A maximum likelihood estimator for left-truncated lifetimes based on probabilistic prior information about time of...

Article *in* Journal of Applied Statistics · December 2017

DOI: 10.1080/02664763.2017.1410527

Rubén Manso
Forest Research - Forestry Commission UK
36 PUBLICATIONS **111** CITATIONS
SEE PROFILE

Rafael Calama
Instituto Nacional de Investigación y Tecnolo...
119 PUBLICATIONS **1,185** CITATIONS
SEE PROFILE

Marta Pardos
Instituto Nacional de Investigación y Tecnolo...
94 PUBLICATIONS **976** CITATIONS
SEE PROFILE

Mathieu Fortin
AgroParisTech
82 PUBLICATIONS **653** CITATIONS
SEE PROFILE

ORIGINAL RESEARCH ARTICLE

A maximum likelihood estimator for left-truncated lifetimes based on probabilistic prior information about time of occurrence

Rubén Manso[a,b], Rafael Calama[c], Marta Pardos[c] and Mathieu Fortin[d]

[a]INRA, UMR 1092 LERFoB, 1 rue de l'Arboretum, 54280, Champenoux, France; [b]Forest Research, Northern Research Station, Roslin, Midlothian, UK, EH25 9SY; [c]INIA-CIFOR. iuFOR. Ctra A Coruña km 7.5, 28040 Madrid, Spain; [d]AgroParisTech, UMR 1092 LERFoB, 14 rue Girardet, 54042 Nancy, France

ARTICLE HISTORY
Compiled December 7, 2017

ABSTRACT
In forestry, many processes of interest are binary and they can be modeled using lifetime analysis. However, available data are often incomplete, being interval- and right-censored as well as left-truncated, which may lead to biased parameter estimates. While censoring can be easily considered in lifetime analysis, left truncation is more complicated when individual age at selection is unknown. In this study, we designed and tested a maximum likelihood estimator that deals with left truncation by taking advantage of prior knowledge about the time when the individuals enter the experiment. Whenever a model is available for predicting the time of selection, the distribution of the delayed entries can be obtained using Bayes' theorem. It is then possible to marginalize the likelihood function over the distribution of the delayed entries in the experiment to assess the joint distribution of time of selection and time to event. This estimator was tested with continuous and discrete Gompertz-distributed lifetimes. It was then compared with two other estimators: a standard one in which left truncation was not considered and a second estimator that implemented an analytical correction. Our new estimator yielded unbiased parameter estimates with empirical coverage of confidence intervals close to their nominal value. The standard estimator leaded to an overestimation of the long-term probability of survival.

KEYWORDS
Lifetime analysis; Left truncation; Likelihood estimation; Marginal likelihood; Bayes' theorem; Gompertz hazard

The published version of this contribution in the *Journal of Applied Statistics* can be found at https://doi.org/10.1080/02664763.2017.1410527.

1. Introduction

In forestry, a number of time-dependent binary processes, such as seed germination and tree survival, are of great interest to modelers. Data from these processes are recorded as the age of an individual when it experiences a change of status (e.g. seed to seedling, alive to dead). This age is usually referred to as *lifetime*. These outcomes can efficiently be modeled using regression based on lifetime analysis [15]. Although

CONTACT Rubén Manso. Email: rmgforestal@hotmail.com, ruben.manso@forestry.gsi.gov.uk

this approach is not widespread in forest science, some contributions can be found in the literature where lifetime analysis was successfully applied [e.g. 16, 26, 28].

Lifetime data are often said to be incomplete, which happens when the exact lifetime of individuals is unknown. This commonly results in *right* and *interval censoring* in the dataset. In forestry, this censoring is usually due to the sampling scheme, such as plots revisited at spaced intervals or individuals surviving beyond experimental follow-up. A third type of data incompleteness is called *left truncation.* Data are said to be left-truncated if some individuals were already alive when they were selected for the experiment.

There are some examples of left-truncated data in the forestry literature [e.g. 6, 10]. However, it is a more common practice to have a set of controlled individuals monitored from individual occurrence [e.g. 11, 24, 25, among others]. This means that pre-existing individuals, i.e. with left-truncated lifetimes, are generally excluded from experiments. In other words, these prospective studies would be limited to the time span where the whole history of individuals can be traced back. Given the short-term nature of many studies due to budgetary reasons, it may prove useful to consider these left-truncated lifetimes.

Data incompleteness imposes restrictions on the basic lifetime methods for regression models, and non-trivial approaches are necessary to obtain unbiased parameter estimates [15]. While the likelihood estimator can be easily adapted to right and interval censoring, left truncation may be more challenging if actual individual lifetime is unknown. In order to illustrate this, let us first define the non-negative random variable T as the lifetime, i.e. time elapsed from the birth of an individual to its death. Let us also define the random variable U as *time of selection*, i.e. time elapsed from the birth of an individual to its inclusion in the experiment. A complementary variable would be V, or *time to event*, i.e. time from selection to death. The latter can be expressed in terms of T and U as $V = T - U$. Unfortunately, time of selection U may be quite difficult to ascertain in many forestry applications as the "birth" event is not usually recorded (e.g. time since a particular seed was released, time since a seedling was established . . .). As a result, T would also be unknown and a trivial modification of the likelihood estimator to take left-truncation into account will be no longer possible.

Theoretical works dealing with left truncation can be found in statistical literature where partial likelihood applies [e.g. 8] or, of course, where U can be precisely asserted [e.g. 27]. When neither is the case, lifetime needs to be inferred from the joint probability of U and V. An equivalent approach used by many is to refer all events to some arbitrary calendar time prior to birth and define R as an additional random variable representing *time of occurrence*, i.e. time elapse from that arbitrary origin to individual birth and $T' = T + R$. In this case T can be imputed from the joint distribution of T' and R. A common procedure is to assume a uniform distribution of R or U, which notably simplifies the mathematical expression of the joint likelihood [e.g. 7]. This approach implies that individuals are equally likely to occur at any time within a given interval. Testing whether this assumption, as known as "stationary hypothesis", holds has been subject of research interest in medical studies [1, 3]. This alternative may be overly simplistic in forestry studies where the process preceding selection is usually more complex (e.g. climate-dependent individual occurrence). Thus, a more general approach is clearly needed.

The objective of this paper is to present and test such an approach. The method is based on the marginalization of the left-truncated likelihood over all possible values of U according to any distribution of the latter. This distribution can be either continuous or discrete. Extending the method to discrete distributions sets the grounds for the use of external models in order to impute the distribution of U. This can be particularly useful in forestry, where a number of pre-existing models for different processes relevant for U, such as seed dispersal or seed germination, are usually available.

This paper is structured as follows. Mathematical developments are provided in Section 2. First, a general background on lifetime analysis, censoring schemes and likelihood definition is given. Second, a hazard functional form is chosen to assess the bias in parameter estimates induced by left truncation. Third, the proposed estimator, which takes left truncation into account, is presented in detail. The new approach, the standard one are evaluated through a simulation study in Section 3, where continuous and discrete distributions of U are considered. A third alternative method that implemented an analytical correction is tested as well. Simulations are carried out across a range of sample sizes and parameterizations of the lifetime distribution. Both main methods are then applied to the case study of stone pine (*Pinus pinea* L.) seedling survival in central Spain, where an external germination model was used to impute the distribution of U. Sections 4 and 5 are devoted to the presentation and discussion, respectively, of the results obtained from the calculations in Section 3.

2. Mathematical developments

2.1. *Lifetime analysis: some definitions*

We followed the definitions provided in [15, p.8]. The cumulative density of T, $F(t)$, yields the probability that the lifetime occurred before time t, such that $F(t) = P(T \leq t)$. The survivor function $S(t)$ represents the probability that a given individual survives until time t, i.e., $S(t) = 1 - F(t)$. The survivor function can also be expressed as:

$$S(t) = \exp\left(-\int_0^t h(x)\mathrm{d}x\right) \tag{1}$$

where $h(t)$ is the instantaneous mortality rate for a given individual at time t, which is conditional on its survival up to t. Function $h(t)$ is usually referred to as the hazard function. Therefore, the probability of surviving up to t is a monotonically decreasing function of the sum of all hazards previous to t. The functional form of the hazard depends upon the assumed parametric distribution of lifetimes. The parameters of the assumed distribution, designated by $\boldsymbol{\beta}$, can be estimated through the maximization of a likelihood function.

2.2. *Censoring schemes and likelihood definition*

The formulation of the likelihood function is conditioned by the sampling scheme. In the unlikely case where lifetimes are precisely observed, the likelihood function is expressed as:

$$\mathcal{L} = \prod_{i}^{n} f(t_i) \tag{2}$$

where $f(t)$ is the probability density function of T and n is the sample size.

