# Continuous time–dependent individual covariates and the Cormack–Jolly–Seber model

# S. J. Bonner & C. J. Schwarz

## Abstract

*Continuous time–dependent individual covariates and the Cormack–Jolly–Seber model.*— The Cormack–Jolly–Seber model provides the basic framework for analyzing the survival of animals in open populations using capture–recapture data. Extensions of this model have already been developed that allow the survival and capture probabilities to vary between individuals based on auxiliary variables, but none can allow for variables that are continuous, time–dependent, and vary among individuals. We summarize a new method for incorporating this type of variable into the Cormack–Jolly–Seber model by modelling the distribution of the unobserved values of the variable conditional on the observed values, given a few basic assumptions about how the variable changes over time. We begin with a hypothetical scenario as motivation for our model and also present the results of two examples used in developing the model.

Key words: Capture–recapture, Cormack–Jolly–Seber model, Auxiliary variables, Continuous time–dependent individual covariates.

## Resumen

*Covariantes continuas individuales dependientes del tiempo y el modelo de Cormack–Jolly–Seber.*— El modelo de Cormack–Jolly–Seber proporciona el marco básico para analizar la supervivencia de animales en poblaciones abiertas utilizando datos de captura–recaptura. Si bien se han desarrollado ampliaciones de este modelo que permiten variar las probabilidades de supervivencia y de captura entre individuos a partir de variables auxiliares, en ninguna de ellas es posible utilizar variables continuas, dependientes del tiempo y que varíen de un individuo a otro. El presente estudio analiza un nuevo método que permite la incorporación de este tipo de variable en el modelo de Cormack–Jolly–Seber mediante la modelación de la distribución de los valores no observados de la variable según los valores observados, tomando como referencia algunas asumciones básicas acerca de cómo la variable cambia con el tiempo. En primer lugar, presentamos un escenario hipotético con objeto de definir el modelo, para posteriormente indicar los resultados de dos ejemplos que utilizamos para su desarrollo.

Palabras clave: Captura–recaptura, Modelo de Cormack–Jolly–Seber, Variables auxiliares, Covariantes continuas individuales dependientes del tiempo.

*C. J. Schwarz & S. J. Bonner, Dept. of Statistics and Actuarial Science, Simon Fraser Univ., 8888 University Drive, Burnaby BC, V5A 1S6 Canada.*

Corresponding author: C. J. Schwarz. E–mail: cschwarz@stat.sfu.ca

## Introduction

The basic model for studying survival probabilities in open animal populations using capture–recapture data is the Cormack–Jolly–Seber (CJS) model (Cormack,1964; Jolly, 1965; Seber, 1965). In the model's original definition, the survival and capture probabilities are assumed to be homogeneous for all animals in the population. In many applications this assumption is not justified and the capture and survival probabilities are allowed to vary as functions of auxiliary variables (Pollock, 2002). These variables may be classified as either continuous or discrete, time–dependent or constant, and individual or external (also called environmental). Here we present a new model aimed specifically at incorporating data from continuous, time–dependent, individual covariates into the CJS framework.

While there are eight possible classifications of variables using the system above, five distinct cases need to be considered separately. The simplest case involves covariates which are both discrete and constant over time. For these variables, both external and individual, the captured animals can be divided into distinct groups based on the value of the variable and survival modelled separately in each group. Two examples are the effect of gender on the European Dipper (*Cinclus cinclus*) and the comparison of survival in two Swift (*Apus apus*) colonies, both from Lebreton et al. (1992). When the variable under consideration is continuous, but still constant over time, the survival and capture probabilities are both modelled using the framework of generalized linear modelling (McCullagh & Nelder, 1989). That is, the survival and capture probabilities are each considered dependent on a linear transform of the variable through a selected link function. An example is the study of the effect of agricultural pesticide use on Sage Grouse *(Centrocercus urophasianus),* which models survival as a function of blood cholinesterase using the log–log link to achieve proportional hazards (Skalski et al., 1993). The same method can be used for time–dependent, external covariates, both discrete and continuous, the only difference being that for such variables the linear predictor changes over time but is the same for all individuals. Early examples incorporating time–dependent, external covariates include studies of the effect of winter temperatures on the survival of the Grey Heron (*Ardea cinerea*) using recoveries of dead birds only (North & Morgan, 1979), and on the survival of the European Starling (*Sturnus vulgaris*) using recapture data (Clobert & Lebreton, 1985).

