Automatic generation of multistate capture-recapture models

Rémi CHOQUET

Abstract: Nowadays multistate capture-recapture models are used extensively in biological studies. They feature movement parameters possibly associated with the quality of sites; they also allow a combination of different sources of information to improve the stability and the accuracy of the estimates. Model refinements potentially yield more precise information on biological status. The integration of extensive biological information in the model increases its complexity, however. Thus without an appropriate tool, model building and selection may be time consuming. In this paper, the author describes a tool called GEMACO dedicated to the automatic generation of design matrices for multistate models by means of a programming language. This symbolic and flexible tool avoids tedious matrix manipulations. It is well suited for the description of complex structures and has been implemented in a generic multistate program called M-SURGE freely available at http://www.cefe.cnrs.fr/BIOM/logiciels.htm.

Génération automatique de modèles multi-états en capture-recapture

Résumé : De nos jours, les modèles de capture-recapture multi-états sont largement utilisés dans les études biologiques. Ils permettent d'associer les mouvements des animaux à la qualité des sites géographiques ; ils autorisent aussi l'intégration de différentes sources d'information susceptibles d'améliorer la stabilité et la précision des résultats. Le raffinement des modèles peut fournir des informations plus précises concernant la biologie des individus. Cependant, la complexité des modèles augmente avec l'intégration d'information biologique supplémentaire. Sans un outil approprié, la construction et la sélection de modèles peut donc s'avérer chronophage. Dans cet article, l'auteur décrit un outil appelé GEMACO pour la génération automatique de matrices de contraintes pour les modèles multi-états à l'aide d'un langage de programmation. Cet outil symbolique flexible permet d'éviter de fastidieuses manipulations matricielles. Il est bien adapté à la description de modèles complexes et est d'ores et déjà implanté dans un gratuiciel appelé M-SURGE disponible à http://www.cefe.cnrs.fr/BIOM/logiciels.htm.

1. INTRODUCTION

Statistical multistate models are used to describe capture-recapture data where individually recognizable animals are followed over time at discrete time periods. Several biological questions can be investigated by suitably defining the states of the model. The dispersal of a species can be looked at when states are sites (see Henaux, Bregnballe & Lebreton 2007). The cost of reproduction and biological trade-offs are investigated when states are reproductive states, i.e., breeder versus non-breeder (see Rivalan et al. 2005). A young animal is a non-breeder since it cannot have offspring; typically the transition probabilities from non-breeder to breeder depend on the age of the animal (see Crespin et al. 2006). The investigation of the true survival and the emigration out of the study area uses states defined as "live capture" and "dead recovery" (Lebreton, Almeras & Pradel 1999).

A capture history is a sequence of observations of length K denoted $\omega = (\omega_1, \ldots, \omega_K)$ where $w_t = 0$ if a unit is missed at time, or occasion, t. When it is caught, $w_t = n$, where n is one of N possible states in which a unit can be captured. The first non-zero entry of ω identifies the time at which the animal is marked and released. The data set contains the frequencies n_ω of each capture history among Ω, the set of all possible capture histories. The data may also contain

information for several groups, such as the males and the females. When there are G groups, the data set contains G frequencies $n_{\omega, g=1, \ldots, G}$ for each history ω.

The reference model for multistate data is the conditional Arnason–Schwarz model (Arnason 1973). Appropriate goodness-of-fit tests were developed for the conditional Arnason–Schwarz model in Pradel, Wintrebert & Gimenez (2003), making its use very reliable. This model is conditional on the first capture; this first capture identifies the cohort. Each unit belongs to one cohort. The parameter vector associated to the conditional Arnason–Schwarz model is denoted by Θ; then $P_\omega(\Theta)$ is the probability of the capture history ω. For a data set $\{n_\omega\}_{\omega \in \Omega}$ featuring only one group, the likelihood L is a product multinomial distribution

$$L(\Theta) = \prod_{\omega \in \Omega} P_\omega(\Theta)^{n_\omega}. \tag{1}$$

For each cohort, one has a multinomial distribution. The likelihood for several groups is the product of the likelihoods for the individual groups. Model selection is used to test biological assumptions (Burnham & Anderson 2002). Thus we need to fit several complex models to this data by deriving maximum likelihood estimates of the parameters. Once one has a method to build Θ and each $P_\omega(\Theta)$, the estimates can be obtained by maximizing (1) using a general purpose maximization routine.

Transitions among states follow a non-homogeneous Markov chain of order one and each $P_\omega(\Theta)$ is evaluated using a product of matrices (Brownie et al. 1993). For multistate models, matrices are a function of basic parameters which are the survival-transition probabilities noted ϕ, and the capture probabilities noted p which are, respectively,

- $\phi_{ij}^{(t)}$ the probability of being alive in state j at time $t + 1$ if alive in state i at time t, $t = 1, \ldots, K - 1$ and $i, j = 1, \ldots, N$.

- $p_j^{(t)}$ the probability of capturing an animal in state j at time t, $t = 2, \ldots, K$ and $i = 1, \ldots, N$.

In matrix notation, $\Phi^{(t)} = (\phi_{ij}^{(t)})$ denotes the $N \times N$ matrix of survival-transition probabilities, $P^{(t)} = (p_j^{(t)})$ denotes the $N \times 1$ vector of capture probabilities. The survival-transition matrix is usually decomposed into the product of a survival matrix and a transition matrix as

$$\Phi^{(t)} = \text{diag}\left(S^{(t)}\right)\Psi^{(t)}, \tag{2}$$

where $S^{(t)} = (s_i^{(t)})$ denotes the $N \times 1$ vector of survival probabilities, the operator diag transforms a vector into a diagonal matrix and $\Psi^{(t)} = (\psi_{ij}^{(t)})$ denotes the $N \times N$ matrix of transition probabilities conditional on survival. The new parameters s, ψ are defined as:

- $s_i^{(t)} =$ the probability that an animal present and alive in state i at time t will be alive at time $t + 1$.

- $\psi_{ij}^{(t)} =$ the probability of being in state j at time $t + 1$ if alive in state i at time t conditional on survival over the interval.

