Incorporating Covariates in the Analysis of Capture-re-encounter Data

Brown, Daniel, McCrea, Rachel and Morgan, Byron
University of Kent, School of Mathematics, Statistics and Actuarial Science
Canterbury, Kent, CT2 7NF, U.K.
E-mail: dib3@kent.ac.uk, R.S.McCrea@kent.ac.uk, B.J.T.Morgan@kent.ac.uk

INTRODUCTION

Methods of analysis of capture-re-encounter data involve fitting probability models to data collected from animals that have been previously given unique markings. Re-encounters can correspond to finding animals dead (recoveries), or capturing/resighting them alive (recaptures). For illustration in this paper we shall consider both the case of finding animals dead and also that of finding animals alive, at various times following marking. The model parameters are appropriate probabilities of annual survival, as well as probabilities of reporting/recapture of dead/live animals, respectively. An illustrative recovery data set is provided in Table 1, which is a subset of one of the data sets analysed in this paper. The full data set is available at www.tibs.org. The data correspond to birds being ringed throughout Britain, and later reported dead to the address on the rings.

Table 1 Recovery data on British grey herons, Ardea cinerea, ringed between 1987 and 1997. Data provided by the British Trust for Ornithology

	Year of recovery											
Year of ringing	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	Never recovered
1987	20	6	3	3	3	1	0	0	1	1	1	539
1988		11	8	5	2	2	0	1	1	0	1	467
1989			18	12	1	4	2	1	3	0	2	484
1990				28	5	2	3	0	2	0	0	556
1991					10	6	3	0	1	1	0	471
1992						21	2	0	0	0	1	541
1993							26	5	2	1	1	537
1994								20	6	3	1	546
1995									23	5	2	653
1996										16	0	494
1997											18	700

A range of models results from different assumptions regarding time- and age-variation of parameters. Simplifications may result from appropriate regressions on time-varying covariates, and the first instance of this being done is to be found in the paper by North and Morgan (1979), for a subset of the heron data illustrated in Table 1. Logistic regression for survival probabilities was used, and this is now often adopted for survival, recapture and reporting probabilities. More recently Gimenez et al (2009) have presented a more flexible approach, based on P-splines. The incorporation of covariates in models for capture-re-encounter data is now widespread (see Morgan,2006), and often complex. For instance, complexity can arise if there are several age-classes for survival, with different covariate regressions in each (see Catchpole et al, 2000). The area is one of current research; see for example Catchpole et al (2008) and Bonner et al (2010), who deal with missing covariate information. In this paper we provide a brief outline of new approaches to three important issues that are still to be resolved. The research is based in part on the thesis of Brown (2010), and will be described fully in papers currently being prepared for journal publication.

USING LOCAL WEATHER COVARIATES

The (global) covariate used by North and Morgan (1979) was the number of days below freezing in each year in Central England. However dead herons are recovered and reported widely, and it is natural to consider whether there are local weather alternatives. The approach to this adopted by Brown (2010) was to use weather data from a range of weather stations, available from the internet, and to smooth these using thin-plate splines. It was then possible to interpolate from the spline surface to obtain a measure of the weather for the location at which dead birds were found. The weather-station data are irregularly spaced, and it is possible to fit a thin-plate spline surface, as we know the longitude and latitude of each station. The fitting of the thin-plate splines is achieved using the contributed R package "fields"[1].

A thin-plate spline surface results from minimising the residual sum of squares as in ordinary least squares, subject to a constraint which governs the smoothness of the fitted function. A penalised residual sum of squares results, which is

$$R(g) = \sum_i \{Y_i - g(z_i)\}^2 + \eta J(g)$$

where z_i are points in two-dimensional space, Y_i is the temperature and g is a suitable smooth function such that $g(z_i) = g_i$ for $i = 1, ..., n$ where g_i is the temperature at the point z_i.

In the two-dimensional case, as here, the smoothness is governed by a roughness penalty which is

$$J(g) = \int\int \left\{ \left(\frac{\partial^2 g}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 g}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 g}{\partial y^2}\right)^2 \right\} dxdy.$$

and a smoothing parameter η. The thin-plate splines method finds the smoothing parameter, η, using generalised cross validation. In the cases where η cannot be found, which are rare, we let $\eta = 0.3$ as it represents a reasonable compromise between over-fitting and under-fitting of the surface to the data.