The difficulty posed in the cases involving time–dependent, individual variables is that the value of the variable is unknown for some events that contribute to the model's likelihood (i.e. capture occasions when a previously captured animal is not observed). In the case of discrete, time–dependent, individual variables, a solution to this problem is the multi–state model which incorporates a Markov chain into the CJS model to describe the movement of individuals between the different values of the

variable (Arnason, 1973; Schwarz et al., 1993). The likelihood contribution for each individual is then adjusted by including the transition probabilities for every pair of consecutive states and summing over all possible values of the unobserved quantities. When the variable is continuous the missing values cannot be modelled by a Markov chain which takes a finite number of values. Thus, it is necessary to develop a new model predicting the distribution of the missing values of the continuous variable. This paper describes one particular model.

As part of the EURING Technical Meeting's proceedings, this paper is not intended to provide mathematical details of our model, of its derivation, or of its parameter estimation. Instead, we provide a heuristic derivation of the model, a brief discussion of the method used to estimate the model parameters, and a description of two examples that were used during the model's development. Full mathematical details of the model and parameter estimation methods can be found in Simon Bonner's M.Sc. report, which is available on–line (Bonner, 2003). A manuscript with full details has been submitted to *Biometrics* (Bonner & Schwarz, 2003)*.*

## Methods

In this paper we consider only the simplest case with one scalar covariate and $k$ equally–spaced capture occasions. The basic idea behind our model is that animals living in the same study area should react similarly to the forces in their environment and so should have similar changes in the value of the covariate. For example, imagine a hypothetical capture–recapture experiment with 5 capture occasions designed to study the relationship between body mass and survival of some animal. One would expect that when food is plentiful and competition low, all animals gain mass and when conditions worsen or food becomes scarce, all animals lose mass. Possible records for four individuals are shown in figure 1. Three of the individuals are captured on all five occasions and show a definite trend in body mass, perhaps in response to environmental stresses. Between occasions two and three the animals gain mass and between occasions three and four they lose mass. The changes are similar for all animals, though small differences do occur as a result of individual variation. The fourth individual is captured only on occasions 2 and 4, and in order to make estimates of the effect of body mass on survival it is necessary to have some inference of the distribution of the individual's mass at the other occasions. The proposed model does this by supposing that the body mass of this individual follows the same trends observed in the other 3 individuals. That is, the body mass rises between capture occasions 2 and 3, drops between capture occasions 3 and 4, and remains almost constant between capture occasions 4 and 5. Note that it is not necessary to make inference about the mass on the first occasion because the CJS model conditions on the animal's first release.
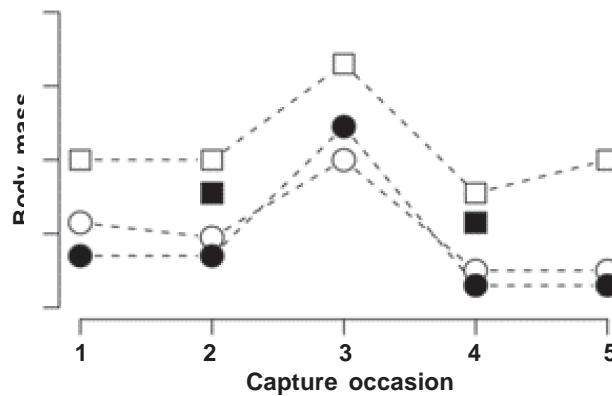
Fig. 1. Hypothetical body mass observations for four individuals in a capture–recapture experiment with five capture occasions. Observations for each individual indicated by a distinct plotting symbol.