Using these parameters, we obtain the predicted frequency for the capture history $w = (1, 0, 2, 1)$ which is $R_1 P_\omega(\Theta)$, where R_1 is the number of animals released on occasion 1 and

$$P_\omega(\Theta) = e_1^\top \Phi^{(1)} \text{diag}\left(1 - P^{(2)}\right)\Phi^{(2)} P^{(3)} e_2 e_2^\top \Phi^{(3)} P^{(4)} e_1.$$

In the basic conditional Arnason–Schwarz model presented above, the parameters can vary with the capture occasion t and the states of origin and of destination; they depend on time factors TIME, FROM and TO where

- FROM = the state of departure, which is identified using the subscript i. This factor does not apply to the capture rate.

- TO = the state of arrival, which is identified using the subscript j. This factor does not apply to the survival rate.

Other factors entering in the formulation of complex models are

- AGE = the number of capture occasions elapsed since the release, superscript a is used to identify this factor;

- COHORT = the date of first capture, this factor will not be used in the sequel;

- GROUP = a classification variable for the animals under study, which is identified using the superscript g.

When the animals are marked young, AGE $= 1$ identifies the young, while AGE > 1 typically corresponds to adults. The examples discussed below involve an age effect and a group effect. Additional indices a and g are used to represent these two factors. Thus $s_i^{(t,a,g)}$ stands for the survival probability at age a for units in group g between capture occasions t and $t + 1$.

Let S, Ψ, Φ, P be respectively the row vectors of the parameters s, ψ, ϕ, p. In the following, the vector $\Theta = (S, \Psi, P)$ and $\Theta = (\Phi, P)$ are called the vector of the biological parameters. The parametrization (S, Ψ, P) (resp. (Φ, P)) is called the separate (resp. combined) parametrization of the conditional Arnason–Schwarz model. Generalized linear models (see McCullagh & Nelder 1989) are widely used for single state capture-recapture models (see Lebreton, Burnham, Clobert & Anderson 1992), and similar ideas apply to multistate models. The umbrella model is then constrained in a linear fashion as

$$f(\Theta) = X\beta,$$

where f is called the link function and the vector β is the vector of mathematical parameters. In practice, survival (apparent survival), transition and capture parameters are constrained independently and X is block diagonal,

$$\begin{pmatrix} f_1(S) \\ f_2(\Psi) \\ f_3(P) \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}; \tag{3}$$

X has a similar structure for $\Theta = (\Phi, P)$.

In this paper, we will address the question of how to implement constraints or, equivalently, how to build X_1, X_2, X_3 for complex models. The classical approach used in capture-recapture (Pradel & Lebreton 1991; White & Burnham 1999) is based on what is called the parameter index matrix approach. It is not well adapted to the multistate framework. To overcome these problems, a new approach based on a language, similar to what exists in general statistical software packages such as S-PLUS (Chambers & Hastie 1992) and R (Ihaka & Gentleman 1996), has been developed for multistate models. This new tool named GEMACO (GEnerator of MAtrices of COnstraints) automatically generates design matrices from a sentence.

The parameter index matrix and its limitations are described in Section 2. Then GEMACO is presented in Section 3. A general algorithm for dealing with redundancy due to aliasing is described in detail. GEMACO is then applied to write the complex example of multisite accession to reproduction in a simple way. Finally in Section 4, we suggest using the concept of product of transition matrices to extend GEMACO to models more complex than the conditional Arnason–Schwarz. This generalization allows one to write a wider range of models than the parameterizations (S, Ψ, P) and (Φ, P) used in existing softwares such as M-SURGE (Choquet et al. 2004) and MARK (White & Burnham 1999).

2. THE PARAMETER INDEX MATRIX APPROACH AND ITS LIMITATIONS

2.1. Description.

The parameter index matrix approach is the first method used to express complex capture-recapture models. It was initially implemented in SURGE (Pradel & Lebreton 1991) and reused in MARK (White & Burnham 1999). It gives an index to each biological parameter (survival, transition, capture). For each type of parameter, one has a triangular array of indices. Each of these indices corresponds to a parameter in the model. For survival, in the triangular tables, a column is a recapture occasion and a row is a cohort effect (i.e., a time of first release). Imposing constraints on the index matrix parameters is difficult. Special programming steps are needed to implement a constraint involving more than two parameters, such as $\beta_1 + \beta_2 = \beta_3$.

Figure 1 presents a model with $K = 5$ capture occasions, two age classes (1 versus 2 to 4 combined) and a AGE by TIME interaction (Pradel et al. 1997). The time varying survival probabilities for animals that have just been captured ($a = 1$) is not the same as those for animals captured previously ($a = 2$). Such a model is appropriate when transient animals are marked. These animals cannot be recaptured since they leave the area; they create a negative bias in the survival estimates at $a = 1$. The parameters β are numbered by occasion of recapture within age-class 1 ($\beta_1, \beta_2, \beta_3, \beta_4$) and age-classes 2 to 4 ($\beta_5, \beta_6, \beta_7$), see Figure 1. The statistical relationship can be summarized by $\mathrm{logit}(s^{(t,1)}) = \beta_t$ for $t = 1, \ldots, 4$ and $\mathrm{logit}(s^{(t,2)}) = \beta_{t+3}$ for $t = 2, \ldots, 4$. Although models described by these triangular structures can be further particularized by constraints, the parameter index matrix approach is not adapted to build additive models (see Examples 1 and 2 below). It is quite complex when several states are involved (see Example 3 below). Since it is not a formal representation, it is neither statistically nor biologically relevant.
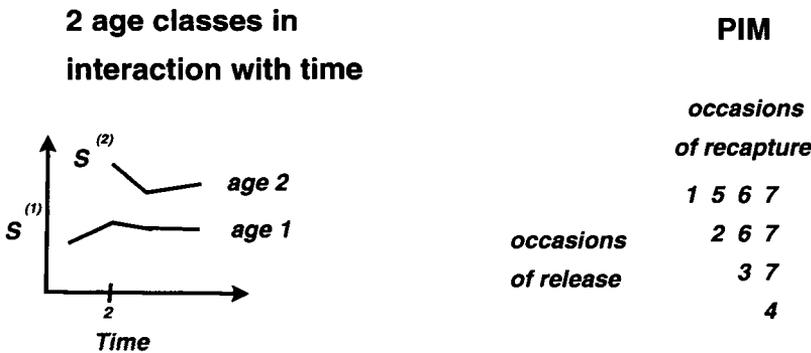

FIGURE 1: Parameter index matrix (PIM) for the survival rate featuring an AGE by TIME interaction, where AGE has two levels (AGE =1 vs AGE >1). It highlights that 7 degrees of freedom are used to model survival: $\mathrm{logit}(s^{(t,1)}) = \beta_t$ for $t = 1, \ldots, 4$ and $\mathrm{logit}(s^{(t,a)}) = \beta_{t+3}$ for $t = 2, \ldots, 4$ and $a = 2, \ldots, 4$.