In an application on Blackbirds, Turdus merula, data were obtained from 44 weather stations throughout Germany, combined with climate indices such as the North Atlantic Oscillation, the NAO. The approach worked well, with spring and summer measures affecting the survival of young birds. Spring and summer temperature logistic regression coefficients were, respectively, 0.36(0.096) and -0.40(0.103), and for Spring rain, the value was -0.14(0.060). These values make biological sense.

As a simulation check of the approach, we present in Table 2 a comparison of the performance of models fitted assuming no covariates, local covariates and global covariates. We can see that it is important to include weather covariates when they are present in the simulations, that the thin-plate spline method performs the best, and the improvement, compared to the use of a global variable, becomes more marked the larger the size of the simulation grid.

Table 2 Δ AIC values in a simulation study to compare the use of local weather covariates smoothed with thin-plate splines, global weather covariates, and no covariates, when mark-recovery data were simulated using local weather covariates. Here r denotes the size of the weather grid.

r	0.1	0.2	0.3	0.4	0.5
local	0	0	0	0	0
global	10.7	43.9	89.1	165.7	229.8
none	312.9	309.3	380.3	408.9	507.7

[1] available at http://www.r-project.org

VARIABLE SELECTION USING THE LASSO

Catchpole et al (1999) undertake variable selection using a method based on the use of score tests. King et al (2006) use reversible-jump MCMC. As an alternative, Brown (2010) considered the merits of the lasso. This may be done in one of two ways. One is by obtaining survival parameter estimates for each year, following model-fitting by maximum likelihood, and then applying a standard lasso approach to the resulting parameter estimates, treating them as the response variables. This is faster and simpler than the alternative approach in which the lasso becomes an integral feature of a mark re-encounter analysis. The two approaches have resulted in similar performance when applied to real examples, by Brown, 2010. For example, for mark-recapture data spanning 16 years on white storks, Ciconia ciconia, with 10 weather stations from Germany, there are potentially 1024 alternative models to consider. Here the lasso approaches readily identified the information from just one weather station as the sole covariate to be included in the model for annual survival. This conclusion agrees with more complex alternative Bayesian approaches to the data.

MODELLING CONDITIONAL DATA

Quite often cohort sizes are either unknown or unreliable; we note that the analysis of North and Morgan (1979) did not involve cohort numbers, which were unavailable at that time. Quite often also, recent years have seen a decline in the reporting probability; see Baillie and Green (1987) and Mazetta (2010). The widely-used computer package for the analysis of capture-re-encounter data, MARK (White and Burnham), only allows a constant reporting probability (model denoted Constant in the results below) for the conditional analysis of mark-recovery data which result when cohort numbers are not used in the analysis. Robinson et al (2007) state that this is the only possibility for conditional analysis, however it has been shown by Cole and Morgan (2010) that models with appropriate time-regressions of the reporting probability may be fitted to mark-recovery data in conditional analyses. Here we propose a scaled-logistic model, in which

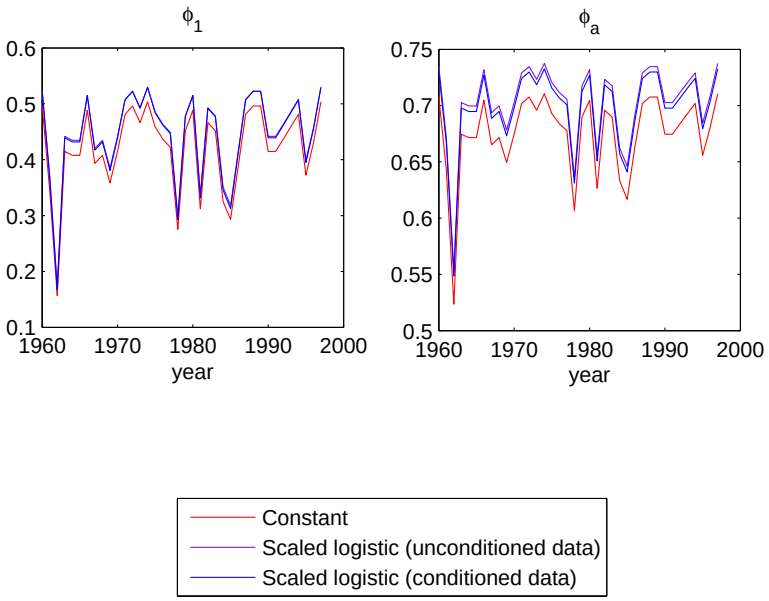
$$\lambda_t = \kappa, \quad t < \tau$$

$$\lambda_t = \frac{2\kappa}{1 + e^{\lambda_\beta(\tau - t)}}, \quad t > \tau$$

where τ is a time to be estimated, indicating the start of a decline in reporting probability. We note that the parameter κ will not enter the conditional analysis due to cancelation.