*Fig. 1. Observaciones hipotéticas de la masa corporal de cuatro individuos en un experimento de captura–recaptura que incluía cinco ocasiones de captura. Las observaciones correspondientes a cada individuo se indican por medio de una línea discontinua.*

More specifically, our model imposes the assumption that for all individuals the change in the value of the variable between capture occasions $t$ and $t + 1$ is normally distributed with mean $\mu_t$, which varies with $t$, and constant variance, $\sigma^2$. Denoting the values of the variable for individual $i$ at capture occasion $t$ by $z_{it}$, this defines the conditional relationship:

$$z_{i,t+1} \mid z_{it} \sim N(z_{it} + \mu_t, \sigma^2) \qquad (1)$$

The values of the covariate for a single individual form a Markov chain in discrete time with continuous state–space, in some sense, the logical extension of the multi–state model to the continuous case. In accordance with our motivating example, the distribution of the change in the covariate for individual $i$ at one capture occasion given the value at the previous occasion is normally distributed with the same mean for all individuals.

Values of the continuous variable are linked to the survival and capture probabilities using the logistic link function (Lebreton et al., 1992). Denoting the capture and survival probabilities as functions of the covariate by $\phi(z)$ and $p(z)$ respectively, the link functions are:

$$\phi(z) = \frac{e^{(\beta_0 + \beta_1 z)}}{1 + e^{(\beta_0 + \beta_1 z)}} \quad \text{and} \quad p(z) = \frac{e^{(\gamma_0 + \gamma_1 z)}}{1 + e^{(\gamma_0 + \gamma_1 z)}} \quad (2)$$

where $(\beta_0, \beta_1)$ and $(\gamma_0, \gamma_1)$ are the vectors of coefficients of the survival and capture curves respectively. Ultimately, these coefficients determine how the covariate affects the survival and capture probabilities.

The model with a single scalar covariate contains $k + 4$ parameters to estimate: $k - 1$ mean differences ($\mu_1, ..., \mu_{k-1}$), the variance parameter ($\sigma^2$),

and the four coefficients of the logistic functions ($(\beta_0, \beta_1)$ and $(\gamma_0, \gamma_1)$). The likelihood function for the model is similar to the likelihood of the basic CJS model with three modifications. First, the survival and capture probabilities are replaced by the functions of the covariate in equation 2. Second, a new product of terms is added for each individual which accounts for the changes in the covariate between each pair of adjacent capture occasions. Third, for each individual it is necessary to integrate their contribution to the likelihood with respect to each unobserved covariate value in order to account for all possible configurations of the missing covariates. The second and third modifications are analogous to summing over all possible transitions for the missing states in the likelihood of the multi–state model. However, the integrals are potentially of dimension $k - 1$ and it is impossible to find maximum likelihood estimates of the parameters analytically. Instead, parameter estimates are computed via a Bayesian estimation scheme using the Metropolis–Hastings algorithm.