2.2. Example 1: An additive model for survival in the presence of transient animals.

We consider the case where survival rates for the two age classes are parallel over time when expressed on the logit scale (see Figure 2). It differs from the model of Figure 1 by the absence of the AGE by TIME interaction. Relations between biological parameters s and mathematical parameters β are defined by $\mathrm{logit}(s^{(t,1)}) = \beta_t$ for $t = 1, \ldots, 4$, and $\mathrm{logit}(s^{(t,a)}) = \beta_t + \beta_5$ for $t = 2, \ldots, 4$ and $a = 2, \ldots, 4$ show clearly that the parameter index matrix approach does not work.

2.3. Example 2: A model with true age.

In this application, we consider a model where the true age (time elapsed since the date of birth) of individuals is needed. This model is used by Loison et al. (1999) to show evidence of senescence in ungulates. For the purpose of this example, we consider three groups of individuals first

marked with known age 1 to 3 with only three occasions of recapture. The group g gives the true ages at the first capture; thus true age is then $a + g - 1$. A basic model where survival depends only on true age (logit($s^{a,g}$) $= \beta_{a+g-1}$ for $a = 1, \ldots, 3$ and $g = 1, \ldots, 3$) can be built with parameter index matrices. A more realistic model has additive effects of true age and time for adults (i.e., $a + g - 1 \geq 2$) and a constant survival for juveniles ($a + g - 1 = 1$)

$$
\begin{aligned}
\text{logit}\left(s^{(1,1,1)}\right) &= \beta_1, \\
\text{logit}\left(s^{(t,a,g)}\right) &= \beta_{a+g-1} + \alpha_t \quad \text{with } t = 2,3 \text{ and } a + g \geq 1.
\end{aligned} \tag{4}
$$

As for the previous example, this model cannot be built with only parameter index matrices.

additive models between

2 age classes and time
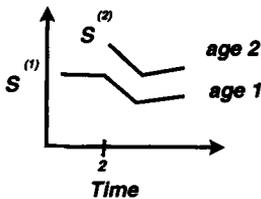
FIGURE 2: An AGE + TIME model that cannot be represented by using only parameter index matrices. The relations logit($s^{(t,1)}$) $= \beta_t$ for $t = 1, \ldots, 4$ and logit($s^{(t,a)}$) $= \beta_t + \beta_5$ for $t = 2, \ldots, 4$ and $a = 2, \ldots, 4$ cannot be built by assigning values to the index matrix parameters of Figure 1.

2.4. Example 3: A model of multisite accession to reproduction.

This model of multisite accession to reproduction is published in Lebreton et al. (2003). The study investigated natal dispersal among three breeding colonies and accession to reproduction for the Roseate Tern. The data were collected in three colonies of the Long Island Sound: Connecticut, New York and Massachusetts. In this example, probabilities of dispersal among the three colonies noted (1,2,3) are related to the biological states breeder and non-breeder. Thus, we consider six states, the first three states are denoted (B1,B2,B3) for breeders on sites (1,2,3) and the three others denoted (N1,N2,N3) for non-breeder on sites (1,2,3). The transitions are described in Table 1 where AGE denotes the true age of a bird. In this example, parameters vary by AGE and states. The probability of a transition from non-breeder to breeder depends on the AGE and on the colony. Building this complex model is hard as there are 36 parameter index matrices, one for each element of the transition matrix. Since there are $K = 11$ capture occasions, each parameter index matrix can have up to $10 \times 11/2 = 55$ parameters for its relationship with AGE and TIME. So clearly, we need a formal tool.

3. THE GENERATOR OF MATRICES OF CONSTRAINTS APPROACH

A brief description of the implementation of GEMACO is given Section 3.1. It focusses on the machinery underlying GEMACO, the kernel of which is the model definition language (MDL). Then a detailed description of GEMACO is provided and the models presented in the examples are cast into this language. Additional details on the basic features of GEMACO are given in Choquet et al. (2004). In Section 3.2, we describe the basic features of the model definition language, inspired by the formal language described in Wilkinson & Rogers (1973). It is similar to several languages used in statistical software packages, such as R and S-PLUS. In Section 3.3, we describe the extended features of the model definition language. They include a general algorithm to deal with redundancy due to aliasing which goes beyond the partial treatment usually found in standard packages. Examples 1 to 3 are then revisited using the model definition language in Sections 3.4, 3.5, and 3.6.

TABLE 1: Transition matrices $\Psi^{(a)}$ for AGE $= 1$ at left top corner, for AGE $= 2, \ldots, 5$ at right top corner, for AGE $= 6$ at left bottom corner, for AGE $= 7, \ldots, 10$ at right bottom corner. These matrices are all different making the parameter index matrix approach very time consuming and error-prone! The notation used is $\overline{x} = 1 - x$. Some transition parameter values are null. In particular, a breeder remains a breeder. More details on the model are given in Lebreton et al. (2003).

