Available online at www.sciencedirect.com
ScienceDirect
Mathematical Biosciences 210 (2007) 508–522
Mathematical Biosciences
www.elsevier.com/locate/mbs

Model-based estimation of the link between the daily survival probability and a time-varying covariate, application to mosquitofish survival data

Samuel Soubeyrand [a,*], Rémy Beaudouin [b], Nicolas Desassis [a], Gilles Monod [b]

[a] *INRA, Unité Biostatistique et Processus Spatiaux, F-84914 Avignon, France*
[b] *INRA, Unité SCRIBE, F-35042 Rennes, France*

Received 7 December 2006; received in revised form 8 June 2007; accepted 22 June 2007
Available online 13 July 2007

Abstract

The survival probability in a group of individuals may evolve in time due to the influence of a time-varying covariate. In this paper we present a model-based approach allowing the estimation of the functional link between the survival probability and a time-varying covariate when data are grouped and time-period censored. The approach is based on an underlying model consisting in non-stationary Markov processes and describing the survival of individuals. The underlying model is aggregated in time and at the group level to handle the group structure of data and the censoring. The aggregation yields a generalized non-linear mixed model. Then, a Bayesian procedure allows the estimation of the model parameters and the description of the link between the survival probability and the time-varying covariate

This approach is applied in order to explore the relationship between the daily survival probability of mosquitofish (*Gambusia holbrooki*) and their time-varying lengths (small mosquitofish die with a higher rate than large ones because they are more affected by predation, cannibalism and environmental stress).

Keywords: Bayesian estimation; Generalized non-linear mixed model; Non-stationary Markov process; Survival analysis; Time-period censoring

* Corresponding author.
E-mail address: samuel.soubeyrand@avignon.inra.fr (S. Soubeyrand).

0025-5564/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2007.06.005

1. Introduction

1.1. Survival probability and time-varying covariate

The daily survival probability (DSP) is defined as the conditional probability that an individual survives at the day of interest given it was alive the day before. The DSP is one minus the so-called hazard probability in discrete-time survival analysis [2]. Note that the content of this paper applies for other time scales: one can consider survival probabilities for a second, a week, or a year. Moreover, one can consider, instead of the dichotomy alive/dead, dichotomies such as healthy/infected, student/worker, or not parent/parent. Here, we consider the daily time scale and the dichotomy alive/dead because of the application.

Depending on the context, the DSP can depend on time-varying (or time-dependent) covariates; in such a case the DSP is non-stationary in time. For example, consider the outbreak of the severe acute respiratory syndrome (SARS) in 2003, the DSP within the population of SARS in-patients was expected to be positively influenced by improvements, with time, of the knowledge about the disease and the management of the authorities [28,29]. Other examples arise in prey-predator systems. Indeed, the survival probability of preys may be affected for instance by (i) the time-varying count of predators, (ii) the time-varying count of preys, (iii) the time-varying size of preys. In the case of infectious diseases, the probability for an individual to stay healthy may be affected by the count of infectious individuals at the date before (autoregressive process). In social science, the probability for an individual to stay a student may be affected by the time-varying policy of the government.

1.2. Grouped and time-period censored data

In order to learn about the relationship between a DSP and a time-varying covariate, one can collect data related to the phenomenon of interest and perform a statistical analysis of these data. But collecting daily data and repeating the measures a large number of days may not be possible. Indeed, each measure may be time and money consuming, and may affect the survival of individuals. That is the reason why, instead of daily measures, one can choose to carry out for several groups of individuals a few measures separated by several days, e.g. one initial measure plus a few additional measures. To get data corresponding to a large range of the time-varying covariate, the groups have to be sampled for different ranges of the covariate. Such a structure for the data generates two problems: (i) the observations are time-period (or interval) censored and (ii) there may be group effects. These two points should be taken into account when a statistical analysis of the data is performed.

Some of the models which have been developed in order to explore the link between survival of individuals and time-varying covariates account for time-period censoring; see for instance [8,17,19,20,23] when time is discrete and [6,13,21] when it is continuous. Among the references adapted to discrete time, some ([17,20,23]) also deal with group effects.

1.3. Proposed approach

In this paper we base our approach on a model which can be viewed as an extension of the one proposed in [23, chapter 3] in the nest survival context, and we combine it with a Bayesian estimation procedure.

The model was built by deriving a model at the individual level and day scale: daily survival processes of individuals were modeled as non-stationary Markov processes depending on the DSP; then the survival processes were aggregated at the group level and in time to obtain, for each group, the conditional distribution of the count of surviving individuals at time t_2 given the count of surviving individuals at time $t_1 < t_2$. The DSP was modeled as a function of (i) a random group effect, constant with time, and (ii) the time-varying covariate. The resulting model for the count of surviving individuals conditional on the count of surviving individuals at the previous observation day is a generalized non-linear mixed model [7,14,25].

This model was used to study the variations in the DSP of mosquitofish (*Gambusia holbrooki*) with fish length: small mosquitofish die with a higher rate than large ones because they are more affected by predation, cannibalism and environmental stress [1]. Accurate quantitative knowledge about the link between the DSP and the length of mosquitofish is needed (i) for improving the knowledge about life history traits of mosquitofish and, more specifically, (ii) for building an individual-based model describing the dynamics of mosquitofish populations in experimental ecosystems (mesocosms). This model, still under construction in research units of the National Institute for Agricultural Research (INRA, France), is being developed to carry out studies in ecotoxicology (mosquitofish is an experimental animal commonly used in ecotoxicology; see for instance [5,9,10,15,24]). A first version of the individual-based model, presented in [12], describes at a daily time step the early stages of the establishment of a mosquitofish population. In that earlier version, the link between fish length and fish survival was modeled such that below a fixed threshold a fish had constant DSP π_1 and above the threshold a fish had constant DSP $\pi_2 > \pi_1$.

A more realistic description of variations in the DSP with fish length is needed in order to improve the individual-based model in its realism. We precisely propose such a description in this article by performing a statistical analysis of experimental data using the model described in the previous paragraph. As we will see in this paper, the accompanying Bayesian procedure allows us to give strong insight into the relationship between the fish DSP and the fish length.

Section 2 presents the grouped and time-period censored data which were collected for the study of the relationship between mosquitofish survival and fish length. Presenting the motivating data allows us to highlight the problems occurring in the statistical analysis when the data are grouped and censored. The survival model and the estimation procedure are detailed in Section 3. The methodology is applied to the mosquitofish survival data set in Section 4. The paper is ended by a discussion in Section 5.