However, the sampling scheme imposed by experimental designs often leads to incomplete data, i.e. lifetimes are not precisely known. [15, p.49] reports three types of data incompleteness: interval censoring, right censoring and left-truncation. The likelihood function can be adapted to take the type of data incompleteness into account. The existing methods are briefly illustrated below.

In the case of interval censoring, the only available information is that lifetime T occurred within a given measurement interval, so that $t_{start} < T \leq t_{end}$, where t_{start} and t_{end} represent the times at which the individual was last observed alive and at which its death was confirmed, respectively. The likelihood function can then be defined as:

$$\mathcal{L} = \prod_{i}^{n} S(t_{start,i}) - S(t_{end,i}) \tag{3}$$

If some observations are right-censored, then our knowledge about their lifetime T is limited to the fact that $t_{end} < T$. Given that unobserved events are complementary to observed events, the likelihood shown in Eq. 3 can be expanded to take these right-censored observations into account as follows:

$$\mathcal{L} = \prod_{i}^{n} (S(t_{start,i}) - S(t_{end,i}))^{q_i} \cdot S(t_{end,i})^{1-q_i} \tag{4}$$

where q_i is a dummy variable that adopts the value of 1 when the death of an individual has been recorded and 0 otherwise.

If on top of right and interval censored times left truncation is present, Eq. 4 need to be adapted. Taking u_i as the time of selection for individual i and $T \geq u_i$, we have

$$\mathcal{L} = \prod_{i}^{n} \Big(\frac{S(t_{start,i}) - S(t_{end,i})}{S(u_i)}\Big)^{q_i} \cdot \Big(\frac{S(t_{end,i})}{S(u_i)}\Big)^{1-q_i} \tag{5}$$

Likelihood 5 holds if the assumption of independent delayed entry is met. This assumption states that the selection of an individual should not depend on individual lifetime. We adopt this assumption throughout. The reader is referred to [15, p.68] for further details on this issue.

2.3. *Bias in parameter estimates due to misinformation of time of selection*

The direct application of likelihood function 5 requires $t_{start,i}$, $t_{end,i}$ and u_i to be known for each individual i. However, the time of occurrence for each individual is some times unknown in left-truncated experiments. As a consequence, the values of those three times remain unknown as well. It can be demonstrated that parameter estimation via likelihood maximization may be biased when those times are not explicitly provided. Although this is true for all lifetime distributions where calendar time is explicitly part of the hazard function, we will focus on Gompertz-distributed lifetimes in the

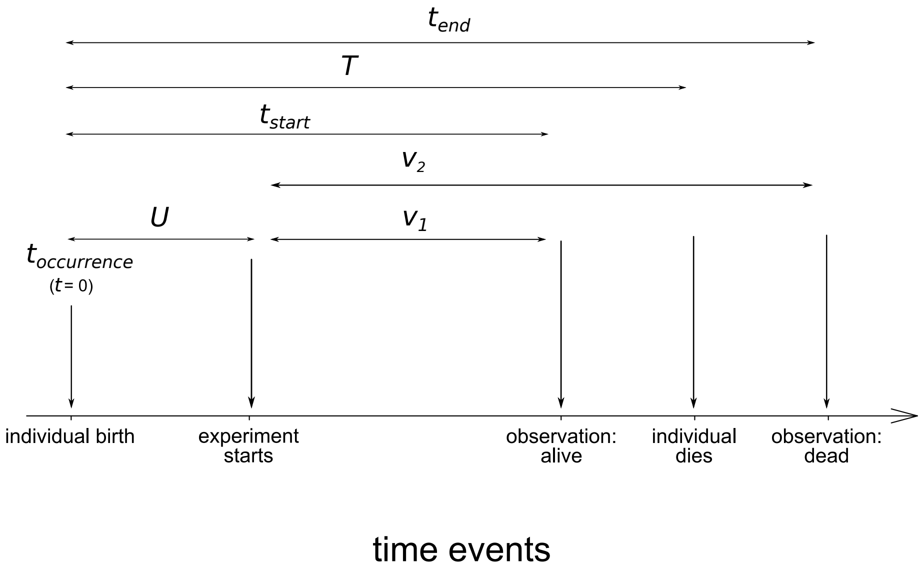

Figure 1. Time events and corresponding time terms used in the present study.

next developments. The Gompertz distribution is convenient for its simplicity while remaining widely applicable.

It is possible to redefine the aforementioned endpoints of the censored observations in terms of the time of selection (U) and the time to event (V). It follows that $t_{start,i} = v_{1,i} + u_i$ and $t_{end,i} = v_{2,i} + u_i$. A scheme with all time terms is detailed in Fig. 1. Developing Equation 5, the likelihood of observation i conditional on u_i is then

$$\mathcal{L}_i \mid u_i = (S(v_{1,i} + u_i \mid T > u_i) - S(v_{2,i} + u_i \mid T > u_i))^{q_i} \cdot (S(v_{2,i} + u_i \mid T > u_i))^{1-q_i} \tag{6}$$

Each conditional survival probability in Eq. 6 can be expressed in terms of hazard through the survival function (see Eq. 1):

$$S(v_{1,i} + u_i \mid T > u_i) = \frac{S(v_{1,i} + u_i)}{S(u_i)} = \exp\left(-\int_{u_i}^{v_{1,i}+u_i} h(t)\mathrm{d}t\right) \tag{7a}$$

$$S(v_{2,i} + u_i \mid T > u_i) = \frac{S(v_{2,i} + u_i)}{S(u_i)} = \exp\left(-\int_{u_i}^{v_{2,i}+u_i} h(t)\mathrm{d}t\right) \tag{7b}$$

The hazard function derived from the Gompertz-distributed lifetime is defined as:

$$h(t) = e^{\beta_0 + \beta_1 t} \tag{8}$$

where $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \beta_1\end{pmatrix}^T$ are estimable parameters.
Substituting Eq. 8 for $h(t)$ in Eq. 7 and integrating leads to

$$S(v_{\cdot,i} + u_i \mid T > u_i) = \exp\left(\frac{-e^{\beta_0 + \beta_1 u_i}}{\beta_1}(e^{\beta_1 t_{\cdot,i}} - 1)\right) \tag{9}$$

Provided that observation i is left-truncated, i.e. $u_i > 0$, the quantity $\beta_1 u_i$ is systematically added to β_0 in the survival function. Therefore, ignoring left-truncation (i.e. assuming that U has a degenerate distribution at 0 so that $u_i = 0$) induces a bias in $\hat{\beta}_0$, whose extent can be approximated as shown in Appendix A. Although this approximation could theoretically be used to correct for the bias in parameter estimates, the estimated variance of these estimates would likely remain biased. Henceforth, we will refer to this corrected estimator as the Analytically Corrected Likelihood Estimator (ACLE).

2.4. *Unbiased likelihood estimation through prior probabilistic information about time of occurrence*

An alternative to obtain unbiased parameter estimates and unbiased estimated variances of them is the maximization of the likelihood in Eq. 6 marginalized over the distribution of u_i. Hereafter, we will refer to this alternative method as Maximum Marginalized Likelihood Estimator (MMLE). Fundamentally, our adapted likelihood function has the same form than that of [7]. The main advantage of the alternative proposed here is that the marginalization of the likelihood allows for any empirical or theoretical version of the distribution of time of occurrence, and not just the uniform case. This flexibility makes the MMLE estimator adaptable to a wider range of common situations in survival studies in forestry.

Provided that the distribution of u_i is subject to the constraint $u_i \geq 0$, the marginalized likelihood becomes:

$$\mathcal{L}_i = \int_0^\infty \Big(\frac{S(v_{1,i} + u_i) - S(v_{2,i} + u_i)}{S(u_i)}\Big)^{q_i} \cdot \Big(\frac{S(v_{2,i} + u_i)}{S(u_i)}\Big)^{1-q_i} \mathrm{pdf}(u_i \mid u_i \geq 0)\mathrm{d}u_i \quad (10)$$

where $\mathrm{pdf}(u_i \mid u_i \geq 0)$ is the truncated density of U. If the distribution of U is discrete, then the likelihood shown in Eq. 10 can be adapted as:

$$\mathcal{L}_i = \sum_{u_i=0}^{\infty} \Big(\frac{S(v_{1,i} + u_i) - S(t_{2,i} + u_i)}{S(u_i)}\Big)^{q_i} \cdot \Big(\frac{S(v_{2,i} + u_i)}{S(u_i)}\Big)^{1-q_i} \mathrm{pmf}(u_i \mid u_i \geq 0) \quad (11)$$

where $\mathrm{pmf}(u_i \mid u_i \geq 0)$ is the truncated probability mass function of U.