	AGE = 1						AGE = 2–5					
	B1(1)	B2(2)	B3(3)	N1(4)	N2(5)	N3(6)	B1	B2	B3	N1	N2	N3
B1(1)	$\sum_{j\neq1}\overline{\psi_{1j}}$	ψ_{12}	ψ_{13}	0	0	0	$\sum_{j\neq1}\overline{\psi_{1j}}$	ψ_{12}	ψ_{13}	0	0	0
B2(2)	ψ_{21}	$\sum_{j\neq2}\overline{\psi_{2j}}$	ψ_{23}	0	0	0	ψ_{21}	$\sum_{j\neq2}\overline{\psi_{2j}}$	ψ_{23}	0	0	0
B3(3)	ψ_{31}	ψ_{32}	$\sum_{j\neq3}\overline{\psi_{3j}}$	0	0	0	ψ_{31}	ψ_{32}	$\sum_{j\neq3}\overline{\psi_{3j}}$	0	0	0
N1(4)	0	0	0	$\sum_{j\neq4}\overline{\psi_{4j}^{(1)}}$	$\psi_{45}^{(1)}$	$\overline{\psi_{46}^{(1)}}$	$\psi_{41}^{(a)}$	0	0	$\overline{\psi_{41}^{(a)}}$	0	0
N2(5)	0	0	0	$\psi_{54}^{(1)}$	$\sum_{j\neq5}\overline{\psi_{5j}^{(1)}}$	$\overline{\psi_{56}^{(1)}}$	0	$\psi_{52}^{(a)}$	0	0	$\overline{\psi_{52}^{(a)}}$	0
N3(6)	0	0	0	$\psi_{64}^{(1)}$	$\psi_{65}^{(1)}$	$\sum_{j\neq6}\overline{\psi_{6j}^{(1)}}$	0	0	$\psi_{63}^{(a)}$	0	0	$\overline{\psi_{63}^{(a)}}$
	AGE = 6						AGE = 7–10					
B1	$\sum_{j\neq1}\overline{\psi_{1j}}$	ψ_{12}	ψ_{13}	0	0	0	$\sum_{j\neq1}\overline{\psi_{1j}}$	ψ_{12}	ψ_{13}	0	0	0
B2	ψ_{21}	$\sum_{j\neq2}\overline{\psi_{2j}}$	ψ_{23}	0	0	0	ψ_{21}	$\sum_{j\neq2}\overline{\psi_{2j}}$	ψ_{23}	0	0	0
B3	ψ_{31}	ψ_{32}	$\sum_{j\neq3}\overline{\psi_{3j}}$	0	0	0	ψ_{31}	ψ_{32}	$\sum_{j\neq3}\overline{\psi_{3j}}$	0	0	0
N1	1	0	0	0	0	0	0	0	0	1	0	0
N2	0	1	0	0	0	0	0	0	0	0	1	0
N3	0	0	1	0	0	0	0	0	0	0	0	1

3.1. Structure and implementation of GEMACO.

GEMACO is used to formulate models for either the vector $\Theta = (S, \Psi, P)$ or (Φ, P). It consists of two components:

1. A language to describe models. The model definition language is presented briefly in Choquet et al. (2004). Inspired by the formal description of linear models in Wilkinson & Rogers (1973), it was applied by Lebreton, Burnham, Clobert & Anderson (1992) in the capture-recapture context, which extends this work to multistate capture-recapture models. One particularity of multistate models is that we need to describe a non-homogeneous Markov chain. Non-homogeneity of Markov chains might occur across time, age (and as a consequence, across cohorts).

2. An interpreter to build design matrices similar to those found in S-PLUS (Chambers & Hastie 1992) and R (Ihaka & Gentleman 1996).

The language and the interpreter were implemented using the lexical analyser generator Flex (Paxson 1994) and a parser generator Bison (Donnelly & Stallman 1994). Two steps are necessary to generate a design matrix from a phrase. The lexical analyzer cuts the sentence describing the model into tokens, made of keywords describing the factors and the operators. Then the parser builds the design matrix by replacing the keywords by their matrix representations and by combining the matrices according to the commands conveyed by the operators. As GEMACO builds the design matrix automatically, the user does not need to know the structure of the matrix, but only has to learn a simple language.

3.2. Basic features of the language.

The basic form of a phrase for an ordinary linear model has a syntax similar to that used in R. A phrase has the following syntax:

$$term_1 \; op_1 \; term_2 \; op_2 \; term_3 \ldots \qquad (5)$$

The term $term_j$ refers to factors (or possibly a combination of factors), while op_j refers to operators.

Built-in keywords for factor.

The main factors TIME, AGE, COHORT, GROUP, FROM, TO are represented in the model definition language by built-in keywords, respectively t, a, c, g, f, to. Each factor is equivalent to a matrix in the context of generalized linear models, where each column is the indicator of one factor level. Systematically, the columns of the matrix sum to the INTERCEPT with keyword i, a vector of ones.

Pooling factor levels.

One particularity of capture-recapture models is the need to address individual levels separately. In Example 1, for the AGE effect, we need to address separately level 1 and levels 2 to 4 combined. This necessary feature is implemented in GEMACO and the general syntax is simply

$$factor(list_1, \ldots, list_k)$$

where $list_i$ is a list of indices.
For Example 1, the phrase

$$S = a(1, 2 : 4) \text{ builds a model where logit} \left(s^{(t,1)}\right) = \beta_1, \text{ logit} \left(s^{(t,a)}\right) = \beta_2 \qquad (6)$$

with $t = 1, \ldots, 4$, $a = 2, \ldots, 4$. In the model definition language, the syntax $2 : 4$ pools the levels 2 to 4.

The traditional operators.

In (5), op_i is an operator. The built-in operators are $(+)$ to include an additional term in a model, $(\cdot)$ the tensor product between two factors, or between a factor and a previously obtained matrix or between two matrices, $(*)$ for the matrix vector product of a factor with a numeric variable. The tensor product gives a model with interaction. For instance, the syntax GROUP.TIME means that the time effects vary independently between groups. For Example 1, the phrase

$$S = i + a + t \text{ builds a model, where logit} \left(s^{(t,a)}\right) = \alpha + \beta_a + \gamma_t \qquad (7)$$

with $t = 1, \ldots, 4$ and $a = 1, \ldots, 4$. At this stage, we can remark that there is a redundancy in the parametrization since the intercept i is in the column space associated to both t and a. The probabilities to remain in states $N1$, $N2$, $N3$ between ages 1 and 5 in Example 3 are built by the phrase

$$\Psi_{i=j+3, j=1,\ldots,3}^{(a=2,\ldots,6)} = a(2\_6).[f(4).to(1) + f(5).to(2) + f(6).to(3)]$$

This builds a submodel, where $\text{logit}(\psi_{ii-3}^{(a)}) = \beta_{i-3,a}$ with $i = 4, \ldots, 6$, $a = 2, \ldots, 6$. The syntax $a(2\_6)$ in the model definition language creates a factor containing the levels 2 to 6 of AGE; it is equivalent to $a(2, 3, 4, 5, 6)$.

Partial answer to aliasing.