2. Motivating data

2.1. Experiments

Survival and mean length were measured for $I = 38$ groups of mosquitofish introduced in mesocosms settled in the years 2002, 2003, 2004 and 2005 in the Agrocampus of Rennes, Brittany, France (48N/1.7W). The mesocosms consisted of circular metal tanks (3.0 m diameter, 0.6 m height) filled with various elements for building fish habitats. In January, the tanks were lined up with polyethylene film (0.2 mm thick), filled with tap water (water depth: 45 cm), and sediments were introduced at a uniform layer (about 5 cm thick). The tanks then sat for a few months

to allow for phyto- and zoo-plankton as well as macroinvertebrates to naturally settle in the tanks. In April, pieces of *Ludwigia peploide* were introduced into the mesocosms. Then, at different times in the summer, groups of mosquitofish were introduced in the mesocosms, one group per mesocosm, and nets with small mesh were installed to protect the fish from bird predation.

The initial sizes of the groups of mosquitofish varied from 5 to 150 individuals, and the initial mean lengths of fish varied from 6.6 to 29.7 mm (see Table 1). In each group, individuals had the same age and thermal history, so their lengths were about the same. For 6 groups ($i \geqslant 33$ in Table 1), fish were captured, counted, measured for length and released in the mesocosms two or three times. Such an experimental procedure is particularly time-consuming because, as mosquitofish survival can be affected by the stress of captures, fish must be captured and manipulated with care. That is the reason why, for 32 other groups ($i \leqslant 32$ in Table 1), fish were captured, counted, measured for length only one time after their introduction in the mesocosms. Forty six captures were done in total for the 38 groups of fish. For each capture we know: the mesocosm and the duration in days since the last capture (or fish introduction), the initial and final fish counts, and the initial and final mean fish lengths (see Table 1).

2.2. Time-period-censored observations

If daily data would have been obtained for each mesocosm i, i.e. if the counts $N_{i,t}$ and the average lengths $\bar{X}_{i,t}$ of the surviving fish would have been measured at times $t_i, \ldots, t_i + q_i$, then we could have estimated the DSP $\pi_{i,t}$ in mesocosm i at day $t = t_i + 1, \ldots, t_i + q_i$ by

$$\hat{\pi}_{i,t} = \frac{N_{i,t}}{N_{i,t-1}}.$$

Then, the functional link between the DSP and the expected length of fish could be inferred, for example, by fitting a regression curve [18] to scatterplot $\mathcal{S}_0 = \{(\bar{X}_{i,t-1}, \hat{\pi}_{i,t}) : i = 1, \ldots, I, t = t_i + 1, \ldots, t_i + q_i\}$.

Unfortunately, the procedure for counting fish and measuring their lengths has many difficulties. First, fish were captured by using fish traps, and capturing all the fish of a mesocosm can take a whole day. Second, the fish can hardly support the stress caused by frequent captures and manipulations. Therefore, the data were obtained for each mesocosm i at days $t_{i,1} < \cdots < t_{i,K_i}$ where $t_{i,1}$ represents the day of fish introduction into the mesocosm. The number of captures, K_i, was often two and no more than four. The time between observations varied from 8 to 127 days. Thus, the observations of the survival process $N_i = \{N_{i,t}: t = 0,1,2,\ldots\}$ and the time-varying covariate $\bar{X}_i = \{\bar{X}_{i,t} : t = 0, 1, 2, \ldots\}$ were time-period censored.

Despite the censoring, we could think about adapting the method proposed in the case of daily observations by making the two following assumptions. First, assume that the DSP is constant between consecutive observation days, then the fish DSP in mesocosm i at day t between the k-th and $(k+1)$-th captures, which occur, respectively, at times $t_{i,k}$ and $t_{i,k+1}$ ($k < K_i$), can be estimated by

$$\tilde{\pi}_{i,t} = \left(\frac{N_{i,t_{i,k+1}}}{N_{i,t_{i,k}}}\right)^{\frac{1}{t_{i,k+1} - t_{i,k}}}.$$

Table 1
Mosquitofish survival data set

Id.	i	N_0	N_1	$\bar{X}_0$	$\bar{X}_1$	Δ_t
1	1	6	6	27.8	39.4	45
2	2	5	5	27.8	47.2	79
3	3	121	119	7.8	22.1	45
4	4	88	84	7.9	25.4	31
5	5	150	127	8.2	20.0	56
6	6	104	67	7.6	23.0	67
7	7	6	4	27.8	43.5	127
8	8	7	6	27.8	45.2	127
9	9	6	6	27.8	45.5	127
10	10	6	6	27.8	46.2	127
11	11	6	6	27.8	42.8	127
12	12	6	6	27.8	44.3	127
13	13	6	5	27.8	43.0	127
14	14	6	6	27.8	42.5	127
15	15	6	6	27.8	44.3	127
16	16	6	5	27.8	44.0	127
17	17	6	6	29.7	46.3	112
18	18	6	5	29.2	43.4	112
19	19	6	5	29.2	44.2	112
20	20	6	5	26.5	44.6	114
21	21	6	6	25.3	46.2	114
22	22	6	5	26.3	44.6	114
23	23	6	5	25.3	41.2	114
24	24	6	5	25.5	43.8	114
25	25	6	5	24.7	42.8	114
26	26	6	4	25.3	42.8	114
27	27	6	6	24.7	45.2	114
28	28	6	6	26.8	43.2	114
29	29	6	6	25.0	45.2	114
30	30	6	5	26.0	43.0	114
31	31	6	6	24.3	44.2	114
32	32	23	23	19.3	21.8	43
33	33	34	34	8.2	17.8	12
34	33	34	34	17.8	20.8	15
35	34	30	29	6.6	10.9	8
36	34	29	28	10.9	20.2	21
37	34	28	26	20.2	22.3	43
38	35	37	36	8.5	13.5	9
39	35	36	36	13.5	27.8	26
40	36	30	23	6.6	9.6	8
41	36	23	23	9.6	19.3	21
42	37	35	29	6.9	13.1	11
43	37	29	28	13.1	25.8	21
44	38	35	24	6.7	9.7	8
45	38	24	21	9.7	19.2	21
46	38	21	17	19.2	22.1	43

Id., identification number; i, mesocosm; N_0 and N_1, initial and final counts of surviving fish; $\bar{X}_0$ and $\bar{X}_1$, initial and final mean lengths of surviving fish; Δ_t, duration in days separating the initial and final measurements.