For a model with two annual survival probabilities, ϕ_1 and ϕ_a, corresponding to first-year and older birds respectively, each logistically regressed on a weather covariate denoting the number of days below freezing in Central England (see Besbeas et al, 2002), we obtain the results shown in Figure 1 when models are fitted to the heron data. For comparison we also provide the results using the cohort numbers (denoted Scaled logistic (unconditional data)). We can see the very good agreement between the analyses with and without using the cohort numbers, and the appreciable bias that results from using the model with assumed constant reporting probability.

Figure 1. A comparison of survival estimates for the heron data for three models described in the text



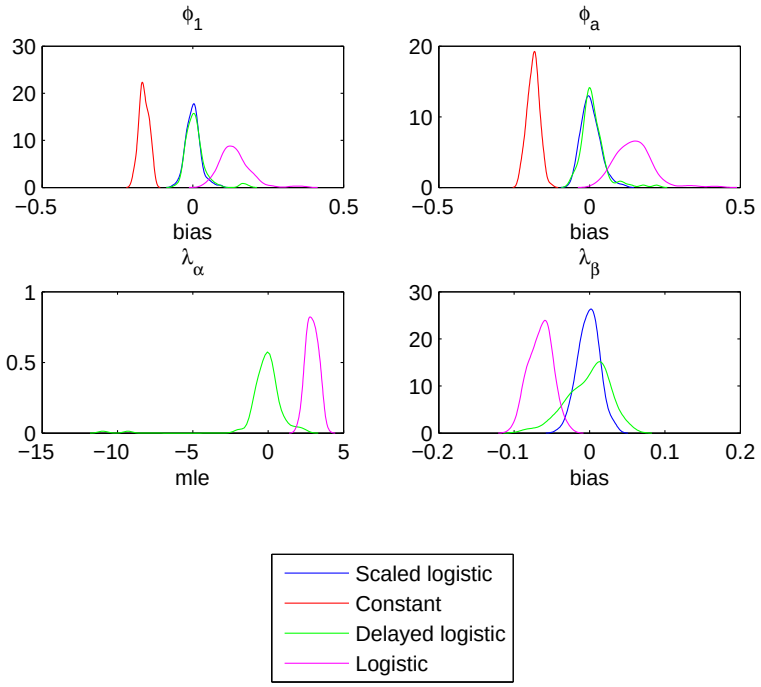
We now present in Figure 2 a simulation study of the bias that can result from using an incorrect model for the reporting probability in conditional analyses only. For comparison, we also include a model in which the reporting probability has a standard logistic regression on time (denoted logistic) and also a model which has the form

$$\lambda_t = \frac{1}{1+e^{\lambda_\alpha t}} \quad t < \tau,$$

$$\lambda_t = \frac{1}{1+e^{\lambda_\alpha + \lambda_\beta (t-\tau)}} \quad t > \tau,$$

which we denote as the Delayed logistic model.
Both the scaled logistic and delayed logistic models produce unbiased estimate of survival probability. As expected, the constant model, which fails to account for any time variation in reporting probability, results in underestimates of survival. Further, we observe that the logistic model, which ignores the period of constant recovery before a decline commences, results in overestimates of the survival probability. It is therefore interesting that a naive approach to modelling the decline in reporting probability over time, by the logistic model, proves to be inadequate. We also note the superior performance of the scaled logistic model compared to the delayed logistic model, with regard to the bias of the slope parameter λ_β.

Figure 2. A simulation study to compare the performance of four alternative models for the reporting probability; mle denotes maximum-likelihood estimate for λ_α, which is only available for two models.