Models $a + t$ and $i + a + t$ in (7) are formally equivalent because the two matrices generated by the phrases $a + t$ and $i + a + t$ span the same linear space. However, some columns are redundant in the direct matrix representation. This kind of redundancy is generally called aliasing. Programs

such S-PLUS or R apply different treatments to cope with aliasing in these simple models. To deal with this overparameterization, GEMACO automatically imposes the constraint $\beta_1 = 0$ where $(\beta_t)_{t=1,\ldots,K-1}$ are the main effects for the factor TIME. In that treatment (see Venables & Ripley 2002), the first column of t is removed in the model $a + t$ and the first columns of both a and t are removed in the model $i + a + t$. This procedure is based on the detection of the INTERCEPT in each of two terms added together (here a and t). However we will see in the advanced features of Section 3.3 that it is only a partial answer to aliasing. Now, after partial treatment of aliasing, (7) becomes:

$$S = i + a + t \text{ builds a model where logit } (s^{(t,a)}) = \alpha + \beta_a + \gamma_t \tag{8}$$

with $t = 2,\ldots,4$ and $a = 2,\ldots,4$.

Redundancy in the Ψ matrix.

The Ψ matrix is row-stochastic, i.e., the sum of each row is equal to one. Thus, only $N \times (N-1)$ parameters out of the $N \times N$ need to be estimated (see Table 1). One redundant parameter has to be chosen in each row. It is selected by the user in GEMACO. Suppose that the redundant parameters are in the diagonal of the Ψ matrix. Probabilities are modeled with a generalized multinomial logit link function,

$$\log\left(\frac{\psi_{ij}}{\psi_{ii}}\right) = \beta_{ij} \quad \text{for } i = 1\ldots,N,\ j = 1,\ldots,N,\ j \neq i \quad \text{and} \quad \psi_{ii} = 1 - \sum_{j=1,j\neq i}^{N} \psi_{ij}.$$

This is important because the model may change depending on which parameter is chosen to be redundant. Although the set of redundant parameters is still present in the list of biological parameters Θ, they are unconstrained since the corresponding rows in X_2 contain only zeros. GEMACO operates in two steps for unconstrained redundant parameters.

1. During the interpretation of the sentence and the building of X_2, redundancy is not taken into account. Redundant parameters are thus "constrained" like the others as this stage. In Example 3, considering transitions from states $B1$, $B2$, $B3$ to states $B1$, $B2$, $B3$, the phrase interpreted with all the transition (redundant or not)

$$\Psi_{i=1,\ldots,3,j=1,\ldots,3} = f(1\_3).to(1\_3) \text{ builds a sub-model where } \log\left(\frac{\psi_{ij}}{\psi_{ii}}\right) = \beta_{ij}$$

with $i,j = 1,\ldots,3$. This model is not correct at this point.

2. Then a step of correction occurs. Rows corresponding to redundant parameters are set to zero in X_2. After the correction and for ψ_{ii} considered as complementaries of the other transitions, the phrase

$$\Psi_{i=1,\ldots,3,j=1,\ldots,3} = f(1\_3).to(1\_3) \text{ builds a sub-model where } \log\left(\frac{\psi_{ij}}{\psi_{ii}}\right) = \beta_{ij} \tag{9}$$

with $i,j = 1,\ldots,3$ and $i \neq j$. The number of columns in the matrix generated by (9) is 6.

Keyword for numeric variables.

The keyword for numeric variables x represents a vector of values which can replace one of the following factors: TIME, AGE, COHORT, GROUP, FROM, or TO. In a model one could for instance replace the factor TIME by the vector of temperatures at the $K-1$ recapture occasions. The value of x corresponding to redundant parameters must be present.

3.3. Advanced features of the language.

The basic features of the model definition language, described in Section 3.2, are similar to the features of the language used in programs like R and S-PLUS, but with predefined factors specific to multistate capture-recapture. However, this is not enough to cover all the multistate models. In Section 2.3, the true age factor is needed in the model. The syntax in Section 3.2 does not allow us to build it. Furthermore in the same example (see (4)), we need an additive effect between true age greater than 2 and TIME and clearly a sophisticated treatment for aliasing is needed. The partial treatment of the previous section cannot solve the problem because the vector space span by levels 2 to 5 of true age do not contain the INTERCEPT. These reasons led us to develop advanced features for the model definition language.

The operator &.

The direct addressing with lists of levels presented in the previous section allows only for the aggregation of parameters belonging to the same main factor. The aggregation operator & makes the aggregation possible of parameters corresponding to levels of different factors. The syntax is

$$term_1 \ \& \ term_2$$

This syntax sums the matrix for $term_1$ with the matrix for $term_2$. If the numbers of columns are not equal, then the last columns of the largest matrix are kept unchanged.

For Example 2, the phrase

$$
\begin{aligned}
S \ = \ & [g(1).a(1) + g(1).a(2)\&g(2).a(1) + g(1).a(3)\&g(2).a(2)\&g(3).a(1) \\
& + g(2).a(3)\&g(3).a(2,3)]
\end{aligned}
\tag{10}
$$

builds a model where $\text{logit}(s^{(a,g)}) = \beta_{a+g-1}$, with $a = 1, \ldots, 3$ and $g = 1, \ldots, 3$. GEMACO generates a matrix with five columns. The first column identifies the true age 1; the command $g(1).a(1)$ gives a value of 1 to the individuals marked at age 1 at their first recapture occasion. The second column is the true age 2; the command $g(1).a(2)\&g(2).a(1)$ identifies the second recapture occasion of individuals of age 2 marked at age 1 and the first recapture occasion of those marked at age 2. The third column, $g(1).a(3)\&g(2).a(2)\&g(3).a(1)$, is associated with the true age 3, while the last two columns represent the true ages 4 and 5.

Definition of shortcuts.

A complex sentence might need to be manipulated repeatedly. To alleviate this problem, it is possible to replace a piece of a sentence by a shortcut. For instance, we can define the shortcut "trueage" as being equal to (10). Then the shortcut "trueage" can be used in combination with any other factor such as "trueage.t". To avoid memory overloads, the matrix associated with a shortcut is not stored. Each time it is used in the sentence, the shortcut "trueage" is replaced with its definition (10). As a consequence, direct addressing of the levels for "trueage" is not possible.

Addressing levels of shortcuts.

To enable the addressing of shortcut levels, we extended the syntax of GEMACO. The syntax for addressing shortcut levels or of any part of a sentence is:

$$\$\{term\}(list_1, list_2, ...)$$

In Example 2, the shortcut "trueage" as defined by (10) has five levels. We may want to try a model where adults of true ages 2 and more have the same survival. The phrase $S = \${trueage}(1, 2 : 5)$ combines the levels 2 to 5 of "trueage." The new model is $\text{logit}(s^{a=1,g=1}) = \beta_1$ and $\text{logit}(s^{a,g}) = \beta_2$ for $a+g-1 \neq 1$ with $a = 1, \ldots, 3$ and $g = 1, \ldots, 3$.

