etudes Biologiques de Chize, France, jenouvrier@cebc.cnrs.fr

B.T. Morgan
Institute of Mathematics, Statistics & Actuarial Science, University of Kent, United Kingdom

Semiparametric regression in capture-recapture modelling

O. Gimenez,[1,2,*] C. Barbraud,[3] C. Crainiceanu,[4] S. Jenouvrier[3] 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]Centre d'Etudes Biologiques de Chizé
CNRS UPR 1934
79360 Villiers en Bois - France

[4]Johns Hopkins University
615 N. Wolfe St. E3037
Baltimore, MD 21205 - USA

December 8, 2004

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

[*] *email:* gimenez@cefe.cnrs-mop.fr

els for estimating 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 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 therefore an important tool for estimating demographic parameters that govern populations.

In the last forty years, a challenging research topic has been the estimation of survival, and when possible, to explain variations using auxiliary variables like e.g. time, age of animal or relevant covariates like temperature or rainfall. Most 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., 2004). The Bayesian approach has been proposed as an alternative (see Brooks et al., 2000 for a review).

To model survival probability, the analysis 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 incorporated, and here we will focus on environmental covariates that vary over sampling occasions but remain constant over individuals, as defined by Pollock (2002). Individual covariates require a separate treatment, and this point will be discussed in the last section. Most frequently, covariates are related to survival by a linear or a quadratic function, on the logit scale. However, this may be unrealistic. For example, it has been shown that using 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). Other covariates that may affect population dynamics in a non-linear way include population density through density-dependence (see e.g. Sinclair, 1989) or age, through senescence defined as a reduction in survival among old individuals (Loison et al., 1999; Catchpole et al., 2004). In many of these examples a nonparametric alternative avoids strong parametric assumptions and could suggest new, scientifically relevant, parametric models.

In this paper we extend the traditional GLM framework using Generalized Additive Models (GAMs) ideas popularized by Hastie and Tibshirani (1990). 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 we show how our approach particularly benefits from the use of Bayesian graphical modeling through Gibbs sampling. Section 5 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 discusses the limits and 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 be actually 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 will be referred to as the CJS model. Schwarz and Seber (1999) and Williams et al. (2002) give reviews of the CJS model and its applications. 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 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

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

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).

Factors possibly affecting both survival and capture probabilities, such as sex or location can be easily handled by considering an m-array for each value of the factor in formula (1) (e.g. Lebreton et al., 1992). Note that age classes can be accommodated in a similar way (Brownie and Robson, 1983).

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 handled 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 is 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. For fitting we used a penalty approach inspired by smoothing splines (Green and Silverman, 1994). To ensure enough flexibility a fixed number of knots is chosen. Following Ruppert (2002), we considered $K = \min\{\frac{1}{4}I, 35\}$ and let κ_k be "equally-spaced sample quantiles" ie the sample quantile of the x_i's corresponding to probabilities $k/(K+1)$. Other choices are possible like 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

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

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. 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 $\mathbf{Z}_i = \{(x_i - \kappa_1)_+^P, \ldots, (x_i - \kappa_K)_+^P\}^T$. Consider the vector $\beta = (\beta_0, \ldots, \beta_P)^T$ as fixed parameters and the vector $\mathbf{b} = (b_1, \ldots, b_K)^T$ as a set of random parameters with $E(\mathbf{b}) = 0$ and $cov(\mathbf{b}) = \sigma_b^2\mathbf{I}_K$. If $\mathbf{b}$ and ε are independent, then an equivalent model representation of the P-spline model in the form of a GLMM is

$$\operatorname{logit}(\phi) \;=\; \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon, \quad \operatorname{cov}\begin{pmatrix} \mathbf{b} \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 \mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I}_I \end{pmatrix} \tag{5}$$

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

Note that the connection between the P-spline model and the mixed model of Equation (5) 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 will consider only one parametric categorical component s with one non-parametric component smoothing a continuous predictor x by linear P-splines. We wish to let the relation between $\operatorname{logit}(\phi_i)$ and x_i vary differently but in parallel according to the variable s_i taking discrete values, i.e.

$$\begin{aligned} \operatorname{logit}(\phi_i) \;=\; & \beta_0 + \gamma s_i + \beta_1 x_i + \\ & \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \varepsilon_i, \quad i = 1, \ldots, I. \end{aligned} \tag{6}$$

Once again, the GLMM representation can 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 (5) can still be used to describe the semiparametric regression just defined in Equation (6) (Ruppert et al., 2003).

