MCEM-Algorithm for estimating lifetime distribution in presence of interval censured observations and a censoring variable. Application to cocoa data.

Patrice Takam Soh
Corresponding author : patricetakam@gmail.com

March 9, 2021

Abstract

We propose here a parametric approach of lifetime estimation in a given stage, based on interval observations and on a censorship whose occurrence is independent of the interest event (changing stage) and which can stop the growth of individuals. Several nonparametric and parametric methods were proposed in the literature in order to take into consideration many types of censured data (right censured data, left censured data, interval censored data,...) which are usually found in medicine data, insurance data, etc. We are interested here in a particular censured data that is observed on the individuals on which the interest event and the censorship were recorded at the same time after occurrence in the interval in which events were observed. This kind of censorship has been observed on the cocoa data when we wanted to estimate the lifetime spent in a given developmental stage of growth. The data used for estimation were weekly observations on a cohort of cocoa fruits and at each observation day, we recorded for each fruit; its developmental stage and its sanitary status. Due to the weekly observations, there were such fruits that had been attacked (Pod rod disease,...) and changed their developmental stage during the same week so that both processes were recorded on the same fruits at the end of week. Here we propose a parametric model in order to take into consideration this type of individuals. We use an MCEM algorithm for estimating the involved parameter using the maximum likelihood.

Keywords: Modeling, lifetime distribution, parametric estimation, MCEM algorithm

1 Introduction

Here our goal is the parametric estimation of lifetime distribution based on the interval censured data. More precisely, we considered a population of individuals whose growth passes through many consecutive stages and we want to estimate the lifetime distribution in each stage. This estimation is complicated due to the interval observations, the fact that individuals were submitted to a censorship whose occurrence stopped their growth and the censorship due to the end of follow-up. The censorship will be referred to PR-censorship and the end of follow-up will be referred to FU-censorship. It has been shown in the literature (ref....) that the time spent in each given stage depends on some factors (henceforth called covariates) that therefore influence the transition from

that stage to the consecutive one.

The estimation of lifetime distribution in a given stage has to do in statistic with the field of survival analysis (Klein and Moeschberger (1997)) where: the interest event is the transition date to the next stage and censorships were: interval observations, presence of PR-censorship and FU-censorship which is a classical right censured data. These three types of censorships were keenly studied in the literature and a non exhaustive list of authors that worked on that topic was presented in (Takam et al. (2013)). In their work, Takam et al. (2013) proposed a non-parametric estimator of survival function with such good properties (continuous and differentiable with time) that took into consideration at the same time, the interval censured observations and the right censured observations (PR and FU censorships). In this work, we propose a parametric estimator that takes into account the previous censorships, in addition a particular censorship which was observed on the individuals on which both events (PR-censorship and interest event) occurred in the same interval and were recorded on the same individuals. This particular censorship exits because between two consecutive observations dates, individuals might change their stage (interest event) and contract PR-censorship. If it is the case, both events are recorded.

Converse to the non-parametric case, we have an additional information to estimate the lifetime in each stage, which is a vector of factors that were continuously recorded in the same period of study. These factors will henceforth be considered as time varying covariates that act and influence the growth of individuals. We then want to estimate a parametric model linking the growth of individuals described by the risk of changing stage, and the time varying covariate. This parametric model falls within the parametric estimation in survival analysis where one of the famous model is the semi-parametric model of Cox (1972).

Generally, in the parametric estimation of lifetime distribution, the objective is to estimate one characteristic of the lifetime distribution (survival function, density function or hazard function) as a function of observed covariates. In the literature, there are two ways to address the issue; either we consider that the distribution of lifetime belongs to a classical family of distribution (For example, Exponential power by Smith (1979), Exponential Family, Cubic exponential, Gamma Mixture with Common Scale Parameter and Gamma Mixture by Glaser (1980)) or we write the characteristic of lifetime (in most of time the hazard function) as a function of the observed covariates (For example, Cox (1972)). In the model proposed by Cox (1972), the covariates depend only on the individuals and not on time. This model was adapted for example by Young (2013) to take into consideration the time dependent covariate. More recently Dupuy (2014) generalized the idea of Cox in the case of missing time-dependent covariates. All these models assume that the hazard function belongs to a known parametric family. We propose here a new parametric family for the cumulative hazard function that is motivated from the data that inspired this work. In fact, we did not find in the literature a parametric distribution that linked the growth of cocoa fruits to the climatic variable