Second, assume that the fish length increases linearly between consecutive captures, then the expected length $\mathbb{E}(X_{i,t})$ for fish in mesocosm i at day $t \in \{t_{i,k}, \ldots, t_{i,k+1}\}$ can be estimated by

$$\tilde{X}_{i,t} = \bar{X}_{i,t_{i,k}} + (t - t_{i,k}) \frac{\bar{X}_{i,t_{i,k+1}} - \bar{X}_{i,t_{i,k}}}{t_{i,k+1} - t_{i,k}}$$

(a more realistic growth assumption will be made in Section 3.4). Using this material, we can plot the fish DSP against the mean fish length: Fig. 1 shows scatterplot $\mathcal{S}_1 = \{(\tilde{X}_{i,t-1}, \tilde{\pi}_{i,t}) : i = 1, \ldots, I, t = t_i + 1, \ldots, t_{i,K_i}\}$ (dots). It also shows scatterplot $\mathcal{S}_2$ (circles) derived from $\mathcal{S}_1$ by only keeping, for each couple (i,k), the point $(\tilde{X}_{i,t-1}, \tilde{\pi}_{i,t})$ for which t is the median day of $\{t_{i,k-1}, \ldots, t_{i,k}\}$. The estimation of the functional link between the DSP and the expected length of fish could be done, as in the case of daily observations, by fitting a regression curve to scatterplot $\mathcal{S}_1$ or $\mathcal{S}_2$.

2.3. Mesocosm dependence and mesocosm effect

However, fitting a regression curve to scatterplots $\mathcal{S}_1$ or $\mathcal{S}_2$, as well as $\mathcal{S}_0$ in the hypothetical case of daily observations, is problematic since there exists a temporal dependence between points corresponding to the same mesocosm.

In addition, the daily survival probability of fish might be submitted to a mesocosm effect. Indeed, the survival of fish is known to be affected, by the densities of predators and vegetation (filamentous algae principally) [1,26]. These densities were precisely varying from mesocosm to mesocosm in our experiments since the settlement process of predators and vegetation was natural. Consequently, motivated by building a realistic simulation model for the mosquitofish life, we want to characterize the variability of the DSP due to the possible mesocosm effect.

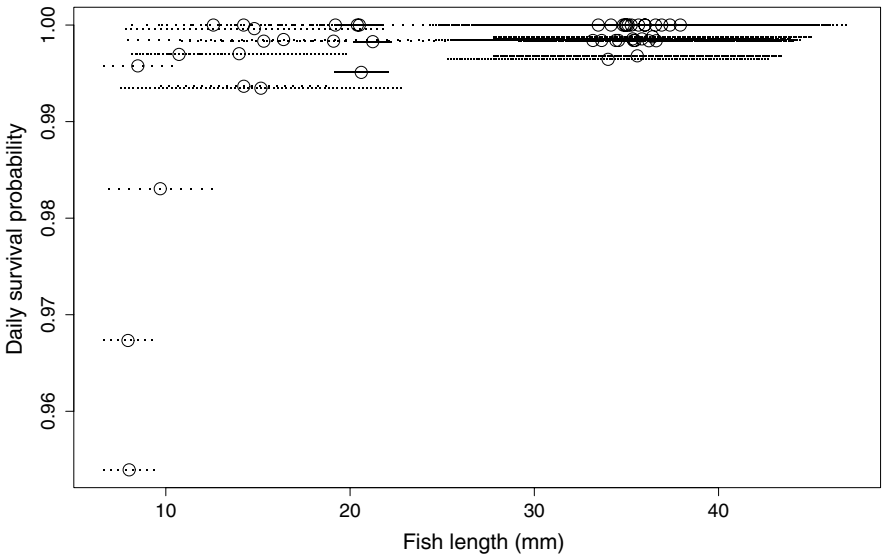

Fig. 1. Two representations of the fish DSP against the mean fish length. Dots, scatterplot $\mathcal{S}_1$; circles, scatterplot $\mathcal{S}_2$.

3. Model and estimation method

We now describe the construction of the model and the estimation method which allowed us to analyze the data set detailed above. Underlying assumptions are labelled A_1, A_2, $A_{3,a-c}$, A_4. The model and the estimation method allow the handling of time-period censoring and group effects. In the following the context of fish survival is used to present the model; nevertheless, as mentioned in the introduction, the model can be applied in other contexts.

3.1. Underlying model at the individual and day scale

Let $S_i = \{S_{i,t}: t = 0,1,\ldots\}$ be the survival process of any fish in mesocosm i: at day 0 the fish is alive and $S_{i,0} = 1$, then $S_{i,t} = 1$ as long as the fish is alive and 0 afterward. The process S_i is assumed to be (A_1) a non-stationary Markov process with transition probabilities at day $t = 1,2,\ldots$

$$\begin{aligned} &\mathbb{P}[S_{i,t} = 1|S_{i,t-1} = 1] = \pi_{i,t} \quad \mathbb{P}[S_{i,t} = 0|S_{i,t-1} = 1] = 1 - \pi_{i,t} \\ &\mathbb{P}[S_{i,t} = 0|S_{i,t-1} = 0] = 1 \qquad \mathbb{P}[S_{i,t} = 1|S_{i,t-1} = 0] = 0, \end{aligned} \tag{1}$$

where $\pi_{i,t}$ is the survival probability for any fish in mesocosm i between days $t-1$ and t; it is called the daily survival probability (DSP) in ecology [4]. Note that 1-$\pi_{i,t}$ is the so-called hazard probability in discrete-time survival analysis [2].

Remark. $\pi_{i,t}$ does not depend on the fish but only on the mesocosm where the fish is and the day; so, fish in a given mesocosm are submitted to the same DSP.

3.2. Model for the count of surviving fish in a mesocosm

Let us now determine the conditional distribution of the count N_{i,t_2} of surviving fish in mesocosm i at day t_2 given the count N_{i,t_1} of surviving fish in the mesocosm at day $t_1 < t_2$.