DISCUSSION AND FURTHER WORK

We have described three new approaches to the analysis of capture-re-encounter data when covariates are involved. These are important in reducing bias in estimates of annual survival, so that we can better understand how wild animals survive, and whether there are changes due to corresponding changes to the climate. Results have only been presented in outline here, and the references to the paper provide far more detail with regard to the background and recent research. There is still more research to do in these areas, for instance, for illustration, the conditional modelling of the heron data considered just one out of a set of alternative models, without due regard to model-selection. The maximum-likelihood estimate of the parameter τ was obtained from a profile log-likelihood, resulting in $\tau = 13$. Similar values have been obtained for the corresponding analysis of data on other British bird species, and good results have been obtained with the scaled logistic model by assuming the same parameters for the decline in reporting probability in a combined analysis of several species, to be reported elsewhere. Further research is needed to determine how viable this conditional approach is when data are sparse, and also whether small values for τ might adversely affect the analysis. The interpolation approach for weather covariates could potentially be extended to include time- as well as space-variation. In the work done here so far, no account has been taken of the errors resulting from the fitting of the spline surface, and this is an area for future research. We note finally that the lasso approaches provide a simple classical solution to the selection of weather covariates in models for mark re-encounter data.

ACKNOWLEDGEMENTS

We thank Ian Jolliffe, Rob Robinson and David Thomson for their input to the research described in this paper.

REFERENCES (RÉFERENCES)

Baillie., S. R. and Green, R. E. (1987) The importance of variation in recovery rates when estimating survival rates from ringing recoveries. Acta Ornithologica, 23, 1, 41–60.

Besbeas, P.T., Freeman, S.N., Morgan, B.J.T. and Catchpole, E.A. (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. Biometrics, 58, 3, 540-547.

Bonner, S. J., Morgan, B.J.T. and King, R. (2010) Continuous covariates in mark-recapture-recovery analysis: A comparison of methods Biometrics, 66, 4, 1256–1265.

Brown, D. I. (2010) Climate modelling for animal survival. Unpublished PhD thesis, University of Kent, Canterbury, England.

Catchpole, E.A., Morgan, B.J.T., Freeman, S.N. & Peach, W.J. (1999) Modelling the survival of British lapwings Vanellus vanellus using weather covariates. Bird Study, 46 (suppl.), 5–13.

Catchpole, E.A., Morgan, B.J.T., Freeman, S.N., Albon, S.D. and Coulson, T.N. (2000) Factors influencing Soay sheep survival. Applied Statistics, 49, 4, 453–472.

Catchpole, E.A., Morgan, B.J.T. and G. Tavecchia, G. (2008) A new method for analysing discrete life-history data with missing covariate values.J. Roy. Statist. Soc., B., 70, 2, 445–460.

Cole, D.J. and Morgan, B.J.T. (2010) Parameter redundancy with covariates. Biometrika, 97, 1002–1005.

Gimenez, O., Barbraud, C., Crainiceanu, C. M., Jenouvrier, S. and Morgan, B.J.T..(2006) Semiparametric regression in capture-recapture modelling. Biometrics. 62, 691–698.

King, R., Brooks, S.P., Morgan, B.J.T. and Coulson, T. (2006) Factors influencing Soay sheep survival: a Bayesian analysis. Biometrics, 62, 211–220.

Mazzetta, C. (2010) Age-specificity in conditional ring-recovery models. JABES, 15, 435–451.

Morgan, B.J.T. (2006) New methods for including covariates in models for the survival of wild animals. pp 50–61, In: IWSM2006. Proceedings of the 21st International Workshop on Statistical Modelling. ISBN 1-86220-180-3.

North, P. M. & Morgan, B.J.T. (1979) Modelling heron survival using weather data. Biometrics, 35, 3, 667–682.

Robinson, R.A., Baillie, S.R. & Crick, H. Q. P. (2007) Weather-dependent survival: implications of climate change for passerine population processes. Ibis,149, 357–364.

White, G.C. and Burnham, K.P. (1999) Program MARK: Survival estimation from populations of marked individuals. Bird Study, 46 (suppl), 120–139.

RÉSUMÉ (ABSTRACT)

Covariates have been used to model parameters in capture-recapture methods in ecology since the paper by North and Morgan (Biometrics, 1979). In this paper we consider a variety of issues that remain to be answered, including how to select covariates, how to deal with spatial information and how to include time-variation in recovery probabilities for conditional analysis of ring-recovery data.