As mentioned before, u_i is the time elapsed since the birth of individual i until it enters the experiment. Let us denote this birth event as E. In forestry, there are many examples of models that predict this event E as a function of time ($Pr(E \mid u_i)$, e.g. probability that a fire occurs in a given year). Because we need $\mathrm{pdf}(u_i \mid u_i \geq 0)$ and $\mathrm{pmf}(u_i \mid u_i \geq 0)$ to sum up to one, they have to be conditional to the occurrence of E. As a consequence, these density and mass functions are instead interpreted as $\mathrm{Pr}(u_i \mid E)$ (e.g. probability that a given year be witness of a fire, providing that a fire occurs). $\mathrm{Pr}(u_i \mid E)$ can be obtained through Bayes' rule from $Pr(E \mid u_i)$ as:

$$\mathrm{Pr}(u_i \mid E) = \frac{\mathrm{Pr}(E \mid u_i)\mathrm{Pr}(u_i)}{\mathrm{Pr}(E)} \quad (12)$$

where $\mathrm{Pr}(u_i)$ is the prior probability of observing u_i and $\mathrm{Pr}(E)$ is the marginal probability of observing event E. Because no date is more likely than another (i.e. all dates are equally likely to 'occur' regardless of any E event), the prior $\Pr(u_i)$ is uniform and can be factored out of Eq. 12. $\Pr(E)$ then reduces to either $\int \Pr(E \mid u_i) \mathrm{d}u_i$ for the continuous case or $\sum \Pr(E \mid u_i)$ for the discrete case. Consequently, if any model is available for the prediction of the probability $\Pr(E \mid u_i)$, then it can be used to define marginalized Likelihoods 10 and 11. Note that our approach is fully frequentist, i.e. no Bayesian inference is carried out. We only use Bayes' rule as a means to assess the terms $\mathrm{pdf}(u_i \mid u_i \geq 0)$ and $\mathrm{pmf}(u_i \mid u_i \geq 0)$ in Likelihoods 10 and 11.

3. Evaluation of the proposed method

The efficiency of the MMLE was evaluated through Monte Carlo simulation and was compared to that of the ACLE and a standard likelihood where u_i is assumed to be 0, i.e., a likelihood that fails to take left truncation into account. We fitted survival models to random samples of individual lifetimes using the MMLE and ACLE methods, and the standard method. This procedure was repeated for different sample sizes and different parameterizations of the lifetime distribution. Subsequently, we applied the MMLE to a real-world case study: *Pinus pinea* L. seedling survival in central Spain.

3.1. *Simulated data*

We ran a simulation study from three populations of Gompertz-distributed lifetimes (Eq. 8) with parameters $\beta_{0,1} = -5$, $\beta_{1,1} = -0.003$; $\beta_{0,2} = -6$, $\beta_{1,2} = 0.003$; and $\beta_{0,3} = -7$, $\beta_{1,3} = 0.001$, respectively, where the second subscript stands for the population. These distributions allow exploring the performance of the proposed methods across a realistic range of parameter values.

For each distribution, we tested three sample sizes, 50, 100 and 200, following a censoring scheme over 1800 days, where v_1 and v_2 were successive measurements taken 180 days apart, such that:

$$v_1 \in (0, 180, 360, 540, 720, 900, 1080, 1260, 1440, 1620)$$
$$v_2 \in (180, 360, 540, 720, 900, 1080, 1260, 1440, 1620, 1800)$$

In order to evaluate the methods for continuous and discrete distributions of U, simulations were adapted accordingly. Each observation of each sample was attributed a lifetime T following the above-mentioned Gompertz distributions. Similarly, each observation was also given a u_i. These u_i were assumed to come from a Gaussian distribution with mean $\mu = 60$ and variance $\sigma^2 = 225$ in the continuous case or from a binomial distribution with a number of trials $n = 120$ and a probability of $\pi = 0.5$ for the discrete case. Note that these distributional assumptions are only made for the sake of the example. The approach is theoretically well suited for any distribution.

Individuals with $T > 1800 + u_i$ were considered as right-censored observations. Given that life history prior to selection is unknown in actual experiments, the lifetimes of those seedlings dying before selection cannot be used in the analysis. In order to have balanced samples, seedlings were generated as many times as necessary to obtain either 50, 100 or 200 observations at the beginning of the experiment.

We simulated 10,000 samples for each combination of lifetime distribution and sample size. Parameters were estimated using a standard maximum likelihood estimator, i.e. not considering left truncation, and using the ACLE and the MMLE methods. With regard to the latter, the survivor function to use in the Likelihood 10, i.e. continuous case, is that of Eq. 9. Alternatively, the survivor function in the discrete case is:

$$S(v_{\cdot,i} + u_i \mid T > u_i) = e^{-\sum_{t=u_i}^{v_{\cdot,i}+u_i} e^{\beta_0+\beta_1 t}} \tag{13}$$

Likelihood functions 10 and 11 were integrated over u_i using numerical methods. In the continuous case, a 15-point Gauss-Hermite quadrature was used [see 12, Section 20.7-3]. Note that this was possible —and convenient— because u_i was normally distributed. For other theoretical distributions, Monte Carlo methods could offer a reasonable alternative. As regards the discrete likelihood, integration was computed over all possible values of u_i.

The simulations were run under the assumption that the true parameters of the distribution of U were unknown and that these parameters were estimated from an independent sample. For each sample of T, an independent sample of the same size was randomly drawn from the distribution of U. This sample was used to obtain estimates of μ and σ^2 for the continuous case and π for the discrete one. The marginal likelihoods were based on those parameter estimates and not the true parameters themselves.

The Monte Carlo simulations provided empirical distributions of $\hat{\beta}_{0,\cdot}$ and $\hat{\beta}_{1,\cdot}$ both when left truncation was not considered and when the ACLE and MMLE were applied. The mean of these distributions was compared to the true parameters $\beta_{0,\cdot}$ and $\beta_{1,\cdot}$ used to simulate the data. The relative bias $B_{\boldsymbol{\beta}\cdot}$ was assessed as:

$$B_{\boldsymbol{\beta}\cdot} = \frac{\boldsymbol{\beta}\cdot - \bar{\hat{\boldsymbol{\beta}}}\cdot}{\boldsymbol{\beta}\cdot} \tag{14}$$

The robustness of the MMLE was additionally tested running the same Monte Carlo simulations but allowing for a deliberate misspecification of the actual distribution of U used to generate the samples. Parameters μ and π were set to 85 and 0.7 in the continuous and discrete cases, respectively.

The estimators were also evaluated as to check whether the empirical coverage of their confidence intervals coincides with their nominal values. This procedure helps determine to what extent an estimator's assumed distribution is a realistic hypothesis. We used the usual assumption that the estimators are asymptotically normally distributed. The nominal coverage is the probability that the true $\boldsymbol{\beta}\cdot$ is contained by the confidence intervals for different levels of significance $(1-\alpha)$, i.e. under the aforementioned assumption, the nominal coverage would be described by a normal distribution such that $N(\hat{\boldsymbol{\beta}}\cdot, \hat{\Omega})$, with $\hat{\Omega}$ the estimated variance-covariance matrix of the parameters. The empirical coverage can be assessed by checking the proportion of simulations where the true $\boldsymbol{\beta}\cdot$ falls within the confidence intervals defined according to that normal distribution for the whole range of $1-\alpha$. If the assumption used to define these confident intervals holds then nominal and empirical coverage should coincide. For example, $\boldsymbol{\beta}\cdot$ would be included in the confidence intervals in approximately 95% of the simulations for $\alpha = 0.05$ should the assumption be met.

Table 1. Evolution of studied cohorts of *P. pinea* seedlings.

	Cohorts					
date	Fall 2004	Fall 2005	Fall 2006	Spring 2007	Spring 2008	Fall 2009
2005-02-28	42	-	-	-	-	-
2005-06-14	3	-	-	-	-	-
2005-12-27	0	34	-	-	-	-
2006-05-19	0	9	-	-	-	-
2007-01-10	0	3	5665	-	-	-
2007-07-03	0	3	2179	170	-	-
2008-02-21	0	2	623	19	-	-
2008-08-27	0	2	519	16	11	-
2009-02-16	0	2	480	14	9	-
2009-07-22	0	2	456	11	7	-
2010-03-25	0	1	422	9	7	40
2011-02-22	0	1	401	9	6	6
2011-11-08	0	0	395	7	6	3

3.2. *Case study*

Data from the *P. pinea* regeneration experimental site monitored by the Forest Research Center of the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CIFOR) were used as a real-world case study. The site is located in a *P. pinea* pure even-aged stand in the Northern Plateau of Spain (41°28' N, 4°43'W), at an altitude of 700 m a.s.l. The area is representative of the average regional conditions, Mediterranean-continental climate and sandy soils. Specifically, the experimental plots were established in a 120-year-old forest of average fertility. Release fellings were carried out in 2002-2003. Among the various experiments that took place at this study site, the trial devoted to the analysis of *P. pinea* seedling emergence and survival is the one we selected as a case study.