From the Markovian assumption A_1, S_{i,t_2} given S_{i,t_1} follows a Bernoulli distribution whose probability is the conditional probability that a fish in mesocosm i is alive at day t_2 given it was surviving at day t_1. This probability is given by

$$\mathbb{P}[S_{i,t_2} = 1|S_{i,t_1} = 1] = \prod_{t=t_1+1}^{t_2} \pi_{i,t}.$$

Assuming that (A_2) survival processes of fish are independent, N_{i,t_2} given N_{i,t_1} follows a binomial distribution with size N_{i,t_1} and probability $\prod_{t=t_1+1}^{t_2}\pi_{i,t}$

$$N_{i,t_2}|N_{i,t_1} \sim \text{Binomial}\left(N_{i,t_1}, \prod_{t=t_1+1}^{t_2} \pi_{i,t}\right). \tag{2}$$

3.3. Model for the daily survival probability

The DSP $\pi_{i,t}$ is modeled as (assumption $A_{3,a}$) a function of the expected length $\mathbb{E}(X_{i,t-1})$ of fish in mesocosm i at day $t-1$

$$\begin{aligned}\pi_{i,t} &= \pi\{\alpha_i, \beta, \mathbb{E}(X_{i,t-1})\} \\ &= \text{logit}^{-1}[\alpha_i + f_\beta\{\mathbb{E}(X_{i,t-1})\}],\end{aligned} \tag{3}$$

where α_i is a real parameter reflecting survival features in mesocosm i and f_β is a real-valued function parameterized by a vector β.

The observed mesocosms are considered as elements of a population and we are interested in characterizing this population. Consequently, $\alpha_1, \ldots, \alpha_I$ are viewed as random effects [14]. They are modeled as (assumption $A_{3,b}$) independent normal variables with mean 0 and standard deviation σ.

The function f_β is assumed ($A_{3,c}$) to take the form

$$f_\beta(x) = \beta_1 + \beta_2 x^{\beta_3}, \tag{4}$$

where $\beta = (\beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is in $\mathbb{R}^3$. Note that when β_2 is zero, then β_3 is no more identifiable and vice-versa. Using Eq. (4) for f_β provides a flexible form for the DSP in Eq. (3).

Thus, the model for the counts of surviving fish is a random-effects model, and Eq. (2) becomes

$$N_{i,t_2}|N_{i,t_1}, \alpha_i \sim \text{Binomial}\left(N_{i,t_1}, \prod_{t=t_1}^{t_2-1} \pi\{\alpha_i, \beta, \mathbb{E}(X_{i,t})\}\right). \tag{5}$$

This model is a generalized non-linear mixed model [7,14,25]. The non-linearity is with respect to both the random effects and the regression parameters. It is due to the product of the DSPs from day $t_1 + 1$ to day t_2, and to parameter β_3. Compared with the model proposed in [23, chapter 3], β_3 is an additional parameter which provides more flexibility and allows a better description of the relationship between the DSP and the time-varying covariate. (see Section 4).

Notation. For n_1, n_2 in $\mathbb{N}$ and α in $\mathbb{R}$, let $p_i^{t_1 \to t_2}(n_2|n_1, \alpha; \beta)$ denote the conditional probability, under model (5), that $N_{i,t_2} = n_2$ given $N_{i,t_1} = n_1$ and $\alpha_i = \alpha$.

3.4. Bayesian estimation of the parameters via an MCMC algorithm

In this subsection, we motivate the choice of a Bayesian estimation procedure based on an MCMC algorithm [7,16] to infer on the parameters of the survival model. The R code of the estimation procedure can be asked to the corresponding author.

Under assumptions (A_1) and (A_2) set in Section 3.2, conditional variables $N_{i,t_{i,k}}|N_{i,t_{i,k-1}}$, $k = 1, \ldots, K_i$, are independent random variables, and the counting processes of surviving fish are independent from mesocosm to mesocosm. Consequently, the likelihood for σ and β is

$$l(\sigma, \beta; \boldsymbol{N}) = \prod_{i=1}^{I} \int_{\mathbb{R}} \phi(\alpha; \sigma) \prod_{k=1}^{K_i - 1} p_i^{t_{i,k} \to t_{i,k+1}}(N_{i,t_{i,k+1}}|N_{i,t_{i,k}}, \alpha; \beta) \mathrm{d}\alpha, \tag{6}$$

where $\boldsymbol{N} = \{N_{i,t} : i = 1, \ldots, I, t = t_{i,1}, \ldots, t_{i,K_i}\}$ and $\phi(\cdot;\sigma)$ denotes the density function of a normal random variable with mean 0 and standard deviation σ. In Eq. (6), $\prod_{k=1}^{K_i-1} p_i^{t_{i,k} \to t_{i,k+1}}(N_{i,t_{i,k+1}}|N_{i,t_{i,k}}, \alpha; \beta)$ is the probability of the survival sequence $N_{i,t_{i,1}}, \ldots, N_{i,t_{i,K_i}}$ observed in mesocosm i, given α and β. Since α is considered as a random effect, this probability is integrated with respect to α. Thus, the integrated probability depends on β and the standard deviation σ of the random effects. We then multiply the integrated probabilities for all the mesocosms to obtain Eq. (6).

In statistics, the likelihood is often used to estimate model parameters. But two problems arise for evaluating the likelihood given by Eq.(6). First, the probability $p_i^{t_{i,k}\to t_{i,k+1}}(N_{i,t_{i,k+1}}|N_{i,t_{i,k}},\alpha;\beta)$ depends on the daily expected values of the fish length $\mathbb{E}(X_{i,t})$, $t = t_{i,k},\ldots,t_{i,k+1}$, that are unknown for the experiments. We only observe the sample means $\bar{X}_{i,t_{i,k}}$ and $\bar{X}_{i,t_{i,k+1}}$ of the fish lengths at the capture days $t_{i,k}$ and $t_{i,k+1}$. So, $p_i^{t_{i,k}\to t_{i,k+1}}(N_{i,t_{i,k+1}}|N_{i,t_{i,k}},\alpha;\beta)$ is computed by approximating $\mathbb{E}(X_{i,t})$ ($t \in \{t_{i,k},\ldots,t_{i,k+1}\}$) by

$$\mathbb{E}(X_{i,t}) \approx X^{\infty} - (X^{\infty} - \bar{X}_{i,t_{i,k}})\left(\frac{X^{\infty} - \bar{X}_{i,t_{i,k+1}}}{X^{\infty} - \bar{X}_{i,t_{i,k}}}\right)^{t/(t_{i,k+1}-t_{i,k})}. \tag{7}$$

This approximation was obtained by assuming that (A_4) the expected fish length satisfies a Bertalanffy's relationship

$$\mathbb{E}(X_{i,t}) = \mathbb{E}(X_{i,t-1}) + \tau\{X^{\infty} - \mathbb{E}(X_{i,t-1})\}, \qquad t = 1, 2, \ldots$$

where X^{∞} is the maximum fish length and τ is the growth rate. So, expression (7) corresponds to a concave growth curve which interpolates observed mean lengths.