The main idea is to define the covariate that acts on an individual in a given stage at a given time as the cumulative values of time varying covariates from the day the individual entered the stage to that given time. This idea was motivated by Kemp et al. (1986) who considered that the growth of an organ depends on the cumulated energy. Our model finds its motivation from the cocoa data where we needed to propose a growth model for cocoa using factors that can influence their growth such as climatic variables (temperature, rainfall,...). For example, the role of the rain is

justified here by Lucia and Rua (2011) who proved that irrigation and nitrogen fertilization have an effect on the growth characteristics of cocoa (Theobroma cacao L.). As models describing biological processes linking developmental stages of cocoa fruits and climatic variables are not known in the literature, we can not use approaches proposed in literature by Génard (2007); Henton et al. (1999), Takashi (2000a); Takashi et al. (2000b); Takashi et al. (2002) and Takashi et al. (2008) who used a given biological decomposition. Dennis et al. (1986) and Kemp et al. (1986) proposed models that take into consideration the global growth from one stage to the next one. They proposed a model of growth through a latent variable that they consider as Brownian motion with drift directed by climate. In that model, the transition from one stage to the next was defined by a threshold of latent variable. But theses models do not take into consideration the fact that the growth depends on the accumulated energy. This hypothesis can be interpreted in the case of cocoa by the fact that the climate affecting the fruits does not variate from one fruit to another, which is not appropriate in our case. We also recall here that the growth is sometimes evaluated by the production. For example, Zuidema et al. (2005) showed through a physiological model of production that 70% of variability of cocoa production is explained by temperature and rainfall.

Our goal here is to build a model of the lifetime distribution that links the cumulative risk of changing stage as an empirical function of climatic variables (room temperature, temperature under cocoa plot and rainfall). The framework is presented in the second section, the MCEM algorithm used for estimation is presented in the third section, the application on the data is presented in the fourth section, the limits of the method is presented in the fifth section and the conclusions and discussions are presented in the sixth section.

2 Framework

2.1 Population

We first recall here that the framework used is the same presented by Takam et al. (2013) which can be summarized as follows. We consider a population of individuals that can passe through K developmental stages (respectively coded stages) during their growth in absence of an accident that can occur. The time T_k spent in a given stage k ($k \in \{1, ..., K\}$) can thus be viewed as a random variable with values in $\mathbb{R}^+$. We are in the case where the values taken by T_k depend on r factors denoted here by $\underline{x} = (x_1, ..., x_r) = (x_o)_{1 \leq o \leq r}$. We wish to estimate the distribution of each T_k as a given function of $\underline{x}$ and some parameters $\underline{\theta}_k$.

The estimation process of T_k was complicated by the fact that, quite often, an infectious agent coded $A1$ could attack individuals in Ω at any age, causing a disease coded PR[1]. Now, an individual with PR symptoms is regarded as a *failed item*, which can be interpreted as a *premature death*. Therefore, its lifetime in its current stage is *right-censored*, and the individual will never reach subsequent stages. In the following, an individual not yet attacked by PR will be said "healthy", and "PR-attacked" or "PR-censored" if not. At the end of study, individuals that were not yet attacked were considered as censured by the follow-up (FU-censured) and this is also another right censured data.

[1]This code comes from our motivating example about cocoa fruits.

2.2 Interpretation of the growth

Each individual i enters the first stage in $(-L, 0]$. In absence of any information about the distribution law of entering date in the first stage, we assume that the entrance is uniform in $(-L, 0]$. This hypothesis is formulated as follows:

Assumption $\mathcal{A}_1$. *The attack occurs uniformly in the interval in which it is recorded, that is*

$$f_{U_{i,1}|U_{i,1}\in(-L,0]} = \frac{1}{L}\mathbb{1}_{(-L,0]} \tag{1}$$

where $U_{i,1}$ is the exact day where i enter the stage 1 and $f_{U_{i,1}}(.)$ is its density distribution.

In absence of attack and when the last day of observation t_p is sufficiently large, i passes through all the stages $\{1, ..., K\}$. In presence of attack, the growth of i is stopped and we admit the following hypotheses.