The seedling survival trial was set up in 2003. It consisted of six 60 m × 80 m plots (0.48 ha each), with a 7.5 m surrounding buffer area. Two regeneration systems were tested, a shelterwood system and the seed-tree system, with three replications each. Additionally, a control plot (no fellings) was also established.

Stand densities ranged from 46 stem·ha^{-1} and a basal area of 7.0 m^{2}ha^{-1} to 75 stem·ha^{-1} and a basal area of 12.3 m^{2}ha^{-1}. In all six plots, twenty 3 m × 3 m subplots were established on a regular grid for a total of 120 subplots. From 2004 to 2011, seedling emergence and survival were measured twice a year in each subplot, except in 2010, when only one measure was taken. At each occasion, new seedlings were identified and dead individuals were recorded. Seedlings usually appeared grouped in autumn or spring cohorts as a result of the favorable periods for germination of *P. pinea* seeds. Recorded individuals from detected cohorts are shown in Table 1. A model of seedling emergence as a function of climate and stand conditions was fitted by [16]. A brief description of this model is annexed to this paper (see Appendix B).

The three types of data incompleteness described in the previous section were present in the dataset. Given the spaced observation intervals, data were interval-censored. A fraction of seedlings survived beyond the follow-up, implying right censoring. Provided that seedling selection did not necessarily coincide with seedling emergence, we can considered that all observations were left-truncated in practice, i.e. $T = U + V$ with $u_i > 0\ \forall i$.

Seedling lifetimes were assumed to follow a Gompertz distribution. For the sake of the example, no covariates were considered to define the hazard. Thus, the survival function used in this case study adopted the same functional form than that of Eq. 13:

$$S_i(t) = e^{-\sum_t e^{\gamma_0+\gamma_1 t}} \tag{15}$$

where $\boldsymbol{\gamma} = \begin{pmatrix}\gamma_0 & \gamma_1\end{pmatrix}^T$ are estimable parameters. Vector $\boldsymbol{\gamma}$ was first estimated through the maximization of the likelihood function in Eq. 4, which assumes no left truncation. The MMLE was then applied. The distribution of u_i conditional on $u_i \geq 0$ was estimated using the model of [16] and Bayes' theorem as shown in Eq. 12.

This model predicts the daily probability that a *P. pinea* seedling i emerges on day j of calendar time as a function of climate variables and the relative location of the seeds with respect to adult trees. Climate data were gathered from a nearby meteorological station (data available at *http://www.inforiego.org*). The first dates allowed for germination were September 1st for the autumn cohorts, and February 1st, for the spring cohorts. It can be safely assumed that by these dates: (i) seed dispersal has been largely completed [17]; and (ii) germination has not yet begun due to climatic limitations [16]. The last possible germination dates were those of seedling selection.

The estimated $\boldsymbol{\gamma}$ from both approaches were used to produce predictions on $S(t)$ with $u_i = 0$. This analysis made it possible to determine the impact of parameter bias on model predictions if the model was applied to independent data. Results were graphically contrasted. The goodness-of-fit of both models was assessed through the models' AICs and visually comparing the evolution of the actual survival fraction for each cohort with the predicted survival probability. In order to allow for this comparison, MMLE predictions were marginalised over the distribution of u_i and the resulting values were standardized so that survival at the time of selection equals 1.

4. Results

4.1. *Simulated data*

The standard maximum likelihood estimator produced biased $\hat{\beta}_{0,\cdot}$ for the first and second populations. Specifically, $\beta_{0,1}$ and $\beta_{0,2}$ were under and overestimated, respectively. In contrast, the bias in the estimate of $\beta_{0,3}$ was negligible. Simulations from the continuous and discrete distributions yielded similar results. The relative bias did not relevantly differ across sample sizes.

The analytical correction applied through the ACLE proved effective in correcting those bias. The relative bias after correction was below or far below 0.5% in most cases. The only exception was $\hat{\beta}_{0,2}$ which appeared to be asymptotically unbiased and not purely unbiased, with $B_{\beta_{0,2}}$ ranging from around -1.6% ($n = 50$) to -0.7% ($n = 200$) in the discrete case and from -1.5% ($n = 50$) to -0.3% ($n = 200$) in the continuous one.

Similarly, there was no apparent bias in the parameter estimates when the MMLE was used. Again, the asymptotical pattern was observed for $\hat{\beta}_{0,2}$, the bias decreasing with increasing sample sizes: -1.6% ($n = 50$) to -0.4% ($n = 200$) in the discrete case; -1.5% ($n = 50$) to -0.3% ($n = 200$) in the continuous one.

Regardless of the population, all estimators were asymptotically unbiased for β_1. Nevertheless, the bias in $\hat{\beta}_{1,2}$ was particularly high for the smallest sample size ($n = 50$) in the continuous and discrete cases ($B_{\beta_{1,2}} > 14\%$). Further details are shown in Table 2.

The bias in the estimates of $\beta_{0,1}$ and $\beta_{0,2}$ were evident in view of the differences

Table 2. Maximum likelihood parameter estimates for the simulation study. True values: $\boldsymbol{\beta}_{\cdot,1} = (-5, -0.003)$; $\boldsymbol{\beta}_{\cdot,2} = (-6, 0.003)$; $\boldsymbol{\beta}_{\cdot,3} = (-7, 0.001)$

		Standard ML			ACLE			MMLE		
Distribution	Estimate	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
	$\widehat{\mathrm{E}[\hat{\beta}_{0,1}]}$	-5.1801	-5.1845	-5.1873	-4.9988	-5.0045	-5.0085	-4.9838	-4.9876	-4.9925
	$\widehat{\mathrm{E}[\hat{\beta}_{1,1}]}$	-0.0030	-0.0030	-0.0030	-0.0030	-0.0030	-0.0030	-0.0031	-0.0030	-0.0030
Discrete	$\widehat{\mathrm{E}[\hat{\beta}_{0,2}]}$	-5.8948	-5.8694	-5.8513	-6.1003	-6.0644	-6.0394	-6.0956	-6.0459	-6.0250
	$\widehat{\mathrm{E}[\hat{\beta}_{1,2}]}$	0.0034	0.0033	0.0031	0.0034	0.0033	0.0031	0.0034	0.0032	0.0031
	$\widehat{\mathrm{E}[\hat{\beta}_{0,3}]}$	-6.9844	-6.9669	-6.9555	-7.0496	-7.0298	-7.0170	-7.0486	-7.0247	-7.0131
	$\widehat{\mathrm{E}[\hat{\beta}_{1,3}]}$	0.0011	0.0010	0.0010	0.0011	0.0010	0.0010	0.0011	0.0010	0.0010
	$\widehat{\mathrm{E}[\hat{\beta}_{0,1}]}$	-5.1649	-5.1744	-5.1761	-4.9795	-4.9919	-4.9945	-4.9835	-4.9957	-4.9963
	$\widehat{\mathrm{E}[\hat{\beta}_{1,1}]}$	-0.0031	-0.0030	-0.0030	-0.0031	-0.0030	-0.0030	-0.0031	-0.0030	-0.0030
Continuous	$\widehat{\mathrm{E}[\hat{\beta}_{0,2}]}$	-5.8881	-5.8504	-5.8348	-6.0921	-6.0416	-6.0200	-6.0921	-6.0470	-6.0212
	$\widehat{\mathrm{E}[\hat{\beta}_{1,2}]}$	0.0034	0.0032	0.0031	0.0034	0.0032	0.0031	0.0034	0.0032	0.0031
	$\widehat{\mathrm{E}[\hat{\beta}_{0,3}]}$	-6.9826	-6.9574	-6.9496	-7.0479	-7.0200	-7.0110	-7.0518	-7.0269	-7.0116
	$\widehat{\mathrm{E}[\hat{\beta}_{1,3}]}$	0.0011	0.0010	0.0010	0.0011	0.0010	0.0010	0.0011	0.0010	0.0010

between empirical and nominal coverage probability (Fig. 2). In contrast, parameters estimated through the MMLE exhibited empirical coverages that were close to their nominal values. Empirical coverage probability for the ACLE estimates of $\beta_{0,\cdot}$ was observed to slightly depart from its nominal coverage. No departure was found in the case of $\beta_{1,\cdot}$.

The fact that the bias tends to zero and the parameters themselves to a given value as the sample size increases seems to indicate that the MMLE is consistent. On top of that, the empirical coverage probability coinciding with its nominal values means that parameter estimates tend to be normally distributed and that the estimated variance-covariance matrix of that distribution can be unbiasedly assessed from the Hessian matrix of the likelihood at the maximum. This would suggest that the estimator's asymptotic normality property is likely to be met as well. Note that these findings are not a formal proof of the asymptotic properties of the MMLE.