4. Bayesian inference

In this section, we will 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 5).

4.1 *Parameter estimation*

The frequentist approach would require maximising the likelihood, which is obtained by integrating the distribution $[\mathbf{R}, \mathbf{m}|\phi, \mathbf{p}]$ 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). We expressed our models in the form of Directed Acyclic Graphs, that are analysed by a Bayesian approach through Gibbs sampling.

4.2 *Bayesian graphical modeling*

In Figure 1, structural relations between the quantities called nodes that form our model are represented by a Directed Acyclic Graph (DAG, see Spiegelhalter, 1998).

[Figure 1 about here.]

Invoking conditional independence properties, the DAG representation leads us to a recursive factorization of the posterior distribution as:

$$\begin{aligned}&[\beta, \mathbf{b}, \varepsilon, \sigma_b^2, \sigma_\varepsilon^2, \mathbf{p}|\mathbf{R}, \mathbf{m}] \\ \propto\ & [\mathbf{R}, \mathbf{m}|\phi, \mathbf{p}][\phi|\beta, \mathbf{b}, \varepsilon][\beta][\mathbf{b}|\sigma_b^2][\varepsilon|\sigma_\varepsilon^2][\sigma_b^2][\sigma_\varepsilon^2][\mathbf{p}]. \qquad (7)\end{aligned}$$

Even if one is only interested in the marginal posterior distribution of some parameters, high-dimensional integrations have to be carried out. In general, such complex integrals are intractable analytically and we will make use of MCMC methods which provide powerful computer-intensive methods for making approximations (e.g. Brooks, 1998). Because of its close relationships with Bayesian graphical modeling, we will make use of Gibbs sampling (Casella and George, 1992). When Gibbs sampling is used for estimating capture-recapture model parameters, 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. In place however, more elaborate algorithms are needed such as adaptive rejection sampling or Metropolis-within-Gibbs sampling (see Gilks, 1996 for a review). We will therefore use software WinBUGS (Spiegelhalter et al., 2003), which performs the latter.

5. 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 we used the whole dataset ($I = 39$, 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.

Briefly speaking, while the NAO is a useful synthesis of climatic variables that might affect ecology in the Northern hemisphere, 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).

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 survival curve (Ruppert et al., 2003). We used $K = 10$ 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

$$\text{logit}(\phi_i^l) \quad = \quad \beta_0 + \gamma\text{SEX} + \beta_1\text{SOI}_i + \sum_{k=1}^{10} b_k \left(\text{SOI}_i - \kappa_k\right)_+ + \varepsilon_i \qquad (8)$$

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 $\phi = (\phi_1^{female}, \ldots, \phi_{39}^{female}, \phi_1^{male}, \ldots, \phi_{39}^{male})^T$. Then, in matrix notation, Equation (8) can be expressed in the form of Equation (5) using

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

$$X = \begin{pmatrix} 1 & 1 & \mathrm{SOI}_1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & \mathrm{SOI}_{39} \\ 1 & 0 & \mathrm{SOI}_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & \mathrm{SOI}_{39} \end{pmatrix}$$

for the fixed effects and

$$\boldsymbol{b} = \begin{pmatrix} b_1 & \dots & b_{10} \end{pmatrix}^T$$

$$Z = \begin{pmatrix} (\mathrm{SOI}_1 - \kappa_1)_+ & \dots & (\mathrm{SOI}_1 - \kappa_{10})_+ \\ \vdots & \vdots & \vdots \\ (\mathrm{SOI}_{39} - \kappa_1)_+ & \dots & (\mathrm{SOI}_{39} - \kappa_{10})_+ \end{pmatrix}$$

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. Note that for illustration, we considered a constant encounter probability p. According to other studies on Snow petrels, complex patterns in encounter probabilities are likely to occur, including sex effects, and this would deserve further attention (Barbraud et al., 2000; Jenouvrier, Barbraud and Weimerskirch, *unpublished results*). We do not anticipate that allowing p to vary would affect conclusions relating to survival.