As discussed in the EURING 2003 short course, Bayesian analysis is particularly well suited for estimation problems involving large proportions of missing data, including capture–recapture experiments. The primary reason for this is that Bayesian statistics does not differentiate between the unknown parameters and the missing data in the same way that frequentist statistics does. Rather, both parameters and missing data are considered as unknown random variables with some underlying distribution. In a Bayesian analysis, inference about the parameters is derived from the posterior distribution, which is proportional to the product of the prior distributions and the model's likelihood function (Carlin & Louis, 1996). In both examples
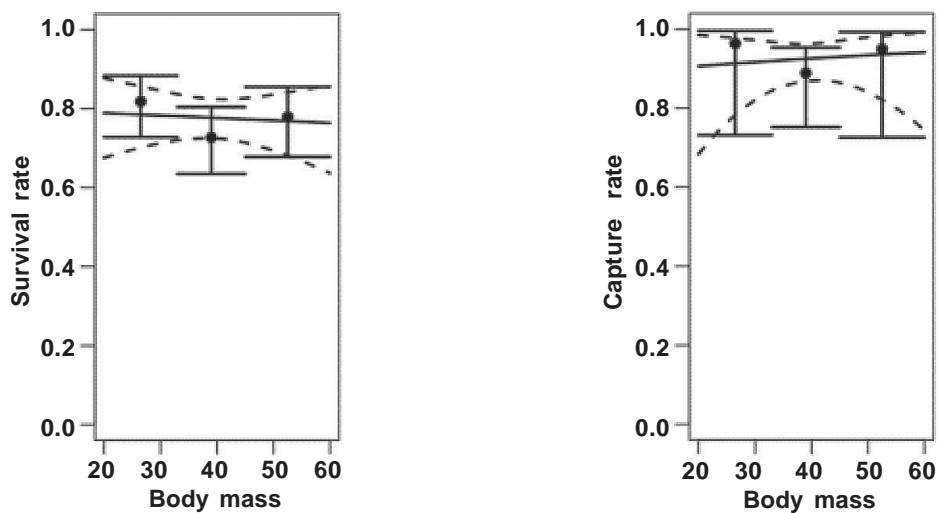
Fig. 2. Estimated survival (left) and capture (right) probabilities as functions of body mass (g) for the Meadow Vole (solid lines) with point–wise 95% credible intervals (broken lines). Probabilities estimated for the multi–state model are shown as point estimates with 95% confidence intervals for each of the three states.

*Fig. 2. Supervivencia estimada (izquierda) y probabilidades de captura (derecha) como funciones de masa corporal (g) para el topillo de Pensilvania (líneas continuas) con intervalos puntuales creíbles al 95% (líneas discontinuas). Las probabilidades estimadas para el modelo multiestado se indican como estimaciones puntuales con intervalos de confianza al 95% para cada uno de los tres estados.*

presented, the prior distributions were chosen to simplify the calculations needed to compute estimates rather than to represent true a priori information. Conjugate priors were used for the expected change on each occasion, $\mu_t$, and the variance, $\sigma^2$, and these were the normal and the inverse Gaussian, respectively. Improper flat priors with equal mass at all points of the real line were chosen for the coefficients of both the survival and capture probability functions.

As is the case with most analyses involving a large number of parameters, the posterior distribution in the examples does not belong to a regular family of distributions and numerical methods were needed to generate parameter estimates. We used the Metropolis–Hastings (MH) algorithm, a specific Markov chain Monte Carlo (MCMC) technique that is common in Bayesian analysis (Chib & Greenberg, 1995). Like all MCMC methods the MH algorithm works by successively simulating values of the parameters in such a way as to generate a Markov chain whose stationary distribution is equal to the posterior distribution. Using selected starting values, the chain is iterated many times until it appears that the tail of the chain has converged in distribution. Values of the parameters from the remaining iterations of the chain are then used as if they formed a random sample from the posterior distribution. In the following examples, the algorithm was iterated 1,000,000 times in total and values from the final 200,000 iterations were used to generate point estimates and credible intervals for each parameter (Louis & Carlin, 1996). Multiple chains were run using different starting values to check the convergence of the chain.