The bias in $\hat{\beta}_{0,\cdot}$ induced by the standard approach implies that the hazard decreases over time faster than it does in reality. As a result, overestimation of the probability of survival was expected. Taking our simulations with the first lifetime distribution and discrete case as an example, the survival probability was overestimated by more than 45% after 1800 days when the left truncation was not taken into account, with $\hat{S}(1800) = 0.156205$ compared to the true probability of $S(1800) = 0.106903$. On the other hand, the MMLE in the continuous case yielded an estimated survival probability of 0.1076451 at $t = 1800$, which can be considered unbiased in practice.

Parameter estimates from the model fit computed to test MMLE robustness are shown in Table 3. $\boldsymbol{\beta_{0,\cdot}}$ were biased for all studied parameter sets in both the discrete and continuous cases with signs opposite to those found under the standard approach. Those bias were lower — though still significant — than in the standard approach for $\beta_{0,1}$, of the same magnitude for $\beta_{0,2}$ and higher for $\beta_{0,3}$. Concerning $\boldsymbol{\beta_{1,\cdot}}$, bias was negligible in all cases. Bias did not notably decrease with increasing sample size.

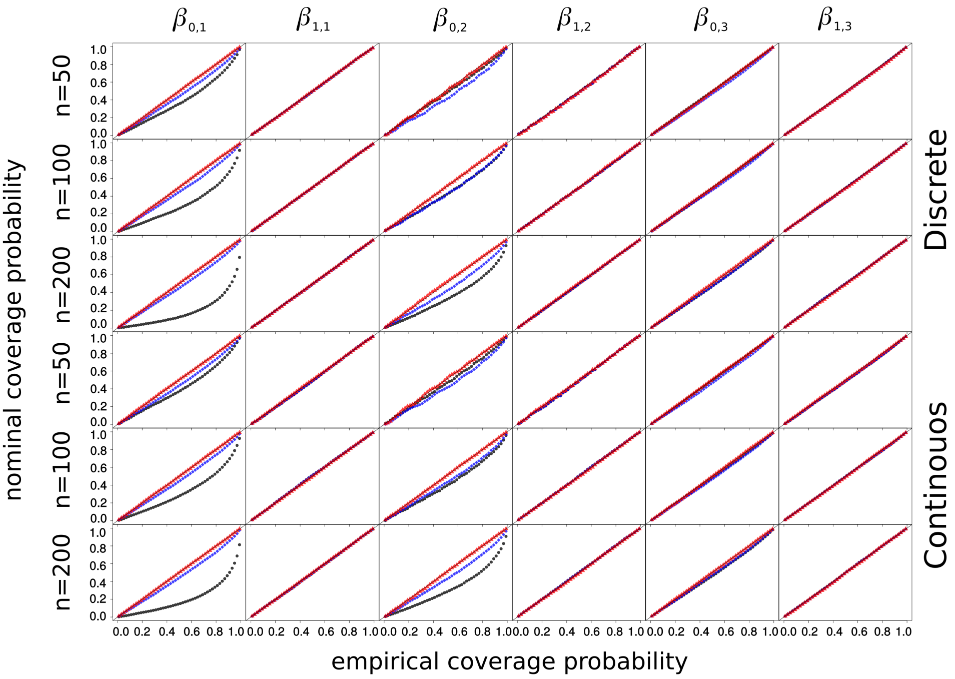

Figure 2. Nominal and empirical coverage probability of parameter estimates based on 10,000 simulations for different sample sizes and different parameterizations when left truncation is not considered in model likelihood (circles), when the ACLE is used (diamonds) and when the MMLE is applied (triangles). Symbols overlap in some panels.

Table 3. Maximum likelihood parameter estimates for the robustness test on the MMLE. True values: $\boldsymbol{\beta}_{\boldsymbol{\cdot},\mathbf{1}} = (-5, -0.003)$; $\boldsymbol{\beta}_{\boldsymbol{\cdot},\mathbf{2}} = (-6, 0.003)$; $\boldsymbol{\beta}_{\boldsymbol{\cdot},\mathbf{3}} = (-7, 0.001)$

Distribution	Estimate	$n = 50$	$n = 100$	$n = 200$
Discrete	$\widehat{\mathrm{E}[\hat{\beta}_{0,1}]}$	-4.9099	-4.9192	-4.9196
	$\widehat{\mathrm{E}[\hat{\beta}_{1,1}]}$	-0.0031	-0.0030	-0.0030
	$\widehat{\mathrm{E}[\hat{\beta}_{0,2}]}$	-7.0740	-7.0483	-7.0366
	$\widehat{\mathrm{E}[\hat{\beta}_{1,2}]}$	0.0011	0.0010	0.0010
	$\widehat{\mathrm{E}[\hat{\beta}_{0,3}]}$	-6.1853	-6.1244	-6.0989
	$\widehat{\mathrm{E}[\hat{\beta}_{1,3}]}$	0.0034	0.0032	0.0031
Continuous	$\widehat{\mathrm{E}[\hat{\beta}_{0,1}]}$	-4.9115	-4.9172	-4.9225
	$\widehat{\mathrm{E}[\hat{\beta}_{1,1}]}$	-0.0031	-0.0030	-0.0030
	$\widehat{\mathrm{E}[\hat{\beta}_{0,2}]}$	-7.0700	-7.0486	-7.0344
	$\widehat{\mathrm{E}[\hat{\beta}_{1,2}]}$	0.0011	0.0010	0.0010
	$\widehat{\mathrm{E}[\hat{\beta}_{0,3}]}$	-6.1746	-6.1229	-6.0972
	$\widehat{\mathrm{E}[\hat{\beta}_{1,3}]}$	0.0034	0.0032	0.0031

Table 4. Maximum likelihood parameters estimates when left truncation was not considered and when the MMLE was used in the *P. pinea* case study. The standard errors appear in parentheses.

Parameter	No left truncation	MMLE
$\hat{\beta}_0$	-4.8421 (0.0369)	-4.5703 (0.0459)
$\hat{\beta}_1$	-0.0029 (0.0002)	-0.0030 (0.0002)

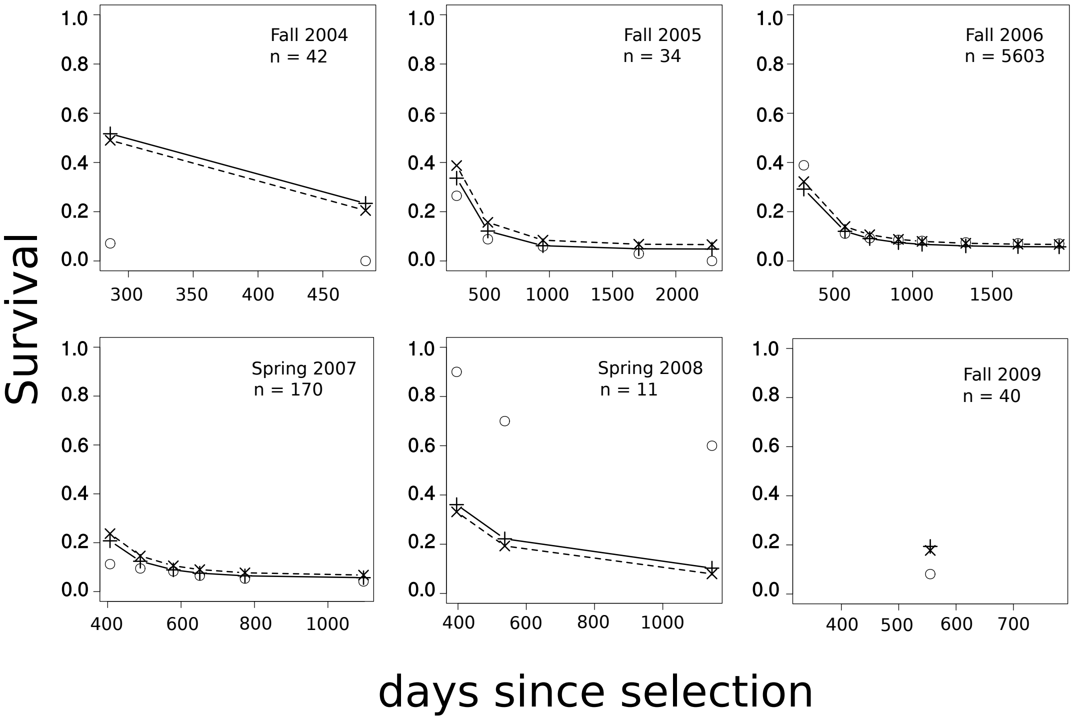

Figure 3. Comparison of the actual survival fractions of *P. pinea* seedlings for all studied cohorts over time (circles) and the respective survival predictions from a model that does not consider left truncation (dashed line) and from a model where likelihood maximization was assisted by the MMLE (solid line).