Assumption $\mathcal{A}_2$. *Knowing the entering day in a given stage and the co-variables $\underline{x}$, the exit day from this stage is independent of the day of attack.*

Assumption $\mathcal{A}_3$. *Knowing the entering day in a given stage and the co-variables $\underline{x}$ and knowing that both events (attack and changing stage) occur in the same interval and are observed, we admit that the changing stage occurred before the attack.*

Now, knowing the vector of covariates $\underline{x}$, the part of each x_o $(1 \leq o \leq r)$ that influences the growth of i in the stage k is defined as follows.

Definition 1. *For a given individual that enters in the stage k at a given date u, the vector of covariates that act on that individual in the stage k until t $(t > u)$ is defined by $\widetilde{x}(t \mid u) = (\widetilde{x}_o(t \mid u))_{o=1,...,r}$ where*

$$\widetilde{x}_o(t \mid u) = \begin{cases} \int_u^t x_o(t)w_o(t)dt & \text{if } t \leq v \\ \int_u^v x_o(t)w_o(t)dt & \text{if } t > v \end{cases} \tag{2}$$

where $w_o(t) \equiv w_o$ is a constant weight that will be chosen such as $\widetilde{x}_o(t_p \mid -L) = 1$, v is the last day the individual was in that stage and t_p is the last day of observation in the study.

Using the definition 1, we can now give the hypothesis on which our model is based.

Assumption $\mathcal{A}_4$. *The cumulative hazard function in the stage k is defined for an individual as follows. Knowing that the individual entered the stage k at u and the covariates $\underline{x}$,*

$$H_k(v \mid u, \underline{x}) = \begin{cases} \sum_{o=1}^{r} \alpha_{k,o} \left[\widetilde{x}_o(v \mid u)\right]^{\mu_{k,o}} & \text{if } v > u \\ 0 & \text{if } v = u \\ \text{not defined} & \text{if } v < u. \end{cases} \tag{3}$$

The formula (3) means we considered in each $x_o(.)$, the part of history which influences the lifetime of the individual in the stage. This hypothesis is motivated by the growth model developed

by Kemp et al. (1986) that suggested that, the growth of an organ (represented here as the cumulative hazard) was usually due to the cumulative energy (represented here by cumulative covariates).

The hypothesis $(\mathcal{A}_3)$ allows to define a family model $\left\{F_\theta, \theta \in \Theta\right\}$ where

- $\theta = (\theta_k)_{1\leq k\leq K}$ with $\theta_k = \left((\alpha_{k,o})_{1\leq o\leq r}, (\mu_{k,o})_{1\leq o\leq r}\right)$ is a vector of $2r$ parameters;
- $\Theta = \underbrace{\mathbb{R}_+^* \times ... \times \mathbb{R}_+^*}_{2rK \text{ times}}$, since all the parameters were supposed non negatives;
- $dF_{\theta_k} = \varphi_{\theta_k} \cdot d\nu$ where ν is the Lebesgue measure and φ_{θ_k} is the the density deduced from $(\mathcal{A}_3)$, its expression is given in section 2.3.

By observing that the mapping $\begin{array}{ccc} \Theta & \longrightarrow & \mathbb{R}_+ \\ \theta & \longmapsto & H_k(v \mid u, \underline{x}) \end{array}$ for fixed values of $\underline{x}$ is an injective map, we then conclude that the proposed model is identifiable (Monfort, 1971).

2.3 Expression of $\varphi_k(.)$ in a current stage k

We first denote the random variable $T_k = V_k - U_k$, the lifetime of a healthy individual in stage k, with U_k which represents the exact entering day in the stage k and V_k the exact exit day from the stage k. As traditional in survival analysis, T_k was taken as a non-negative random variable and was assumed to be independently and identically distributed with probability density φ_{θ_k}.

The relationship between the survival function $\mathcal{S}_k(\cdot)$ and the cumulative hazard function $H_k(\cdot)$ allowed to write

$$\mathcal{S}_k(v \mid u, \underline{x}) = \Pr(V_k > v \mid U_k = u, \underline{x}) = \exp\left(-H_k(v \mid u, \underline{x})\right) \text{ for } v \geq u \geq 0. \tag{4}$$

The expression of the density φ_k is given by the following lemma.