Dealing with zero probabilities: the keyword others.

The keyword *others* is particularly useful to set many parameters equal to a constant value. To be valid, it must always appear at the end of as sentence

$$model + others$$

as it accounts for all remaining parameters that are not constrained in *model*.

The transition matrices of Example 3 are sparse and using this keyword makes model definition simple. The phrase $f(1\_3).to(1\_3) + others$ builds the 6×6 transition matrix for AGE = 7–10 in Table 1. This is the same model as (9) but with an extra parameter making all the other transitions constant. The extra parameter is fixed by the user to a value such as zero or one in the probability scale, or to a constant value in the transformed scale.

A general treatment for aliasing.

One difficulty for the automatic generation of models is aliasing. Thus the natural question is why bother about aliasing? At least, two main reasons can be put forward:

- Redundancy leads to numerical instabilities.

- Interpretation of mathematical parameters may become difficult and risky.

The user is frequently unaware of redundancy. Dealing with redundancy in a three way table as $t.g+t.a+a.g$ is not trivial even for an advanced user. And as we saw, capture-recapture models can be much more complex. So, there is clearly a need for an automatic treatment. The partial answer to aliasing given in Section 3.2 does not work for $t.g+t.a+a.g$ because of the complex relationship between these three factors.

We present now a more general treatment for aliasing in the form of Algorithm 1. GEMACO deletes redundancy on the design matrix X by a simple variant of the Gram–Schmidt algorithm (see Golub & Van Loan 1996). The Gram–Schmidt algorithm computes the QR factorization of a matrix $X = (x_1, \ldots, x_n)$ with Q an orthogonal basis of X. The partial orthogonal basis of X is built step by step. We used this property to detect redundancy and to eliminate the redundant columns. If at a step k of the orthogonalization (Algorithm 1, step 3:), x_k is contained in the vector space spanned by the first $k - 1$ columns of X (Algorithm 1, step 4:) then x_k is redundant and is rejected (Algorithm 1, step 6:).

Algorithm 1 Delete redundancy in the $r \times s$ matrix $X = (x_1, \ldots, x_s)$

1: $k = 1; l = 0$ {initialisation}
2: at step k, $Q_l = (q_1, \ldots, q_l)$ an orthogonal basis of $span(x_1, \ldots, x_{k-1})$
3: {compute the orthogonal complement of x_k to $span(Q_l)$} $q_{l+1} = x_k - \sum_{m=1}^{l} \langle x_k, q_m \rangle q_m$
4: $r_k = \|q_{l+1}\|_2$
5: **if** $r_k = 0$ **then**
6:     set x_k equal to the null vector
7: **else**
8:     $q_{l+1} = \frac{q_{l+1}}{r_k}$, {normalization}
9:     l=l+1
10: **end if**
11: $k = k + 1$
12: **if** $k <= s$ **then**
13:     goto 2:
14: **end if**
15: remove all the null vectors from X

If we use $\widetilde{X}$ to denote the matrix resulting from the application of Algorithm 1 to X, then $\widetilde{X}$ and X generate the same linear space (i.e., $span(\widetilde{X}) = span(X)$). Depending on the sentence and/or because of the automatic treatment of redundancy by Algorithm 1, the nature of the vector β may change without control. Column indices that are deleted in $X^{(1)}$ associated with $t.g+t.a+a.g$ are

not the same in $X^{(2)}$ associated with the yet equivalent model $a.g+t.a+t.g$. A question then arises naturally. For two equivalent sentences, are the two models equivalent after the treatment of redundancy in Algorithm 1? As $\mathrm{span}(\widetilde{X}^{(1)}) = \mathrm{span}(\widetilde{X}^{(2)})$ and since these matrices are of full rank, there exists a square non-singular matrix A such that $\widetilde{X}^{(1)} = \widetilde{X}^{(2)}A$. Furthermore, we have the following proposition.

PROPOSITION 1. *Let $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ be the unique maximum likelihood estimator of the following models:*
Model 1: $f(\widehat{\Theta}^{(1)}) = X^{(1)}\hat{\beta}^{(1)}$
Model 2: $f(\widehat{\Theta}^{(2)}) = X^{(2)}\hat{\beta}^{(2)}$
where the inverse link function f^{-1} is well defined. If $X^{(1)} = X^{(2)}A$ with A non-singular, then models 1 and 2 are equivalent, that is $\widehat{\Theta}^{(1)} = \widehat{\Theta}^{(2)}$.

This result, whose proof is given in the Appendix, demonstrates the independence of Algorithm 1 over the syntax of the model. Now, we go back to the examples.

Example 1 revisited. Following (6), the model described in Section 2.2 can be easily built with the following sentence
$$S = a(1, 2 : 4) + t$$

Example 2 revisited. We have already written half of the model described in Section 2.3 with the definition of *trueage* as (10). Thus with GEMACO, we can build the model (4) with the following phrase:
$$S = \${trueage}(1) + \${trueage}(2 : 5).[\${trueage}(2\_5) + t]$$

During the building of X_1 in equation (3) with the previous phrase, Algorithm 1 automatically deletes the first level of TIME.

Example 3 revisited. To help design the process of transitions, we defined the three following shortcuts *BtoB, NtoN, NtoB*:

- *BtoB* replaces *f(1_3).to(1_3)* to represent the breeder to breeder transitions.

- *NtoN* replaces *f(4_6).to(4_6)* to represent the non-breeder to non-breeder transitions.

- *NtoB* replaces *[f(4).to(1)+f(5).to(2)+f(6).to(3)]* to represent the non-breeder to breeder transitions.

Then, we build the transition design matrix(X_2) in equation (3) corresponding to transitions given in Table 1. The sentence for the transition, cut into three pieces to help us understand it, is:
$$\begin{aligned} \Psi \;=\;& BtoB + a(1).NtoN + a(2\_5).NtoB + \\ & a(6 : 10).\${NtoB}(1 : 3) + \\ & others. \end{aligned}$$

This sentence is composed of three phrases. The first gives the transition parameters that need to be estimated, the second phrase lists the transition probability that is equal to 1 (in Table 1, at AGE=6, transitions from Ni to Bi occur with probability one for $i = 1, 2, 3$), and the third phrase lists constants which have to be fixed to zero in the probability scale. For fixed values in the probability scale, M-SURGE uses the identity link rather than the original logit link that is used for the mathematical parameters. Thanks to shortcuts *BtoB, NtoN, NtoB*, this sentence is very short, even though the model is complex.