Second, although probabilities $p_i^{t_{i,k}\to t_{i,k+1}}(N_{i,t_{i,k+1}}|N_{i,t_{i,k}},\alpha;\beta)$ can be computed using the proposal made above, likelihood (6) cannot be evaluated analytically because of the integrals over α. This difficulty is circumvented in [23, chapter 3] by maximizing an approximation of the likelihood where the integrals are assessed using Gaussian quadrature. Instead, we specify prior distributions for σ and β and apply an MCMC algorithm yielding posterior joint distributions for σ, β but also the group effects α_i ($i = 1,\ldots,I$).

The MCMC algorithm includes a Metropolis–Hasting within Gibbs algorithm [16]. At each iteration, the unknowns (i.e. $\log\sigma$, β_1, β_2, β_3 and the α_is) are updated using Gaussian transition kernels. The kernels are centered on the values of the unknowns at the previous iteration and their standard deviations were 0.1 for $\log\sigma$, β_1, β_3 and the α_is, and 5 for β_2. The regression parameters β_1, β_2, β_3 are updated simultaneously as well as the group effects α_is.

Compared with the estimation procedure used in [23], the Bayesian approach we propose has two advantages. First, one obtains estimates for the random effects which can be used in a post-analysis (using additional covariates) to better understand the mesocosms characteristics. Second, the assessment of the uncertainty and the correlation structure of the parameter estimators are not based on asymptotic results.

Remark. The model can also be fitted via a Monte–Carlo expectation-maximization (MCEM) algorithm [27]. The uncertainty and the correlations of the parameter estimators can be assessed by applying a parameteric bootstrap procedure [11]. Such an approach has the same advantages than the Bayesian approach presented above, but it is much more time consuming.

4. Application to mosquitofish survival data

4.1. Estimation

We applied the Bayesian estimation procedure using constant improper priors for $\log\sigma$, β_1, β_2 and β_3. Fig. 2 shows the two-dimensional posterior distributions of the parameters. In each panel,

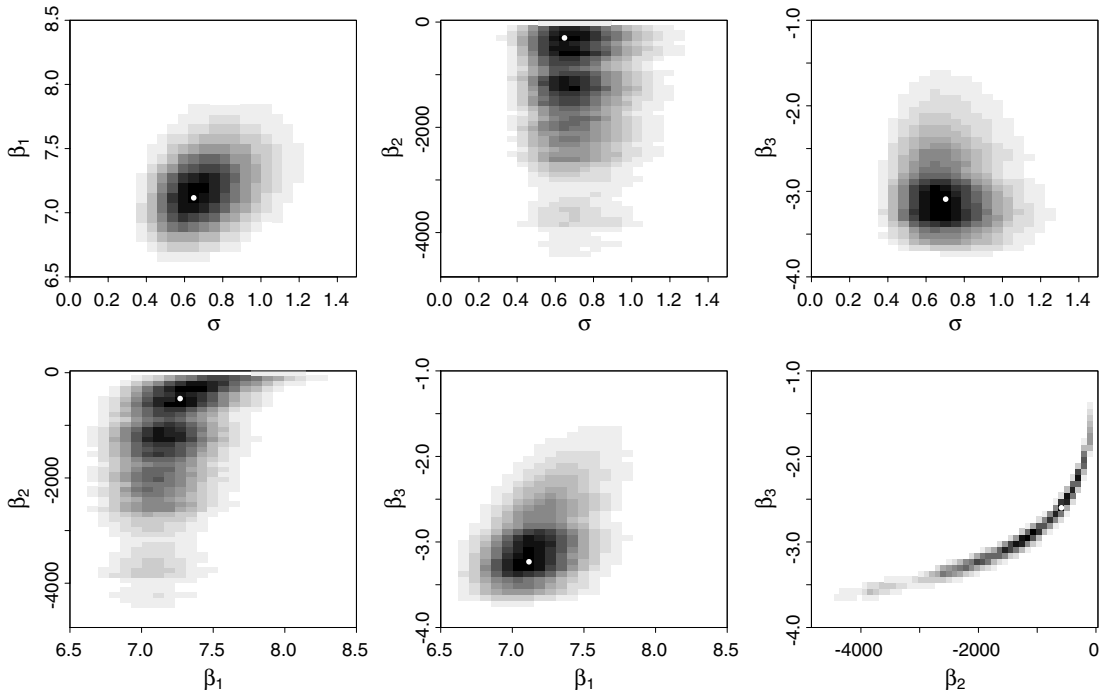

Fig. 2. Bivariate histograms representing the two-dimensional posterior distributions of the standard deviation σ and the regression parameters β_1, β_2 and β_3 (high densities are in black). White dots correspond to the posterior modes.

the posterior mode is marked by a white dot. The four-dimensional posterior mode corresponds to $\hat{\sigma} = 0.630$, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (7.30, -295, -2.25)$. Note on the bottom right panel that the MCMC runs for β_2 and β_3 are strongly dependent. This feature will be discussed in the perspective of sensibility analysis (see Section 4.3).

Using the Bayesian procedure, we fitted the submodel with $\sigma = 0$ fixed (no mesocosm effect) and the submodel with $\beta_3 = 1$ fixed (f_β is linear). We compared the fits using two criteria, the deviance information criterion [DIC, 22] and the posterior expectation of the Akaike's information criterion [EAIC, 3]; lower these criteria are, better the fit is. We computed the DIC and AIC differences, say ΔDIC and ΔEAIC, between the complete model and each of the two submodels. For the complete model against the submodel with $\sigma = 0$: $\Delta\text{DIC} = -39$ and $\Delta\text{EAIC} = -49$. For the complete model against the submodel with $\beta_3 = 1$: $\Delta\text{DIC} = -15$ and $\Delta\text{EAIC} = -11$. Obviously the complete model has lower criteria, and according to the rule given in [22], the differences are large enough to state that the submodels have considerably less support than the complete model. Consequently, both the mesocosm effects and the non-linearity of f_β allows a more accurate description of the relationship between the fish DSP and the fish length.

4.2. Residual analysis

We performed a residual analysis to check whether the parametric link we specified in Eqs. (3) and (4) is suitable or not. We used standardized residuals for the ratios $N_{i,t_k}/N_{i,t_{k-1}}$ ($i = 1, \ldots, I$ and

$k = 2, \ldots, K_i$) and we plotted them against the estimated fish lengths (given by (7)) at the mean time between t_{k-1} and t_k.