Lemma 1. *Using the fact that the mapping $v \longmapsto H_k(v \mid u, \underline{x})$ is continuous and differentiable at $v \in \mathbb{R}_+$ knowing u and $\underline{x}$, φ_{θ_k} is given by*

$$\varphi_{\theta_k}(v \mid u, \underline{x}) = \mathcal{S}_k(v \mid u, \underline{x}) \cdot \sum_{o=1}^{r} \beta_{k,o}(v) \cdot [\widetilde{x}_o(v \mid u)]^{\mu_{k,o}-1} \tag{5}$$

where $\beta_{k,o}(v) = \alpha_{k,o} \cdot \mu_{k,o} \cdot w_o \cdot x_o(v)$ with $w_o = 1/\int_{-L}^{t_p} x_o(t)dt$.

Proof. We can write that, $\forall v > u$, we have

$$\begin{aligned} \frac{\partial H_k}{\partial v}(v \mid u, \underline{x}) &= \sum_{o=1}^{r} \alpha_{ko} \cdot \mu_{k,o} \cdot [\widetilde{x}_o(v \mid u)]^{\mu_{ko}-1} \cdot \frac{\partial \widetilde{x}_o(v \mid u)}{\partial v} \\ &= \sum_{o=1}^{r} \alpha_{k,o} \cdot \mu_{k,o} \cdot w_o \cdot [\widetilde{x}_o(v \mid u)]^{\mu_{ko}-1} \cdot x_o(v) \\ &= \sum_{o=1}^{r} \beta_{k,o}(v) \cdot [\widetilde{x}_o(v \mid u)]^{\mu_{k,o}-1}. \end{aligned}$$

Using the classical relation between $\mathcal{S}_k$ and the density φ_k, we have

$$\begin{aligned}\varphi_{\theta_k}(v \mid u, \underline{x}) &= -\frac{\partial \mathcal{S}_k}{\partial v}(v \mid u, \underline{x}) = \mathcal{S}_k(v \mid u, \underline{x}) \cdot \frac{\partial H_k}{\partial v}(v \mid u, \underline{x}) \\ &= \mathcal{S}_k(v \mid u, \underline{x}) \cdot \sum_{o=1}^{r} \beta_{k,o}(v) \cdot [\widetilde{x}_o(v \mid u)]^{\mu_{k,o}-1} \qquad (6)\end{aligned}$$

□

3 Parameter estimation

3.1 Observations

In order to estimate $\underline{\theta}_k$, we have a sample of N individuals that were recorded during p dates of observation from the origin $t_0 = 0$ to $t_p = pL$ with $p \geq 2$ ($p \in \mathbb{N}^*$) and $L > 0$. At each observation day $t_j = jL$ ($0 \leq j \leq p$), we recorded for each individual: its stage s_i^j and the presence of attack a_i^j ($a_i^j = 0$ if presence of attack and 1 if not) on i. The record information is the vector

$$w_i = \left((s_i^0, a_i^0), ..., (s_i^{p_i}, a_i^{p_i})\right)$$

where $s_i^j \in \{1, ..., K\}$, $a_i^j \in \{0, 1\}$ and p_i is the last day of observation of i ($p_i \in \{1, ..., p\}$) in the study.

In order to estimate the parameters in each stage, we will represent the observations of each individual in each stage k and this will be based on the following notations:

- $e_{i,k}$ the first day where i was observed in the stage k,
- $e_{i,k}^+$ the last day where i was observed in the stage k (where $e_{i,k}^+ = p_i L = t_{p_i}$ if i exit from the study in the stage k),
- $u_{i,k}$ the exact day where i entered in the stage k,
- $v_{i,k}$ the exit day of i from the stage k,
- $T_{i,k} = v_{i,k} - u_{i,k}$ the exact duration of i in the stage k,
- M_i the exact day where i was attacked in the study.
- $\underline{x}$ the time varying covariate; that is the factors that can influence the lifetime in each stage k. In our motivation example, these variables were represented by the temperature and rainfall that were collected in the study site and every day during the period of study.