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

$$[p] = Beta(A_p, B_p), \qquad [\varepsilon_i] = N(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, I$$

$$\begin{aligned} [\beta_0], [\beta_1], [\gamma] &= N(0, \sigma_\beta^2), \\ [b_k] &= N(0, \sigma_b^2), \quad k = 1, \ldots, K, \end{aligned}$$

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_β^2 was set to 10^6 and priors for hyperparameters were chosen as

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

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 1100000, discarding the first 100000 as burn-in. These simulations took approximatively 100 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 good mixing and moderate autocorrelation. 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 clear that survival is nonlinearly related to the SOI covariate (Figure 2). When the SOI increases, survival first decreases and then clearly increases, being maximal for higher SOI values. 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.]

Note that the encounter probability was around 53%, in agreement with a recent study on Snow petrels (Jenouvrier, Barbraud and Weimerskirch, *unpublished results*).

6. 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. There are several directions in which the methodology used here could be extended in future research.

The simplest way of looking at patterns of variation in survival would be to consider a nonparametric trend with time. Due to their importance in a context of global warming, we focused on the effect of a climate indicator on survival but other covariates are possible such as e.g. age or density (see Section 1). Individual covariates may affect survival as well. For example, survival of small birds is often believed to increase with increasing body mass (e.g. Covas et al., 2002). In such a case, sufficient statistics like the m-array no longer exist and we would have to evaluate the probability of each individual capture history. This is the subject of ongoing work. Individual covariates changing over time can exhibit missing values when individuals are undetected. In this regard, the Bayesian approach can be easily extended by specifying a probability distribution for the covariate (Pollock, 2002).

Of course, we dealt with survival only, but other demographic parameters can be considered. A Bayesian approach has already been used by Dupuis (1995) and King and Brooks (2003) for example to estimate dispersion parameters and by Brooks et al. (2004) to estimate growth rate, so that nonparametric and semiparametric modeling of those demographic parameters could be easily performed.

Here, we did not consider interactions. In our example, an interaction between sex and the climatic covariate would have consisted in considering different smooth functions for male and female individuals. To implement an interaction between two continuous covariates can be achieved using bivariate smoothing (Ruppert et al., 2003) but would depend on the nature of the covariates involved. For example, it would be interesting to include an interaction between 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). However, because of numerical problems, we expect radial smoothers to be a better choice (Crainiceanu et al, 2004b). In the study of spatial patterns in demographic parameters, geographical covariates would be worth considering, for example latitude and longitude coordinates. In such cases, the close relationships between splines and kriging (Cressie, 1993) would be useful in extending our approach (Ruppert et al., 2003). The possibility to include geographical and non-geographical covariates within the framework we developed here is the object of ongoing work.

Model selection is another important research topic. King and Brooks (2002a; 2002b) successfully made used of Reversible Jump MCMC (Green, 1995) in order to select among a large set of models for estimating survival and dispersion parameters. However, this approach is not yet implemented in standard software and a particular treatment is therefore needed for each model considered. Another option is the deviance information criterion (DIC) introduced by (Spiegelhalter et al., 2002). The DIC is a Bayesian analogue to the Akaïke information criterion penalizing the fit of a model measured by the deviance with the complexity of a model represented by its number of parameters. This criterion is available in WinBUGS. Model selection provides a test for nonlinearity by contrasting the DIC value for a model that fits the relevant covariates nonparametrically with the DIC value for the corresponding model that fits the terms linearly. Barry et al. (2003) recommend caution in using the DIC for comparing models. However, for the Snow petrels example, the semiparametric model of Equation (8) has a DIC value of 17030.6, while for the model where the covariate is entered linearly, we found a DIC value of 18349.3; this very large difference suggests that nonlinearities were needed to represent variation in survival.

Goodness-of-fit has not been considered here, but Bayesian p-values may be obtained, as explained in Brooks et al. (2000).