4.2. *Case study*

Concerning the case study, the model fit using the MMLE proved to be better than that of the standard approach with AICs of 12,127.96 and 12,288.43, respectively. Parameter estimates from both models are shown in Table 4. Interestingly, the parameter associated with time was negative, suggesting that there was an increasing resistance of seedlings over time.

The Gompertz assumption for the distribution of the lifetimes behaved reasonably well for most cohorts (Fig. 3). Model predictions were less accurate shortly after selection, which may explain the overestimation of survival for that originated during the fall of 2004. Survival probability corresponding to the 2008 cohort was largely underestimated.

The probability of surviving up to 1800 days was 0.066530 when left truncation was not considered, and 0.032025 when the MMLE was applied. Providing that the latter results from unbiased parameter estimates, overestimation of seedling survival would be over 100% for the former at the end of the *P. pinea* experiment (Fig. 4).

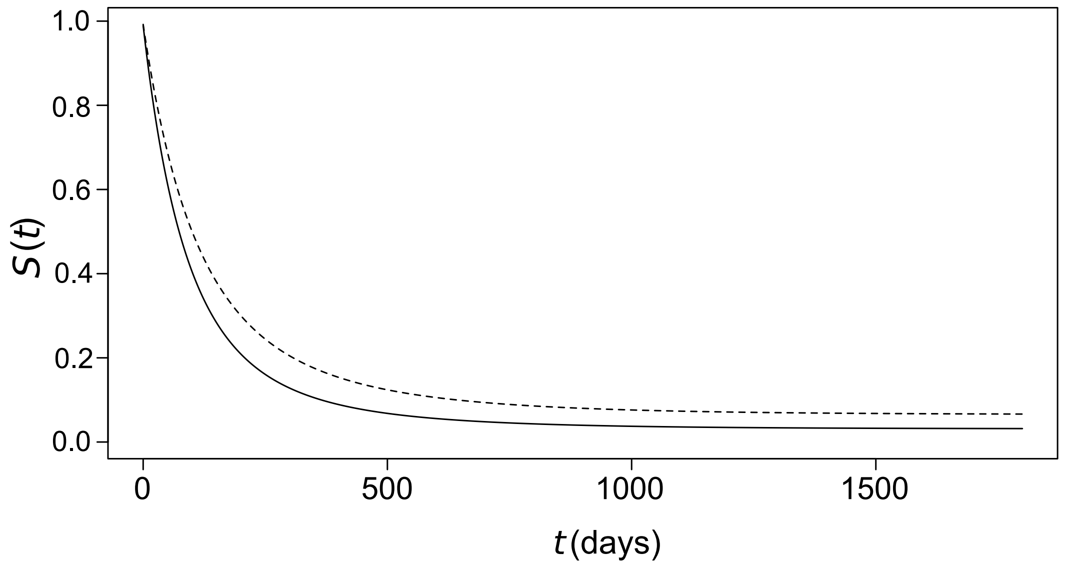

Figure 4. Expected value of the probability of *P. pinea* seedling survival from a model fitted without considering left truncation (dashed line) and from a model where likelihood maximization was assisted by the MMLE (solid line).

Beyond that time, survival does not notably change, since the hazard asymptotically tends to 0. The survival fraction when $t \to \infty$ can be analytically approached as $\lim_{t\to\infty} S(t) = e^{-\int_{t=0}^{\infty} e^{\gamma_0+\gamma_1 t}} = e^{-e^{\gamma_0}/\gamma_1}$. At the limit, the uncorrected approach predicts a survival probability of 0.065819 whereas the MMLE only reaches 0.031694.

5. Discussion

The MMLE technique presented here proved adequate to deal with left-truncated data when exact time of individual occurrence is unknown. The current approach extends the use of models based on lifetime techniques to a data type potentially present in survival studies in forestry. One aspect contributing to the originality of this technique is that it makes it possible to estimate the distribution of time of selection from external models.

The MMLE is also advantageous with regard to the ACLE method. Because the MMLE consists in a maximum likelihood estimation *sensu stricto*, an estimate of the variance-covariance matrix of the parameters is available from the Hessian at likelihood maximum. This is not the case when the ACLE is used, since this method just corrects the biased maximum likelihood estimates. Using the estimated variance-covariance matrix from these biased estimates as a proxy for the unknown variance-covariance matrix of the corrected parameters may result in unrealistic confidence intervals. Actually, this substitution is very likely to cause the empirical coverage probability distortion observed in the present study for the ACLE estimates. On the contrary, the noticeable agreement between empirical and nominal coverage probability in the case of MMLE not only indicated that this estimator was unbiased, but that confidence intervals were correctly estimated as well.

MMLE parameter estimates proved sensitive to a moderate misspecification of the distribution of U. Robustness was tested by increasing the mean of the true normal distribution by roughly 30%. This caused the bias to have the opposite sign to that attained using the standard approach. This is not surprising as the latter can be seen as a fit where time of selection is reduced —suppressed in fact—. The magnitude of the bias was very variable depending on the set of parameters though, being sometimes even higher than that detected under the standard approach. Given that this behaviour is difficult to ascertain, no strong conclusions can be drawn in this regard.

The MMLE approach is of interest when calendar time constitutes part of the hazard. This is the case of the Gompertz model used in the present paper, where calendar time is explicit in the model baseline. An advantage of the MMLE is that it can be easily extended to any model type that presents time-dependent hazards, such us accelerated failure-time (AFT) models, for example. In contrast, proportional hazards models that assume an exponential distribution of lifetimes are examples where the MMLE would make no sense. Because calendar time does not affect the hazard in these models, time of occurrence is not required and parameter estimates from standard likelihood maximization would be unbiased in the event of left truncation.

Calendar time is also the most likely reason why some parameter estimates, namely $\beta_{1,2}$, appeared unexpectedly *biased* with small sample sizes. Although these biases had little impact on the inference, we want to recall that no bias at all was expected in the estimation of $\beta_{1,\cdot}$ when the standard approach was used. However, the second lifetime distribution tested in our simulation study implies a sharp decrease of the survival probability over time. This means that the proportion of individuals that die before the beginning of the experiment is potentially higher than under other parameterizations. Because maximum likelihood estimates are only asymptotically unbiased, remaining individuals at time of selection when sample size is small may be insufficient for the algorithm to converge to the true parameter values. The same rationale applies for the unexpected *unbiased* estimate of $\beta_{0,3}$ through the standard likelihood. The parameterization of the lifetime distribution used in this case led to low rates of mortality over time. In turn, individuals dying before time of selection may be relatively small, resulting in data being close to those of a left truncation-free experiment. In consequence, the estimation bias tends to disappear.

There are three essential points to take into account to attain valid inference from our approach. The first is whether or not the distribution of time of selection is accurate enough. Ideally, a predicted empirical distribution like the one shown in our case study is to be used. Provided that the model that gives rise to such a distribution is unbiased, predictions of time of selection should closely mimic reality. Otherwise, a theoretical distribution could be a convenient alternative. In this case, even if we had a good prior knowledge about the occurrence process, the challenge to parameterize the distribution would still remain. One possibility would be to arbitrarily fix the value of the parameters, as in our simulation example. This would be a realistic option but it may be hindered by the estimator's weak robustness. It therefore seems sensible to recommend awareness on any potential and significant misspecification of the distribution of time of selection. Another possibility would have been to estimate them out of the maximisation of the likelihood. This alternative would somehow have resembled the use of random effects in mixed models, which can in turn be seen as a true Bayesian approach: on the one hand, strong assumptions are made on the random effect distribution (e.g. Gaussian-distributed; [14]); on the other hand, the likelihood is integrated over this distribution and one of its parameters (i.e. the variance) is estimated together with the fixed effects [22, p.62]. While this approach may have been feasible for a theoretical distribution of U, estimation of all parameters of a more complex model, like the one of our case study, would have led to a very complex, likely intractable likelihood.

The second issue regarding inference is our assumption of independent delayed entry. Violation of this assumption arises when time of selection and time to event are not independent. As a result, short-living individuals are less likely to be included in the experimental sample because they have more chances of being already dead at the time of selection [15]. This would cause more resistant individuals to be favored in the dataset. In theory, our case study may not be free of this kind of bias as very short-living individuals could not be considered. However, there are not evidences that suggest that the process causing seedlings to perish is different in individuals where $T < U$ than in those that were actually observed, especially in such a short period of time. Consequently the same hazard function could be used in Eq. 5 for $S(t)$ and $S(u)$, implying in turn that the assumption should be met. If U had been much higher, most excluded individuals would have been at the seedling stage whereas most selected ones would have been saplings. This transition implies dramatic physiological changes that affect survival. In this case, the sample would not be representative of the whole population, causing the independent delayed entry assumption to fail.