## Results

### Meadow Voles

The primary data set used in developing the model concerned the effect of body mass on the survival of the North American Meadow Vole (*Microtus pennsylvanicus*). The data set contained observations of 515 voles captured at the Patuxent Wildlife Research Center, Maryland, on four capture occasions in the fall of 1981 and spring of 1982. To satisfy the assumption that all animals change weight in a similar manner, captures of juvenile animals were removed from the data set. A vole was considered juvenile if its mass was less than 22 g (Nichols et al., 1992) and for each individual, only the observations where the mass was actually less than this mark were deleted. No individual was observed with a mass less than 22 g after being captured with a mass greater than 22 g. Records for individuals captured only on the last occasion were also removed because they do not contribute to the likelihood function. The final data set comprised 450 captures of 199 voles.
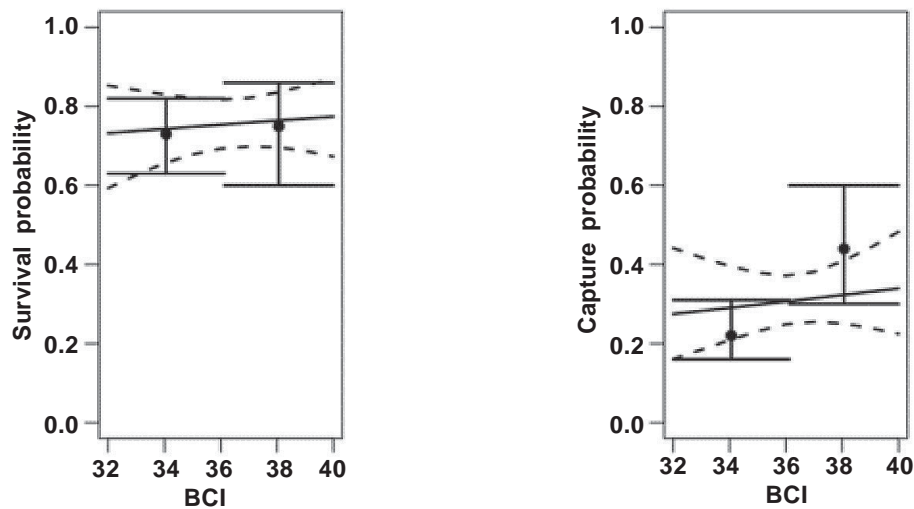
Fig. 3. Estimated survival (left) and capture (right) probabilities as functions of body condition index (g/mm) for the Lesser Snow Goose (solid lines) with point–wise 95% credible intervals (broken lines). Probabilities estimated for the multi–state model are shown as point estimates with 95% confidence intervals for each of the two states.

*Fig. 3. Supervivencia estimada (izquierda) y probabilidades de captura (derecha) como funciones de condición corporal (g/mm) para el ansar nival (líneas continuas) con intervalos puntuales creíbles al 95% (líneas discontinuas). Las probabilidades estimadas para el modelo multiestado se indican como estimaciones puntuales con intervalos de confianza al 95% para cada uno de los dos estados.*

The estimated capture and survival probabilities as functions of the voles' body mass are in shown figure 2 along with point–wise 95% credible intervals. The estimated survival probability decreases slightly with the weight of the animal and the estimated capture probability increases slightly, though neither of these effects is significant at the $\alpha = 0.05$ level. A constant survival probability near 0.8 and constant capture probability near 0.9 appears to fit all animals.

For comparison, a multi–state model using discrete mass classes was also fit to the data. Following Nichols et al. (1992), the body mass for the adult voles was divided into three states (22–33 g, 34–45 g, and > 45g) and estimates of the capture and survival probabilities for each state (fig. 2), along with the transition probabilities, were computed using Program MARK (White & Burnham, 1999). To match the assumptions of our model, transition probabilities in the multi–state model were allowed to change over time, but survival and capture probabilities were assumed constant. For both survival and capture there is considerable overlap of the 95% confidence intervals for the point estimates from the multi–state model and the point–wise 95% credible intervals for our continuous probability functions over the observed range of body mass. This suggest no discrepancies between the two methods. As with our method, the multi–state model shows no significant differences in either the survival or capture probabilities.

## Snow Geese

The second data set used in developing the model contained information on 31,240 Lesser Snow Geese (*Chen carulescens*) captured in Northern Manitoba over a 19–year period. Capture occasions occurred on an annual basis when the adult geese underwent post–breeding molt and became flightless for a short period of time. In this experiment, the researchers also weighed the birds and collected vital body measurements at each capture.