We made the implicit assumption that the covariates were measured without error. However, in trying to exhibit density-dependence phenomena for example, covariates such as population sizes are often subject to measurement error. Among several methods for dealing with imperfect measurements in regression models (e.g. Carroll et al., 1995), we regard the approach proposed by Carroll et al. (1999) and its recent Bayesian extension (Berry et al., 2002) as the most promising, since the regression function is modeled with P-splines, while the covariate is treated as a latent random variable and integrated out using MCMC methods.

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.

Berry, S., Carroll, R. and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* **97**, 160–169.

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.

Brooks, S. P., King, R. and Morgan, B. J. T. (2004). A Bayesian approach to combining animal abundance and demographic data. *Animal Biodiversity and Conservation* **27**.

Brownie, C. and Robson, D. (1983). Estimation of time-specific survival rates from tag- resighting samples: a generalization of the jolly-seber model. *Biometrics* **39**, 437–453.

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.

Carroll, R., Ruppert, D. and Stefanski, L. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, New York. USA.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.

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, N. (1999). Bayesian inference for small-sample capture-recapture data. *Biometrics* **55**, 245–264.

Choquet, R., Reboulet, A. M., Pradel, R., Gimenez, O. and Lebreton, J.-D. (2004). M-SURGE : new software specially designed for multistate capture-recapture models. *Animal Biodiversity and Conservation* **27**.

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

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.

Covas, R., Brown, C. R., Anderson, M. D. and Brown, M. B. (2002). Stabilizing selection on body mass in the sociable weaver philetairus socius. *Proceedings of the Royal Society of London Series B - Biological Sciences* **269**, 1905–1909.

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* .

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York. USA.

Dupuis, J. (1995). Bayesian estimation of movement and survival probabilites from capture-recapture data. *Biometrika* **82**, 761–772.

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. (1995). Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika* **82**, 711–732.

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.

King, R. and Brooks, S. P. (2002a). Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika* **89**, 785–806.

King, R. and Brooks, S. P. (2002b). Model selection for integrated recovery/recapture data. *Biometrics* **58**, 841–851.

King, R. and Brooks, S. P. (2003). Survival and spatial fidelity of mouflons: The effect of location, age, and sex. *Journal of Agricultural Biological and Environmental Statistics* **8**, 486–513.

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.

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 fodd web. *Nature* **387**, 897–900.

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

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.

Sinclair, A. R. E. (1989). Population regulation in animals. In Cherrett, J. M., editor, *Ecological concepts: the contribution of ecology to an understanding of the natural world.*, pages 197–241. Blackwell Scientific Publishers.

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. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Appl. Statist.* **47**, 115–133.

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.

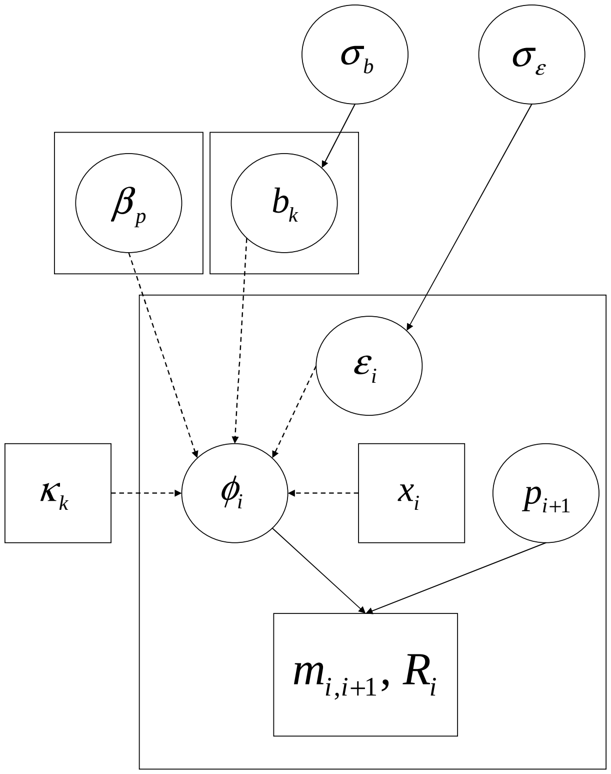

Figure 1. Graphical model for the nonparametric survival model (see Equation (3)): the large rectangle represents repetition over sampling occasions; variables are shown within circles; constants are shown within squares; stochastic dependencies are denoted by full arrows whereas logical dependencies are denoted by broken arrows. Here, $i = 1, \ldots, I$, $p = 0, \ldots, P$ and $k = 1, \ldots, K$.