A third issue is the sample size. The construction of the confidence intervals relies on the assumption of asymptotically normally distributed parameter estimates. With small samples, it can reasonably be expected that the empirical coverage of the confidence intervals will not reach its nominal value. Although accuracy tended to increase along n, our tests showed that MMLE performs relatively well with sample size as small as $n = 50$. At $n = 200$, bias could be considered as negligible. Nevertheless, this may not be the case when other distributions of lifetimes are used. Consequently, care should be taken when generalizing the approach.

Focusing now on our case study, the implications of correctly setting the likelihood in the presence of left truncation are evident. The seemingly small differences in long-term survival between both approaches can however be crucial in this and other systems where seedling mortality is considerable. In their comprehensive study on *P. pinea* natural regeneration, [18] built a multistage model to predict seedling establishment. The model consisted of several submodels, each of them related to a key regeneration process: seed production, dispersal, germination, predation and seedling mortality. In that study, the mortality submodel was fitted to a reduced database through the non-truncated version of the likelihood presented in the current paper. The resulting multistage model was used to optimize felling intensity in order to guarantee natural regeneration success. In view of the overestimation in survival derived from the non-truncated approach, current optimised fellings are surely optimistic. Given the extremely low survival expectation for *P. pinea* seedlings, prediction errors like those presented here may encourage managers to favor silvicultural schemes that actually prevent acceptable regeneration standards from being reached.

Very few unmarked individuals were found dead when they were observed for the first time. In this regard, the notable frost tolerance of *P. pinea* seedlings [21] suggests that winter mortality of individuals must only be of limited extent. Therefore, the predicted survival probability for autumn cohorts observed in mid-winter is expected to be well defined as we have little loss of information due to left-truncation. However, seedlings of this species have been reported to be much more susceptible to drought [5, 19]. Consequently, lifetimes of individuals that emerged in spring but were not observed until late summer could be more prone to exhibit $T < U$. In our example, only the 2008 spring cohort was thus concerned. As a matter of fact, predictions on seedling survival for this cohort were highly overestimated when the MMLE was considered. Despite not invalidating the overall results due to its small weight in the likelihood, caution must definitely be taken in examples where a relevant number of individuals are not expected to survive beyond the time of selection.

Prediction of a close-to-stationary survival fraction different from 0 is possible due to the parameterization of the Gompertz hazard attained in this study. According to [5], *P. pinea* seedlings gain resistance against abiotic stress over time, which translates into a decreasing hazard with seedling age. The more negative γ_1 is, closer to zero the limit of the hazard gets as t approaches infinity. The consequence is that the asymptote of the probability distribution is less than 1. This means that some individuals will never die as seedlings -exactly what was observed-. This phenomenon could have been regarded as right-censored observations. The problem is that we would have had to set a prior critical threshold probability - or age - beyond which individuals are considered as fully established as adult trees. Conversely, allowing negative parameter values provides a more flexible approach. Classic literature on lifetime analysis supports this alternative [2], which is common to other AFT models.

The present approach is obviously not limited to seedling survival studies. Almost all processes involved in natural regeneration imply data that can be explored through survival analysis techniques. For example, if seed germination is taken as the event of interest, time from seed release to germination can be considered as a lifetime [e.g. 16]. Having a good knowledge of pre-selection lifetime history would be of interest as seeds of a number of species need a period of dormancy to germinate [4]. Because it would be impossible to start the experiment at the exact time the seed rain starts, left-truncation would occur and the exact time of occurrence (seed release) would normally be unknown. However, models that predict the probability of seed release do exist for a number of species [e.g. 17], and the MMLE then applies. The same would be true if the event of interest were seed predation. Another interesting application could be the prediction of occurrence of forest wildfires. Aside from climate and human activities, fuel accumulation over time has been reported as a possible factor contributing to fire risk [23]. Very often, current models predict fire risk as a function of fuel type and abundance through a logistic approach [e.g. 9]. When information on fuel status is not available, logistic models would be of no use. Lifetime analysis could then be a convenient alternative. The time elapsed between the last major disturbance and wildfire would be the lifetime in this case. Given that the monitoring usually begins some time after a major disturbance, these lifetimes would be left-truncated. However, it is unlikely that the time of those disturbances is precisely known for each target stand. If their distribution can be inferred from national forest inventories or other sources, MMLE applies. Other processes such as individual-tree mortality are less adaptable to MMLE. In principle, the framework is adequate (time-dependent hazard, unknown time of occurrence). Unfortunately, the use of the MMLE is hindered by the generally long time elapsed between time of occurrence and time of selection, which makes the independent delayed entry assumption likely to fail.

Geolocation information

The site where the case study of this paper is based on is located in the Northern Plateau of Spain (41°28' N, 4°43'W).

Acknowledgements

We are much obliged to the INIA-CIFOR field staff and, most of all, to Guillermo Madrigal, Mar Conde, Enrique Garriga and Santiago de Blas. Special thanks to Jesús Silveira for advice in preliminary mathematical developments. We are also grateful to the Forest Service of the Junta de Castilla y León for permission and support to carry out the field experiment. We wish to express our gratitude to two anonymous reviewers whose comments significantly contributed to improve the quality of this paper.

Funding

This work was supported by the INIA under Grant RTA2007-00044; the European Forest Institute - Mediterranean office - (EFIMED) under Grant EFIMED STSV-2012; and the Agence National de la Recherche (ANR) under Grant ANR-11-LABX-0002-01.

References

[1] V. Addona and D. Wolfson, *A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up*, Statistics in Medicine 12 (2006), pp. 267–284.

[2] P.D. Allison, *Event history analysis: regression for longitudinal event data*, no. no. 07-046 in Sage university papers. Quantitative applications in the social sciences, Sage Publications, Beverly Hills, Calif, 1984.

[3] M. Asgharian, D. Wolfson, and X. Zhang, *Checking stationarity of the incidence rate using prevalent cohort survival data*, Statistics in Medicine 25 (2005), pp. 1751–1767.

[4] C. Baskin and J. Baskin, *Seeds. Ecology, biogeography, and evolution of dormancy and germination*, Academic Press, San Diego, 2001.

[5] R. Calama, J. Puértolas, R. Manso, and M. Pardos, *Defining the optimal regeneration niche for Pinus pinea L. through physiology-based models for seedling survival and carbon assimilation*, Trees 29 (2015), pp. 1761–1771. Available at http://link.springer.com/10.1007/s00468-015-1257-5.

[6] C. Collet and G. Le Moguédec, *Individual seedling mortality as a function of size, growth and competition in naturally regenerated beech seedlings*, Forestry 80 (2007), pp. 359–370.

[7] R.J. Cook and P.J. Bergeron, *Information in the sample covariate distribution in prevalent cohorts*, Statistics in Medicine 30 (2011), pp. 1397–1409. Available at http://doi.wiley.com/10.1002/sim.4180.

[8] S. Datta, G.A. Satten, and J.M. Williamson, *Consistency and Asymptotic Normality of Estimators in a Proportional Hazards Model with Interval Censoring and Left Truncation*, Annals of the Institute of Statistical Mathematics 52 (2000), pp. 160–172. Available at http://link.springer.com/10.1023/A:1004197201989.

[9] J.R. González, M. Palahí, A. Trasobares, and T. Pukkala, *A fire probability model for forest stands in Catalonia (north-east Spain)*, Annals of Forest Science 63 (2006), pp. 169–176. Available at http://www.edpsciences.org/10.1051/forest:2005109.

[10] S.R. Green, M.A. Arthur, and B.A. Blankenship, *Oak and red maple seedling survival and growth following periodic prescribed fire on xeric ridgetops on the cumberland plateau*, Forest Ecology and Management 259 (2010), pp. 359–370.

[11] E.N. Hane, *Indirect effects of beech bark disease on sugar maple seedling survival*, Canadian Journal of Forest Research 33 (2003), pp. 807–813.

[12] G.A. Korn and T.M. Korn, *Mathematical handbook for scientists and engineers. Definitions, theorems, and formulas for reference and review*, Dover Publications, Mineola, New York, USA, 1968.

[13] T. Kuuluvainen and T. Pukkala, *Effect of scots pine seed trees on the density of ground vegetation and tree seedlings*, Silva Fennica 23 (1989), pp. 159–167.

[14] N.M. Laird and J.H. Ware, *Random-Effects Models for Longitudinal Data*, Biometrics 38 (1982), p. 963. Available at http://www.jstor.org/stable/2529876?origin=crossref.

[15] J.F. Lawless, *Statistical models and methods for lifetime data*, 2nd ed., Wiley series in probability and statistics, Wiley-Interscience, Hoboken, N.J, 2003.

[16] R. Manso, M. Fortin, R. Calama, and M. Pardos, *Modelling seed germination in forest tree species through survival analysis. The Pinus pinea L. case study*, Forest Ecology and Management 289 (2013), pp. 515–524. Available at http://linkinghub.elsevier.com/retrieve/pii/S0378112712006251.