4. BEYOND CURRENT PARAMETERIZATIONS

The separate parametrization noted (S, Ψ, P) permits constraining the parameters for survival and transition independently. But other parameterizations are possible in multistate models, where survival-transition matrix elements are functions of biological parameters.

This section presents two examples where there is a need for a more general model building tool than the one outlined here. It must be able to consider new parameterizations. Building an enlarged class of models can be done in two steps. First, the general survival-transition matrix is decomposed into several elementary matrices of transitions (each elementary matrix correspond to a type of parameter). The second step applies GEMACO to each elementary transition matrix. In these two examples, the survival-transitions matrices cannot be decomposed as equation (2), and as a consequence their parameters cannot be constrained under the conditional Arnason–Schwarz model. Still an alternative decomposition is possible in both cases, if a wider class of components is considered than those of the basic conditional Arnason–Schwarz model.

4.1. A version of the Arnason–Schwarz model incorporating site fidelity.

We first consider a version of the Arnason–Schwarz model in which the probability of transition conditional on survival is further subdivided into a probability of leaving the site (the complement of site fidelity noted f), and a probability of moving to another site conditional on leaving (see Grosbois & Tavecchia 2003).

$$
\Phi = \begin{pmatrix}
s_1 f_1 & s_1(1-f_1)\psi_{12} & s_1(1-f_1)(1-\psi_{12}) \\
s_2(1-f_2)\psi_{21} & s_2 f_2 & s_2(1-f_2)(1-\psi_{21}) \\
s_3(1-f_3)\psi_{31} & s_3(1-f_3)(1-\psi_{31}) & s_3 f_3
\end{pmatrix}.
$$

The problem is how to constrain fidelity (f) and settlement (ψ) separately. The current implementation given in Grosbois & Tavecchia (2003) is very technical and limited to three states. It uses dummy occasions. In the classical separate formulation of the Arnason–Schwarz model, the set of possible states for an individual is the same for both elementary matrices (survival and transition conditional on survival). However, we will show that relaxing this constraint by allowing a different set of states at each of the elementary steps, enables a decomposition with three matrices. Let the sets of states be

$$
\begin{aligned}
\mathbb{E}^{(0)} &= \{\text{site 1, site 2, site 3}\} \\
\mathbb{E}^{(1)} &= \{\text{site 1, site 2, site 3}\} \\
\mathbb{E}^{(2)} &= \{\text{staying in 1, leaving 1, staying in 2, leaving 2,} \\
&\qquad \text{staying in 3, leaving 3}\}
\end{aligned}
$$

We define the following three matrices for survival, fidelity and settlement. A first elementary transition matrix for survival maps $\mathbb{E}^{(0)}$ into $\mathbb{E}^{(1)}$. It is of dimension 3×3:

$$
\text{diag}\,(S) = \begin{pmatrix}
s_1 & 0 & 0 \\
0 & s_2 & 0 \\
0 & 0 & s_3
\end{pmatrix}.
$$

The second elementary matrix for site fidelity given survival maps $\mathbb{E}^{(1)}$ into $\mathbb{E}^{(2)}$. It is of dimension (3×6). Let f_i be the probability of remaining in site i given survival, we have

$$
F = \begin{pmatrix}
f_1 & 1-f_1 & 0 & 0 & 0 & 0 \\
0 & 0 & f_2 & 1-f_2 & 0 & 0 \\
0 & 0 & 0 & 0 & f_3 & 1-f_3
\end{pmatrix}.
$$

The third elementary matrix for movement conditional on emigration maps $\mathbb{E}^{(2)}$ back to $\mathbb{E}^{(0)}$. It is of dimension 6×3:

$$\Psi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \psi_{12} & 1 - \psi_{12} \\ 0 & 1 & 0 \\ \psi_{21} & 0 & 1 - \psi_{21} \\ 0 & 0 & 1 \\ \psi_{31} & 1 - \psi_{31} & 0 \end{pmatrix}.$$

We can easily show that

$$\Phi = \text{diag}\,(S) \times F \times \Psi.$$

So, this decomposition allows us to constrain survival, fidelity and settlement independently. Thus with an appropriate tool to decompose Φ, we would be able to apply GEMACO separately to these three components of the model.

4.2. A mixture of multistate captures and recoveries.

We consider a standard model for mixture of information of recapture on 2 sites with recoveries elsewhere. The survival-transition matrix with the 3 states {"Alive in site 1," "Alive in site 2," "Newly dead"} is

$$\Phi = \begin{pmatrix} s_1 \psi_{11} & s_1 (1 - \psi_{11}) & 1 - s_1 \\ s_2 (1 - \psi_{22}) & s_2 \psi_{22} & 1 - s_2 \\ 0 & 0 & 0 \end{pmatrix}.$$

The problem is how to constrain survival and transition separately. To model parameters s and ψ separately, we introduce a stochastic matrix for survival instead of a vector. We then have $\Phi = S \times \Psi$, where

$$S = \begin{pmatrix} s_1 & 0 & 1 - s_1 & 0 \\ 0 & s_2 & 1 - s_2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\Psi = \begin{pmatrix} \psi_{11} & (1 - \psi_{11}) & 0 & 0 \\ (1 - \psi_{22}) & \psi_{22} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and S and Ψ are transition matrices between the states {"Alive in site 1," "Alive in site 2," "Newly dead," "Dead"}.

5. CONCLUSION

We have seen that GEMACO allows the automatic generation of multistate models. This tool solves several difficult problems:

1. The generation of the set of multistate models.

2. The automatisation of the treatment of aliasing.

3. The simplification of the model syntax by allowing the definition of shortcuts by the user. The levels of these shortcuts can now be addressed directly.

Several improvements can be considered in the future. First, we plan to develop a generator of shortcuts. It would be very useful to generate automatically keywords specific to the application under consideration. For example, the closed robust design models discussed by Kendall, Pollock & Brownie (1995) can be written as multistate models. Keywords appropriate for the robust design would allow for an easy implementation of these models. Second, a possible extension is to permit the constraint that two parameters of different kinds be equal; such an application can be found in Lebreton, Morgan, Pradel & Freeman (1995). Third, we think also that this tool can be further improved by allowing the definition of new formulations. We gave two examples in Section 4, and some others exist in multistate (see Conn, Kendall & Samuel 2004; Schmidt, Feldmann & Schaub 2005), but also potentially much more so in multievent models as defined in Pradel (2005). So, we would like to go beyond classical combined and separate parameterizations by permitting further decompositions.