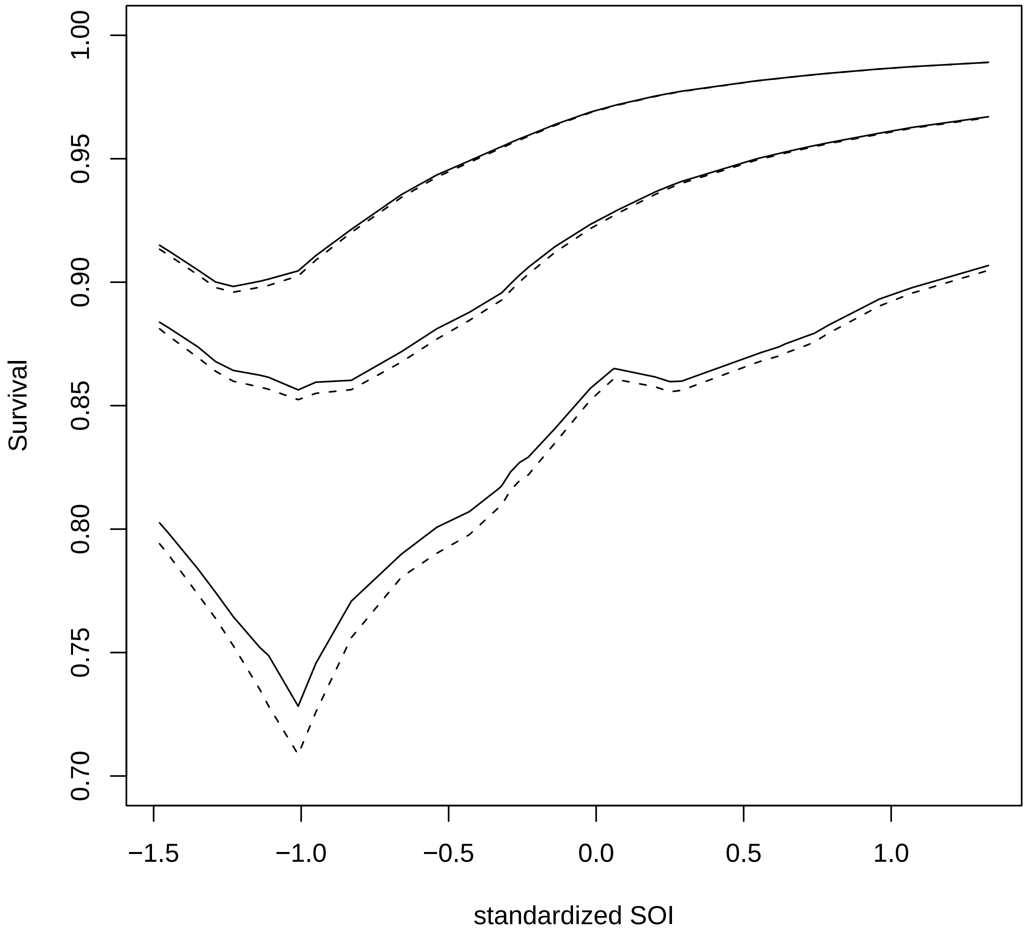

Figure 2. Annual variations in survival of male (solid line) and female (dashed line) Snow petrels, as a function of the standardized Southern Oscillation Index (SOI) using the semiparametric model (Equation (8)). Medians with 95% pointwise credible intervals are shown.

Table 1

Posterior medians, standard deviations, and 95% credible intervals for the semiparametric model applied to the Snow petrels data set (see Equation (8)).

Parameter	Median	St. Dev.	95% Cred. Int.
β_0	0.22	2.66	[-4.79;4.01]
γ	-0.22	0.09	[-0.40;-0.05]
β_1	-4.36	2.07	[-7.64;-1.85]
σ_b	3.76	1.66	[1.33;7.26]
σ_ε	8.85	1.38	[6.46;11.75]
p	0.53	0.01	[0.52;0.55]