We want to represent each item in the stage k by its entry interval I_k^{entry} and exit interval I_k^{exit} since this will facilitate the expression of the contribution of each individual to the likelihood of observations. In fact, the contribution of i to the partial likelihood of observations in the stage k is given by

$$\Pr(V_{i,k} \in I_{i,k}^{\text{exit}} \mid U_{i,k} \in I_k^{\text{entry}}, \underline{x}).$$

For a given individual i, that has been observed at least once in the stage k,

- Either i has been observed healthy until the last day of observation t_p; $u_{i,k} \in (e_{i,k} - L, e_{i,k}]$, $v_{i,k} > t_p$ and $T_{i,k} > t_p - u_{i,k}$.
- Either i has been observed healthy until the first day of observation in the stage $k+1$; $u_{i,k} \in (e_{i,k} - L, e_{i,k}]$, $v_{i,k} \in (e_{i,k+1} + L, e_{i,k+1}]$ and $T_{i,k} = v_{i,k} - u_{i,k}$.
- Either i has been observed attacked in the stage k; $u_{i,k} \in (e_{i,k} - L, e_{i,k}]$, $M_i \in (e^+_{i,k} - L, e^+_{i,k}]$, $v_{i,k} > M_i$ and $T_{i,k} > M_i - u_{i,k}$.
- Either i has been observed attacked during the transition interval from the stage k to $k+1$; $u_{i,k} \in (e_{i,k} - L, e_{i,k}]$, $M_i \in (e^+_{i,k} - L, e^+_{i,k}]$, $e_{i,k+1} - L < v_{i,k} < M_i$ and $e_{i,k+1} - L - u_{i,k} < T_{i,k} < M_i - u_{i,k}$.

These observations can be represented by $Y_{i,k} = \left(I^{\text{entry}}_{i,k}, I^{\text{exit}}_{i,k}, \Delta_{i,k}\right)$ with $I^{\text{entry}}_{i,k} = (e_{i,k} - L, e_{i,k}]$ and $I^{\text{exit}}_{i,k}$ is given as a function of $\Delta_{i,k}$ where

$$\Delta_{i,k} = \begin{cases} 1 & \text{if } I^{\text{exit}}_{i,k} = (t_p, +\infty) \\ 2 & \text{if } I^{\text{exit}}_{i,k} = (e_{i,k+1} - L, e_{i,k+1}] \\ 3 & \text{if } I^{\text{exit}}_{i,k} = (M_i, +\infty) \\ 4 & \text{if } I^{\text{exit}}_{i,k} = (e_{i,k+1} - L, M_i). \end{cases} \tag{7}$$

In the following, we will denote the exit interval from the stage k by $I^j_{i,k}$ for $j = 1, 2, 3, 4$ where

$$I^j_{i,k} = \begin{cases} (t_p, +\infty) & \text{for } j = 1 \\ (e_{i,k+1} - L, e_{i,k+1}] & \text{for } j = 2 \\ (M_i, +\infty) & \text{for } j = 3 \\ (e_{i,k+1} - L, M_i) & \text{for } j = 4 \end{cases} \tag{8}$$

3.2 Partial likelihood observations

We aim to evaluate the contribution of individuals to the partial likelihood of observations. If we denote by $\mathcal{J}_k$ the set of individuals that were observed at least once in the stage k, they can be organized in two groups as follows $\mathcal{J}_k = \mathcal{J}^h_k \cup \mathcal{J}^a_k$ where :

- $\mathcal{J}^h_k$ those on which the attack has not been observed in the exit interval from the stage k,
- $\mathcal{J}^a_k$ those on which the attack was observed in the exit interval from the stage k.

The contribution of i is then given by the following lemmas where we have separated the cases of $\mathcal{J}^h_k$ for healthy individuals and $\mathcal{J}^a_k$ for attacked individuals.

Lemma 2. *If we denote by $\mathcal{L}^h_{i,j}$ (with $j = 1, 2$), the contribution of i, knowing that i was healthy at exit from the stage k, to the likelihood of observations, we prove that it is given by*

$$\mathcal{L}^h_{i,k,j} = \begin{cases} \displaystyle\int_{e_{i,k}-L}^{e_{i,k}} \mathcal{S}_k(t_p \mid u, \underline{x}) f_{U_{i,k}|I^{entry}_{i,k}}(u \mid \underline{x})\ du & \text{for } j = 1 \\ \displaystyle\int_{e_{i,k}-L}^{e_{i,k}} \left[\mathcal{S}_k(e'_{i,k+1} \mid u, \underline{x}) - \mathcal{S}_k(e_{i,k+1} \mid u, \underline{x})\right] f_{U_{i,k}|I^{entry}_{i,k}}(u \mid \underline{x})\ du & \text{for } j = 2. \end{cases} \tag{9}$$