Given α_i and β, we define

$$r_{i,k} = \frac{\frac{N_{i,t_{i,k}}}{N_{i,t_{i,k-1}}} - \mathbb{E}\left(\frac{N_{i,t_{i,k}}}{N_{i,t_{i,k-1}}} \middle| N_{i,t_{i,k-1}}, \alpha_i; \beta\right)}{\mathbb{V}\left(\frac{N_{i,t_{i,k}}}{N_{i,t_{i,k-1}}} \middle| N_{i,t_{i,k-1}}, \alpha_i; \beta\right)^{1/2}} = \frac{N_{i,t_{i,k}} - N_{i,t_{i,k-1}} v_{i,k}(\alpha_i, \beta)}{[N_{i,t_{i,k-1}} v_{i,k}(\alpha_i, \beta)\{1 - v_{i,k}(\alpha_i, \beta)\}]^{1/2}}, \tag{8}$$

where (see Eq. (5))

$$v_{i,k}(\alpha, \beta) = \prod_{t=t_{i,k-1}}^{t_{i,k}-1} \pi\{\alpha, \beta, \mathbb{E}(X_{i,t})\}.$$

$r_{i,k}$ was computed for each vector (α_i,β) obtained in the MCMC run, and the standardized residual $\hat{r}_{i,k}$ we finally used is the posterior median of $r_{i,k}$.

Fig. 3 (left) shows residuals $\hat{r}_{i,k}$ obtained under the complete model. Fig. 3 (right) shows residuals $\hat{r}_{i,k}$ obtained under the submodel with $\beta_3 = 1$, i.e. f_β is a linear function of the fish length. The horizontal dashed lines are at quantiles 0.05 and 0.95 of the normal distribution (if fish counts are large, $\hat{r}_{i,k}$ approximately have centered normal distributions with variance equal to one). The residual plot in the right panel suggests that the explanatory variable, i.e. the fish length, should be transformed. Contrastively, the residual plot in the left panel does not suggest such a transformation. This qualitative observation partly explains the differences in the DIC and EAIC values provided in the previous subsection. To conclude, the parametric link between the fish DSP and the fish length we specified in Eqs. (3) and (4) seems to be suitable.

Remark. There are three clusters of points at the right hand side of each residual plot. These clusters correspond to observations with 0 (top cluster), 1 (middle cluster) or 2 (bottom cluster) dead fish in small groups of fish (see Table 1).

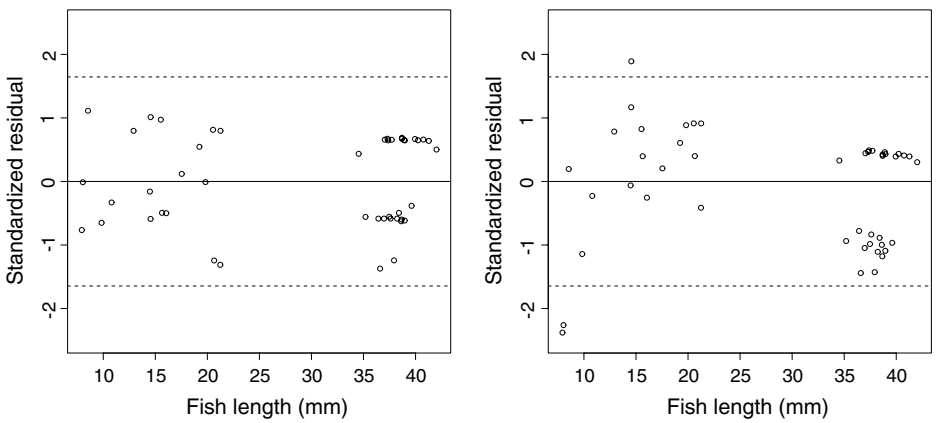

Fig. 3. Residual plots obtained under the complete model (left) and under the submodel with $\beta_3 = 1$ (right).

4.3. Functional link between the DSP and the fish length

Let g_β denote the conditional DSP for a fish with expected length x given the mesocosm effect is zero

$$g_\beta(x) = \pi(0, \beta, x) = \text{logit}^{-1}\{f_\beta(x)\}.$$

Fig. 4 shows $g_{\hat{\beta}}$ (solid line), where $\hat{\beta}$ is the posterior mode given in Section 4.1, together with the posterior median (dotted line) and the 95%-credibility pointwise envelope of $g_\beta(x)$ for fish lengths x in the range of observed fish lengths. This plot shows that the DSP quickly increases with the fish length and once fish are larger than about 15 mm, their DSP is constant up to the mesocosm effect.

The estimate $g_{\hat{\beta}}$ will be incorporated in the individual-based model mentioned in the introduction. The present study also provides insight into how to analyze the sensibility of the model output with respect to β (one of the aims of the individual-based model is to see how variations of mosquitofish life history traits (described by parameters in the model) influence population dynamics (i.e. model output)). We saw in Fig. 2 that the MCMC runs for β_2 and β_3 are strongly dependent. The sensiblity analysis should account for this dependence. The linear regression model $\beta_3 = 0.75 - 0.52\log(-\beta_2)$ fitted to the MCMC runs reflects quite well the dependence ($R^2 = 0.98$). Thus, the sensibility of the model with respect to the components of β can be performed with β_1 approximately varying in $[6.9, 7.3]$, β_2 in $[-1000, -200]$ and β_3 satisfying the equation given above.

4.4. Illustration: paths of the count of surviving fish

To better understand the behavior of the fitted model, we simulated paths of the count of surviving fish. For all the simulated paths, the count of fish at day 0 is 50, and the expected fish length satisfies the Bertalanffy's relationship