Many of these extensions are feasible using a new tool that is under development (see Choquet, Rouan & Pradel 2008) which can be combined with GEMACO. This tool will allow one to generate multievent models (Pradel 2005) and would be a generalization of multistate models. Finally, we think that models with fixed and random effects applied to capture-recapture could also be generated in a similar way.

APPENDIX

Proof of Proposition 1. By definition and because f is invertible, $L(\Theta^{(1)}) = L \circ f^{-1}(X^{(1)}\beta^{(1)})$ and the maximum likelihood estimator $\widehat{\Theta}^{(1)}$ satisfies:

$$
\begin{aligned}
L(\hat{\Theta}^{(1)}) &= \min_{\beta} L \circ f^{-1}(X^{(1)}\beta) = \min_{\beta} L \circ f^{-1}(X^{(1)}AA^{-1}\beta) \\
&= \min_{\tilde{\beta}} L \circ f^{-1}(X^{(2)}\tilde{\beta}) = L(\widehat{\Theta}^{(2)}).
\end{aligned}
$$

As L is unimodal, $\widehat{\Theta}^{(1)} = \widehat{\Theta}^{(2)}$.

ACKNOWLEDGEMENTS

REFERENCES

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

C. Brownie, J. Hines, J. Nichols, K. Pollock & J. Hestbeck (1993). Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics*, 49, 1173–1187.

K. Burnham & D. Anderson (2002). *Model Selection and Inference: A Practical Information-theoretic Approach*, 2nd edition. Springer, New York.

J. Chambers & T. Hastie (1992). *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove, California.

R. Choquet, A.-M. Reboulet, R. Pradel, O. Gimenez & J.-D. Lebreton (2004). M-SURGE: New software specifically designed for multistate recapture models. *Animal Biodiversity and Conservation*, 27, 177–185.

R. Choquet, L. Rouan & R. Pradel (2008). Program E-SURGE: A software application for fitting multievent models. *Environmental and Ecological Statistics*, in press.

P. Conn, W. Kendall & M. Samuel (2004). A general model for the analysis of mark-resight mark-recapture, and band-recovery data under tag loss. *Biometrics*, 60, 900–909.

L. Crespin, M.-P. Harris, J.-D. Lebreton, M. Frederiksen & S. Wanless (2006). Recruitment to a seabird population depends on environmental factors and on population size. *Journal of Avian Biology*, 75, 228–238.

C. Donnelly & R. Stallman (1994). *Bison: The Yacc-Compatible Parser Generator.* Free Software Foundation, Cambridge, Massachusetts: http://dinosaur.compilertools.net/bison/index.html

G. Golub & C. Van Loan (1996). *Matrix Computations*, 3rd edition. The Johns Hopkins University Press, Baltimore.

V. Grosbois & G. Tavecchia (2003). Modeling dispersal with capture-recapture data: Disentangling decisions of leaving and settlement. *Ecology*, 84, 1225–1236.

V. Henaux, T. Bregnballe & J.-D. Lebreton (2007). Dispersal and recruitment during population growth in a colonial bird, the great Cormorant *Phalacrocorax Carbo Sinensis. Journal of Avian Biology*, 38, 44–57.

R. Ihaka & R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.

W. L. Kendall, K. H. Pollock & C. Brownie (1995). A likelihood-based approach to capture-recapture estimation of demographic parameters under the robust design. *Biometrics*, 51, 293–308.

J.-D. Lebreton, T. Almeras & R. Pradel (1999). Competing events, mixtures of information and multistratum recapture models. *Bird Study*, 46, 39–46.

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

J.-D. Lebreton, J. Hines, R. Pradel, J. Nichols & J. Spendelow (2003). Estimation by capture-recapture of recruitment and dispersal over several sites. *Oikos*, 101, 253–264.

J.-D. Lebreton, B. J. T. Morgan, R. Pradel & S. N. Freeman (1995). A simultaneous survival rate analysis of dead recovery and live recapture data. *Biometrics*, 51, 1418–1428.

J.-D. Lebreton & R. Pradel (1991). *User's Manual for Program SURGE (version 4.1).* Technical report, Center for Evolutionary and Functional Ecology, Montpellier, France.

A. Loison, M. Festa-Bianchet, J.-M. Gaillard, J. T. Jorgenson & J. M. Jullien (1999). Age-specific survival in five populations of ungulates: Evidence of senescence. *Oikos*, 80, 2539–2554.

P. McCullagh & J. Nelder (1989). *Generalized Linear Models.* Chapman & Hall, New York.

V. Paxson (1994). *Flex: A Fast Scanner Generator.* Free Software Foundation, Cambridge, Massachusetts.

R. Pradel (2005). Multievent: An extension of multistate capture-recapture models to uncertain states. *Biometrics*, 61, 442–447.

R. Pradel, A.-R. Johnson, A. Viallefont, R.-G. Nager & F. Cezilly (1997). Local recruitment in the greater Flamingo: A new approach using capture-mark-recapture data. *Ecology*, 78, 1431–1445.

R. Pradel, C. M. A. Wintrebert & O. Gimenez (2003). A proposal for a goodness-of-fit test to the Arnason–Schwarz multisite capture-recapture model. *Biometrics*, 59, 43–53.

P. Rivalan, A.-C. Prevot-Julliard, R. Choquet, R. Pradel, B. Jacquemin & M. Girondot (2005). Trade-off between current reproductive effort and delay to next reproduction in the leatherback sea Turtle. *Oecologia*, 145, 564–574.

B. Schmidt, R. Feldmann & M. Schaub (2005). Demographic processes underlying population growth and decline in *Salamandre Salamandre. Conservation Biology*, 19, 1149–1156.

W. Venables & B. Ripley (2002). *Modern Applied Statistics with S*, 4th edition. Springer, New York.

G. White & K. Burnham (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study*, 46 (supplement), 120–139.

G. Wilkinson & C. Rogers (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society Series C*, 22, 392–399.

Rémi CHOQUET: remi.choquet@cefe.cnrs.fr
Centre d'écologie fonctionnelle et évolutive
Centre National de la Recherche Scientifique
1919, route de Mende, FR-34293 Montpellier cedex 5, France