Here we have examined the effect of a body condition index (BCI) defined as the ratio of body mass to culmen length (g/mm). Because the MH algorithm for computing the parameter estimates was computationally intensive, the analysis was restricted to a 6–year subset of the original data for the purpose of model development. Further, the records for several geese were missing values for either body mass or culmen length on at least one occasion and all records for these individuals were removed. As in the previous example, observations for juvenile birds, identified at the time of capture, and for birds captured only on the 6th occasion were also removed from the data set. The final data set contained a total of 474 observations for 314 geese.

As in the previous example, the results show no significant effect of BCI on either the survival or capture probability of the geese (fig. 3). The 95%

credible intervals are relatively wide and suggest a survival probability of 0.75 and capture probability of 0.3 for all birds.

As above, a multi–state model was also fit for comparison with our model. In this case the geese were divided into 2 states using the median observed value of the BCI. Again, there is significant overlap between the 95% credible intervals for the continuous capture and survival probability functions and the 95% confidence intervals of the point estimates from the mulit–state model (fig. 3). However, while the estimated survival probability function passes very close to the survival estimates for the multi–state model, there appears to be some discrepancy between the estimated capture probability function and multi–state capture estimates. We believe that this might be an indication that the logistic link function is not appropriate in this case. If there actually was a significant change in the capture probability of the magnitude suggested by the estimates from the multi–state model, then the continuous capture probability function would need to be very steep near the median BCI value. For a logistic curve, this would result in very low capture probabilities for birds with low BCI and very high capture probabilities for geese with high BCI. Instead, the best fitting logistic function may be fairly flat and the true effect may not be observed. Another reason for the discrepancy may be the relatively small size of the data set caused by removing the observations with missing data. Both the 95% credible intervals for the continuous function and 95% confidence intervals for the multi–state estimates are very wide, indicating that the capture probability estimates produced by both models are highly variable. We are continuing to develop the model in order to address both of these issues.

## Discussion

The method presented here provides a general framework for incorporating the effect of continuous, time–dependent, individual covariates into open population capture–recapture models. The model assumes that changes in the covariate are similar for all individuals in the population and constructs the distribution of the unknown values of the auxiliary variable conditional on the observed values. This information is then incorporated into the CJS model likelihood using selected link functions in order to estimate the survival and capture probabilities as continuous functions of the variable. In the specific model described, changes in the covariate are assumed to be normally distributed with time–dependent mean and constant variance, and the logistic function is used as the link for both the capture and survival probabilities. We believe that this model presents significant advantages over the current approach of categorizing continuous covariates and incorporating them into a multi–state model.

The differences between the proposed model and the multi–state model are similar to the differences between a model using completely separate, static groups of animals and one incorporating information on a continuous, but constant, variable. The main advantage of the proposed model is that it can incorporate continuous covariates without requiring them to be categorized. Although continuous covariates may used in a multi–state model by dividing their range into discrete intervals which are treated as distinct states (see Nichols et al., 1992, for example), the divisions may be arbitrary and different categorizations may lead to different conclusions about the variable's effect. Using a small number of divisions may impose unrealistic assumptions about the similarity of individuals in the same state and obscure the underlying changes in the capture or survival probabilities. Using a larger number of divisions may lead to problems with model identifiability. In some situations, the results of a multi–state model may also be more difficult to interpret. In particular, if a large number of divisions is used then it becomes hard to describe changes in the covariate over time based on the transition rates. In comparison, our model produces direct information about changes in the variable's distribution. Comparing different models fit using different sets of auxiliary variables will also provide a simple way to test the effect of each variable on the animals' survival probability.