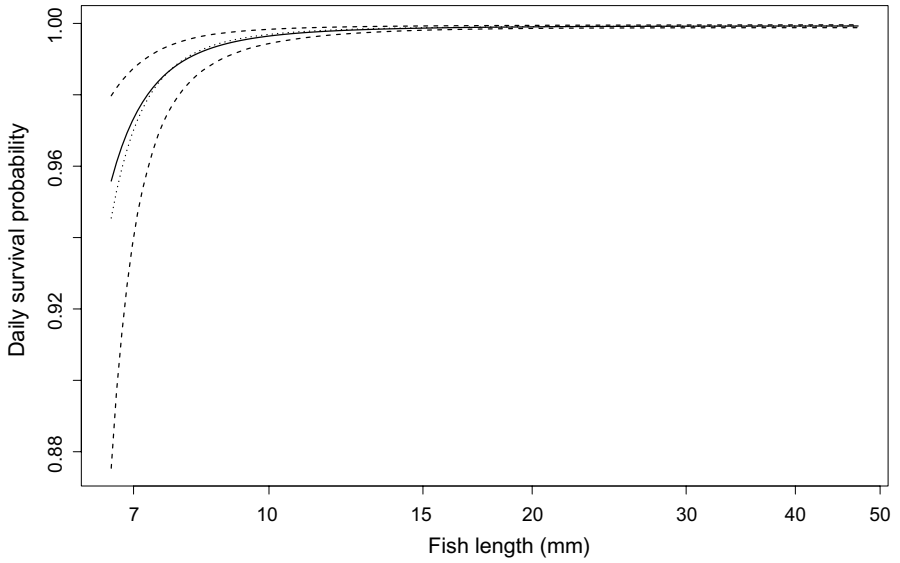

Fig. 4. Estimate $g_{\hat{\beta}}$ (solid line), posterior median (dotted line) and 95%-credibility pointwise envelope (dashed lines) of g_β for fish lengths varying between 6.6 and 47.2 mm.

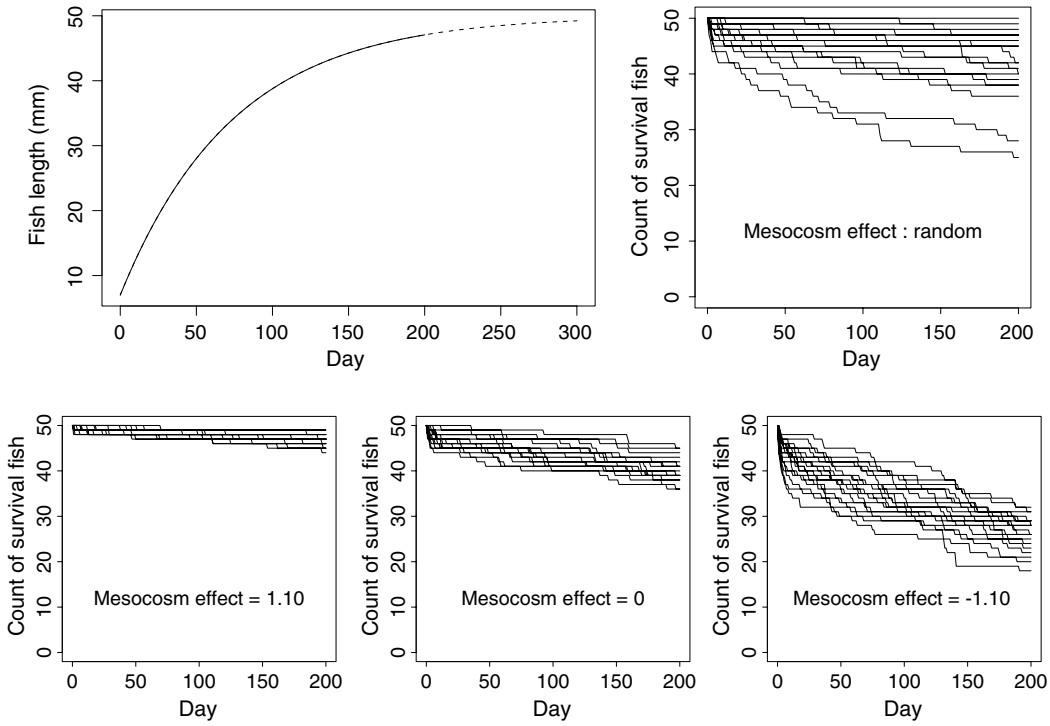

Fig. 5. Simulation, under the fitted model, of paths of the count of surviving fish (top right and bottom) given the daily evolution of the fish length (top left). Top right: 20 paths corresponding to 20 different mesocosm effects drawn from the law $N(0,\hat{\sigma})$; bottom: 20 paths corresponding to a mesocosm effect fixed at quantiles $q_{N(0,\hat{\sigma})}(0.975) = 1.10$ (left), $q_{N(0,\hat{\sigma})}(0.5) = 0$ (middle) or $q_{N(0,\hat{\sigma})}(0.025) = -1.10$ (right).

$$\mathbb{E}(X_{i,t}) = \mathbb{E}(X_{i,t-1}) + \tau\{X^{\infty} - \mathbb{E}(X_{i,t-1})\}, \qquad t = 1, 2, \ldots$$

where the initial length is $\mathbb{E}(X_{i,0}) = 7$ mm, the growth rate is $\tau = 0.013$ mm per day and the maximum length is $X^{\infty} = 50$ mm. Fig. 5 shows the daily evolution of the fish length (top left). It also shows simulated paths of the count of surviving fish: in the top right panel 20 mesocosm effects were generated from the law $N(0, \hat{\sigma})$ and for each mesocosm effect a path was simulated; in the bottom panels, the mesocosm effect was fixed at different quantiles of the law $N(0, \hat{\sigma})$ and 20 paths were simulated for each value of the mesocosm effect.

Under the fitted model, the paths of the count of surviving fish can be very variable. For instance, in the top right panel, there is a path without death during 200 days and another one with about 25 deaths during the same period. The bottom panels show the strong influence of the mesocosm effect. Including mesocosm effects in the individual-based model of Ref. [12] will allow the authors to reflect a significant source of variability of the population dynamics.

5. Discussion

We have presented a generalized nonlinear mixed model and proposed a Bayesian estimation procedure using an MCMC algorithm in order to learn about the relationship between

the survival probability of individuals and a time-varying covariate when data are grouped and time-period censored. The model was built by aggregating in time and at the group level an underlying model adapted to the individual and day scale. Moreover, the model integrates a flexible parametric form for the daily survival probability (DSP). Thus, the link between the DSP and the time-varying covariate can be more accurately explored. The model we used can be viewed as an extension of the nest survival model first developed in [8] and improved in [23].

We have applied this approach in order to explore the relationship between the mosquitofish DSP and the fish length. This study will be useful for improving the simulation model proposed in [12] which includes, at present, a quite rough description of this function. The characterization of the mesocosm effect will also be useful for improving the "realism" of the simulation model. Besides, we saw that the results obtained can be used in the perspective of a sensibility analysis.

It would be interesting to assess what covariates could be linked with the mesocosm effects. Plotting the estimated mesocosm effects against the experimentation years or the experimentation months did not result in a significant relationship (not shown). The possible link between the mesocosm effect and the predator and vegetation densities [see Section 2 and references 1,26] remain to be explored.