[17] R. Manso, M. Pardos, C.R. Keyes, and R. Calama, *Modelling the spatio-temporal pattern of primary dispersal in stone pine (Pinus pinea L.) stands in the Northern Plateau (Spain)*, Ecological Modelling 226 (2012), pp. 11–21. Available at http://linkinghub.elsevier.com/retrieve/pii/S0304380011005709.

[18] R. Manso, T. Pukkala, M. Pardos, J. Miina, and R. Calama, *Modelling Pinus pinea forest management to attain natural regeneration under present and future climatic scenarios*, Canadian Journal of Forest Research 44 (2014), pp. 250–262. Available at http://www.nrcresearchpress.com/doi/abs/10.1139/cjfr-2013-0179.

[19] C. Mayoral, R. Calama, M. Sánchez-González, and M. Pardos, *Modelling the influence of light, water and temperature on photosynthesis in young trees of mixed Mediterranean forests*, New Forests 46 (2015), pp. 485–506. Available at http://link.springer.com/10.1007/s11056-015-9471-y.

[20] C.E. McCulloch, S.R. Searle, and J.M. Neuhaus, *Generalized, Linear, and Mixed Models*, 2nd ed., Wiley series in probability and statistics, Wiley, Hoboken, N.J, 2008.

[21] M. Pardos, J. Climent, H. Almeida, and R. Calama, *The role of developmental stage in frost tolerance of Pinus pinea L. seedlings and saplings*, Annals of Forest Science 71 (2014), pp. 551–562. Available at http://link.springer.com/10.1007/s13595-014-0361-9.

[22] J. Pinheiro and D. Bates, *Mixed effects models in S and S-PLUS*, Springer, New York, 2000.

[23] J. Piñol, K. Beven, and D. Viegas, *Modelling the effect of fire-exclusion and prescribed fire on wildfire size in Mediterranean ecosystems*, Ecological Modelling 183 (2005), pp. 397–409. Available at http://linkinghub.elsevier.com/retrieve/pii/S0304380004004995.

[24] C. Puerta-Piñero, J.M. Gómez, and F. Valadares, *Irradiance and oak seedling survival and growth in a heterogeneous environment*, Forest Ecology and Management 242 (2007), pp. 462–469.

[25] E. Rodríguez-García, F. Bravo, and T.A. Spies, *Effects of overstorey canopy, plant–plant interactions and soil properties on mediterranean maritime pine seedling dynamics*, Annals of Forest Science 262 (2011), pp. 244–251.

[26] C.E.J. Rose, D.B. Hall, B.D. Shiver, M.L. Clutter, and B. Borders, *A multilevel approach to individual tree survival prediction*, Forest Science 52 (2006), pp. 31–43.

[27] M. Woodroofe, *Estimating a Distribution Function with Truncated Data*, The Annals of Statistics 13 (1985), pp. 163–177. Available at http://projecteuclid.org/euclid.aos/1176346584.

[28] J.M. Yoder, E.A. Marschal, and D.A. Swanson, *The cost of dispersal: predation as a function of movement and site familiarity in ruffed grouse*, Behavioral Ecology 15 (2004), pp. 469–476.

Appendix A. Assessment of bias in parameter estimates

If $\mathcal{L}(\boldsymbol{\beta}, u_i)$ is the conditional likelihood 6, then the maximum likelihood estimator of $\boldsymbol{\beta}$ is the vector $\hat{\boldsymbol{\beta}}_{ML}$ that satisfies this condition:

$$\sum_i \frac{\partial ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ML}} = \mathbf{0} \tag{A1}$$

Using a first-order Taylor expansion at $u_i = 0$ provides a more tractable approximation of this condition:

$$\sum_i \frac{\partial ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ML}, u_i=0} + \sum_i u_i \frac{\partial^2 ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta} \partial u_i}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ML}, u_i=0} \approx \mathbf{0} \tag{A2}$$

When left truncation is not considered in the likelihood function, i.e., the u_i are arbitrarily set to 0, then the maximization of the function provides the estimator $\hat{\boldsymbol{\beta}}^*_{ML}$ which ensures that $\sum_i \frac{\partial ln(\mathcal{L}(\boldsymbol{\beta},u_i))}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*_{ML}, u_i=0} = \mathbf{0}$. Given that the expectation of u_i is strictly positive, the condition A2 is never met when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^*_{ML}$.

Corrected parameter estimates $\hat{\boldsymbol{\beta}}_{CORR}$ can be approximated using a first-order Taylor expansion around $\hat{\boldsymbol{\beta}}^*_{ML}$ such that

$$\sum_i \frac{\partial ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*_{ML}, u_i=0} + \sum_i \frac{\partial^2 ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*_{ML}, u_i=0} (\hat{\boldsymbol{\beta}}_{CORR} - \hat{\boldsymbol{\beta}}^*_{ML})$$
$$+ \sum_i u_i \frac{\partial^2 ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta} \partial u_i}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*_{ML}, u_i=0} \approx \mathbf{0} \tag{A3}$$

Since the first term is equal to $\mathbf{0}$, Eq. A3 simplifies to

$$\mathbf{H}(\hat{\boldsymbol{\beta}}_{CORR} - \hat{\boldsymbol{\beta}}^*_{ML}) = -\sum_i u_i \frac{\partial^2 ln(\mathcal{L}(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta} \partial u_i}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*_{ML}, u_i=0} \tag{A4}$$

where $\mathbf{H} = \sum_i \frac{\partial^2 ln(\mathcal{L}(\boldsymbol{\beta},u_i))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ is the Hessian matrix of the log-likelihood function evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^*_{ML}$ and $u_i = 0$. The u_i are unobserved. However, if their expectations can be somehow estimated, then a feasible estimator is obtained by substituting $\hat{\mathrm{E}}[u_i]$ for u_i:

$$\hat{\boldsymbol{\beta}}_{CORR} = \hat{\boldsymbol{\beta}}^*_{ML} - \mathbf{H}^{-1} \sum_i \hat{\mathrm{E}}[u_i] \frac{\partial^2 ln(\mathcal{L}_i(\boldsymbol{\beta}, u_i))}{\partial \boldsymbol{\beta} \partial u_i}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*_{ML}, u_i=0} \tag{A5}$$

Note that $-\mathbf{H}^{-1}$ is actually the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}^*_{ML}$ [cf. 20, p. 359]. The estimator A5 only requires the estimated expectations of the u_i and not the u_i themselves. Theoretically, this estimator can provide unbiased parameter estimates. However, variance-covariance matrix is unknown. The estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}^*_{ML}$, i.e. $-\mathbf{H}^{-1}$ can be used as a substitute although it is likely suboptimal.

Appendix B. Seedling emergence model

The following developments are based on [16]. In this paper, the probability of a given *P. pinea* seed remaining latent at time v was assessed through lifetime analysis. Let define this event as L, with associated probability $\Pr(L \mid v)$. Provided that $\Pr(L \mid v)$ was location-dependent and that v was conceived on a daily basis (i.e. discrete time), let us also assign indices j and k to the location and the day, respectively. The probability that one seedling emerges — event E — from that seed can be then expressed as $\Pr(E \mid v) = 1 - \Pr(L \mid v)$. Put in terms of the corresponding hazard function it results:

$$\Pr(L \mid v) = e^{-\sum_{k=0}^{k=v} h(\boldsymbol{z_{jk}}, \boldsymbol{\zeta})} \tag{B1}$$

where $\boldsymbol{z_{jk}}$ is a vector of possibly time-dependent covariates at location j and $\boldsymbol{\zeta}$ a vector of estimable parameters. Under the assumption of an exponential distribution of time to emergence v, the hazard $h(\boldsymbol{z_{jk}}, \boldsymbol{\zeta})$ was defined as:

$$\begin{aligned} h(\boldsymbol{z_{jk}}, \boldsymbol{\zeta}) = (0.0250 + b) \cdot (1 - e^{8.4053 IPOT_j}) \cdot (1 - e^{0.0127 fr_k}) \cdot e^{\frac{-(temp_k - 14.1700)^2}{2 \cdot 0.7486}} \\ + 0.9603 h(\boldsymbol{z_{j,k-1}}, \boldsymbol{\zeta}) \end{aligned} \tag{B2}$$

where $IPOT_j$ is the influence potential as described by [13], a competition index informing on shading conditions of location j; fr_k is the number of days since the last frost event; $temp_k$ is the daily temperature; and b is a year random effect set to zero in the present case study. The additive term is an auto-regressive factor.

In order to use $\Pr(E \mid v)$ as an estimator of $\mathrm{pmf}(u \mid u \geq 0)$, a relationship between v and u needs to be established. The period where seeds are available for germination is actually $g = v + u$ (i.e. time elapsed since germination conditions may be reasonably met up to seedling emergence — v — plus time of selection — u —). As a result, g is known and u can be assessed as $u = g - v$.