The main disadvantage of our approach is that it imposes assumptions on both the distribution of the covariate and on the relationship between the covariate and the survival and capture probabilities. If these assumptions are not satisfied then the model will fit the data poorly and provide erroneous conclusions. In particular, the use of the logistic link function in the current model may be inappropriate in some situations, e.g. if the survival or capture probabilities are not monotonic functions of the auxiliary variable. If this is the case and a suitable link function cannot be found, then the multi–state model, which makes no assumptions about the relationship between the capture and survival probabilities for the different states, will be more appropriate.

As future work, we envisage several extensions of our model to accommodate different assumptions about the behaviour of the covariate and the relationship between the covariate and the survival and capture probabilities. One of the basic assumptions of our model is that the changes in the covariate have the same distribution for all individuals in the population. As in both examples described, this assumption would likely not be satisfied when considering changes in body measurements for both adult and juvenile animals. A simple extension would use two values of the expected change for each pair of capture occasion, one for adults and one for juvenile. It should also be possible to fit models in which the expected change in the auxiliary variable depends on the value of the variable itself. For example, modelling the behaviour of the variable on the

log scale would generate a multiplicative model in which the expected change in the value of the covariate is proportional to its current value. Another constraint of the proposed model is that the survival and capture probabilities for a given value of the covariate are assumed constant over time. Time–dependent survival probabilities could be based on the Cox proportional hazards model for survival analysis, which allows for variation in survival over time under the assumption that the relative hazard for two different values of the auxiliary variable is constant (Cox, 1972). The same model could be used for the capture probability as well.

## Acknowledgments

## References

Arnason, A. N., 1973. The estimation of population size, migration rates, and survival in a stratified population. *Research in Population Ecology*, 15: 1–8.

Bonner, S. J., 2003. Continuous Time–Dependent Individual Covariates in the Cormack–Jolly–Seber Model. M. Sc. Thesis, Simon Fraser Univ.. (Available online at: www.stat.sfu.ca/alumni)

Bonner, S. J. & Schwarz, C. J., 2004. An Extension of the Cormack–Jolly Seber model for continuous covariates with application to *Microtus pennsylvanicus*. *Biometrics* (in press).

Carlin, B. P. & Louis, T. A., 1996. *Bayes and Empirical Bayes Methods for Data analysis*. Chapman and Hall, London.

Chib, S. & Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. *The American Statistician,* 49: 327–335.

Cormack, R. M., Estimation of survival from the sighting of marked animals. *Biometrika,* 51: 429–438.

Cox, D. R., 1972. Regression models and life–tables. *Journal of the Royal Statistical Society*, 34: 187–220.

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

Clobert, J. & Lebreton, J.–D., 1985. Dépendence de Facteurs de Milieu dans les Estimations de Taux de Survie par Capture–Recapture. *Biometrics*, 41: 1031–1037.

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

McCullagh, P. & Nelder, J. A., 1989. *Generalized linear models*. Second edition. Chapman and Hall, New York, U.S.A.

Nichols, J. D., Sauer, J.R., Pollock, K. H. & Hestbeck, J. B., 1992. Estimating transition probabilities for stage–based population projection matrices using capture–recapture data. *Ecology*, 73: 306–312.

North, P. M. & Morgan, B. J. T., 1979. Modelling Heron survival using weather data. *Biometrics*, 35: 667–681

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

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

Schwarz, C. H., Schweigert, J. F. & Arnason, N. A., 1993. Estimating migration rates using tag–recovery data. *Biometrics*, 49: 177–193.

Skalski, J. R., Hoffmann, A. & Smith, S.G., 1993. Testing the Significance of Individual and Cohort Level Covariates in Animal Survival Studies. In *Marked Individuals in the Study of Bird Population* (J.–D. Lebreton & P. M. North, Eds.). Bikhäuser Verlag, Boston.

White, G. C. & Burnham, K. P., 1999. Program mark–survival estimation from populations of marked animals. *Bird Study*, 46 Supplement: 120–138.