In the model construction we assumed that survival processes of individual fish are independent (assumption A_2 in Section 3.2). In the context of mosquitofish this assumption is not strong, but in other contexts it can be. For instance, instead of the occurrence of the death of a mosquitofish, consider the occurrence of a transmissible disease affecting humans (e.g. AIDS, SARS), animals (e.g. avian flu) or plants (e.g. wheat rusts), and suppose one aims to link the DSP (the time step of interest can be different from the day) with the time-varying intensity of prevention measures for example. In such a context, assumption A_2 is clearly not suitable. Indeed, the "survival" process of an individual can affect the "survival" processes of other individuals since the disease is transmissible. The model might be adapted by conditioning the transition probabilities (1) not only on the past state of the individual of interest but also on the past states for the other individuals. The consequences on the construction of the model and on the estimation procedure of such a modification should be explored.

Acknowledgements

The authors would like to thank G. Bounaud and C. Sevellec for technical assistance for the experiments. This work was supported in part by the Programme National d'Ecotoxicologie (PNETOX., France).

References

[1] L.W. Botsford, B. Vondracek, T.C. Wainwright, A.L. Linden, R.G. Kope, D.E. Reed, J.J. Cech Jr., Population development of the mosquitofish, *Gambusia affinis*, in rice fields, Environ. Biol. Fish. 20 (1987) 143.

[2] J.M. Box-Steffensmeier, B.S. Jones, Event History Modeling, A Guide for Social Scientists, Cambridge University (2004).

[3] S.P. Brooks, Discussion on the paper by spiegelhalter, best, carlin and van der linde, J. R. Stat. Soc. B 64 (2002) 616.
[4] C.R. Brown, M. Bomberger Brown, Group size and ectoparasitism affect daily survival probability in a colonial bird, Behav. Ecol. Sociobiol. 56 (2004) 498.
[5] J.A. Cabral, J.C. Marques, S.N. Nielsen, Modeling mosquitofish (*Gambusia holbrooki*) response to genapol 0xd-080, a non ionic surfactant, in rice fields, Ecol. Eng. 16 (2001) 537.
[6] L.V.A. Chalita, E.A. Colosimo, J.R. de Souza Passos, Modeling grouped survival data with time-dependent covariates, Commun. Stat.: Simul. Comput. 35 (2006) 975.
[7] M. Davidian, D.M. Giltinan, Nonlinear Models for Repeated Measurement Data, Chapman & Hall, London, 1995.
[8] S.J. Dinsmore, G.C. White, F.L. Knopf, Adv anced techniques for modeling avian nest survival, Ecology 83 (2002) 3476.
[9] C.J. Doyle, R.P. Lim, Sexual behavior and impregnation success of adult male mosquitofish following exposure to 17 beta-estradiol, Ecotoxicol. Environ. Safety 61 (2005) 392.
[10] V. Drèze, G. Monod, J.P. Cravedi, S. Biagianti-Risbourg, F. Le Gac, Effects of 4-nonylphenol on sex differentiation and puberty in mosquitofish (*Gambusia holbrooki*), Ecotoxicology 9 (2000) 93.
[11] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
[12] V. Ginot, S. Gaba, R. Beaudouin, F. Aries, H. Monod, Combined use of local and anova-based global sensitivity analyses for the investigation of a stochastic dynamic model: application to the case study of an individual-based model of a fish population, Ecol. Model. 193 (2006) 479.
[13] C. Kooperberg, D.B. Clarkson, Hazard regression with interval-censored data, Biometrics 53 (1997) 1485.
[14] C.E. McCulloch, S.R. Searle, Generalized, Linear and Mixed Models, Wiley, New York, 2001.
[15] M. Mulvey, M.C. Newman, A. Chazal, M.M. Keklak, Genetic and demographic responses of mosquitofish (*Gambusia holbrooki* girard 1859) populations stressed by mercury, Environ. Toxicol. Chem. 14 (1995) 1411.
[16] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer, New York, 1999.
[17] J.J. Rotella, S.J. Dinsmore, T.L. Shaffer, Modeling nest-survival data: a comparison of recently developed methods that can be implemented in mark and sas, Animal Biodiv. Conserv. 27 (2004) 187.
[18] D. Ruppert, M.P. Wand, R.J. Carroll, Semiparametric regression, Cambridge University, 2003.
[19] C.C. Schwartz, M.A. Haroldson, G.C. White, Survival of cub and yearling grizzly bears in the Greater Yellowstone Ecosystem, 1983–2001, Wildlife Monogr. 161 (2006) 25.
[20] T.L. Shaffer, A unified approach to analyzing nest success, The Auk 121 (2004) 526.
[21] Y.H. Sparling, N. Younes, J.M. Lachin, O.M. Bautista, Parameteric survival models for interval-censored data with time-dependent covariates, Biostatistics 7 (2006) 599.
[22] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. van der Linde, Bayesian measures of model complexity and fit (with discussion), J. R. Stat. Soc. B 64 (2002) 583.
[23] S.E. Stephens, The influence of landscape characteristics on duck nesting success in the Missouri Coteau Region of North Dakota. Ph.D. thesis, Montana State University, 2003.
[24] G. Toft, T.M. Edwards, E. Baatrup, L.J. Guillette, Disturbed sexual characteristics in male mosquitofish (*Gambusia holbrooki*) from a lake contaminated with endocrine disruptors, Environ. Health Persp. 111 (2003) 695.
[25] E.F. Vonesh, V.M. Chinchilli, Linear and Nonlinear Models for the Analysis of Repeated Measurements, Dekker, New York, 1997.
[26] T.C. Wainwright, R.G. Kope, L.W. Botsford, J.J. Cech Jr., Implications of laboratory mosquitofish experiments for population development in rice fields, Proceedings of the California Mosquito and Vector Control Association 52 (1984) 110.
[27] G.C.G. Wei, M.A. Tanner, A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms, J. Am. Stat. Assoc. 85 (1990) 699.
[28] P.S.F. Yip, K.F. Lam, E.H.Y. Lau, P.-H. Chau, K.W. Tsang, A comparison study of realtime fatality rates: severe acute respiratory syndrome in hong kong, singapore, taiwan, toronto and beijing, china, J. R. Stat. Soc. A 168 (2005) 233.
[29] P.S.F. Yip, E.H.Y. Lau, K.F. Lam, R.M. Huggins, A chain multinomial model for estimating the real-time fatality rate of a disease, with an application to severe acute respiratory syndrome, American Journal of Epidemiology 161 (2005) 700.