Proof. It is immediate, we write

$$\begin{aligned} \mathcal{L}^h_{i,k,j} &= \Pr\big(V_{i,k} \in (t_p, +\infty) \mid U_{i,k} \in (e'_{i,k}, e_{i,k}], \underline{x}\big) \\ &= \int_{e_{i,k}-L}^{e_{i,k}} \int_{t_p}^{+\infty} f_{V_{i,k}}(v \mid u, \underline{x}) f_{U_{i,k}|I^{\text{entry}}_{i,k}}(u \mid \underline{x})\ \text{d}u\text{d}v \end{aligned}$$

□

Lemma 3. *If we denote by $\mathcal{L}^a_{i,j}$ (with $j = 3, 4$), the contribution of i, knowing that i was attacked at exit from the stage k, to the likelihood of observations, we prove that it is given by*

$$\mathcal{L}^a_{i,k,j} = \begin{cases} \frac{1}{L}\displaystyle\int_{e_{i,k}-L}^{e_{i,k}} \int_{e^+_{i,k}-L}^{e^+_{i,k}} \mathcal{S}_k(m \mid u, \underline{x}) f_{U_{i,k}|I^{entry}_{i,k}}(u \mid \underline{x})\ du dm & \text{for } j = 1 \\ \frac{1}{L}\displaystyle\int_{e_{i,k}-L}^{e_{i,k}} \int_{e^+_{i,k}-L}^{e^+_{i,k}} \left[\mathcal{S}_k(e'_{i,k+1} \mid u, \underline{x}) - \mathcal{S}_k(m \mid u, \underline{x})\right] f_{U_{i,k}|I^{entry}_{i,k}}(u \mid \underline{x})\ du & \text{for } j = 2. \end{cases} \tag{10}$$

Proof. It is straightforward by witting that

$$\begin{aligned} \mathcal{L}^a_{i,k,j} &= \Pr\big(V_{i,k} \in (e'_{i,k+1}, e_{i,k+1}] \mid U_{i,k} \in (e'_{i,k}, e_{i,k}], M_i \in (e^+_{i,k} - L, e^+_{i,k}], \underline{x}\big) \\ &= \frac{1}{L}\int_{e_{i,k}-L}^{e_{i,k}} \int_{e^+_{i,k}-L}^{+\infty} \Pr\big(V_{i,k} \in (e'_{i,k+1}, e_{i,k+1}] \mid U_{i,k} = u, M_i = m, \underline{x}\big) f_{U_{i,k}|I^{\text{entry}}_{i,k}}(u \mid \underline{x})\ \text{d}u\text{d}m \end{aligned}$$

since M_i is uniform in $(e^+_{i,k} - L, e^+_{i,k}]$. It comes that

$$\mathcal{L}^a_{i,k,j} = \frac{1}{L}\int_{e_{i,k}-L}^{e_{i,k}} \int_{e^+_{i,k}-L}^{e^+_{i,k}} \Pr\big(V_{i,k} \in (e'_{i,k+1}, e_{i,k+1}] \mid U_{i,k} = u, \underline{x}\big) f_{U_{i,k}|I^{\text{entry}}_{i,k}}(u \mid \underline{x})\ \text{d}u\text{d}m$$

since $V_{i,k}$ and M_i are independents knowing $U_{i,k}$ and $\underline{x}$. □

Proposition 1. *Under the hypothesis that individuals are independent, the partial likelihood of observations is given by*

$$\mathcal{L}(\underline{y} \mid \underline{\theta}_k, \underline{x}) = \prod_{i \in \mathcal{J}^h_k} \left\{ \prod_{j=1}^{2} \{\mathcal{L}^h_{i,k,j}\}^{\mathbb{1}_{\Delta_{i,k}=j}} \right\} \prod_{i \in \mathcal{J}^a_k} \left\{ \prod_{j=3}^{4} \{\mathcal{L}^h_{i,k,j}\}^{\mathbb{1}_{\Delta_{i,k}=j}} \right\} \tag{11}$$

Proof. We easily obtain the partial likelihood of observations by calculating separately the case where $i \in \mathcal{J}_k^h$ and the case where $i \in \mathcal{J}_k^a$. In the expression of (11), $\mathcal{L}_{i,k,j}^h$ and $\mathcal{L}_{i,k,j}^a$ are respectively given by lemma 2 and lemma 3. □

The main difficulty to estimate this likelihood resides to the fact that it contains many integrals which have to be approximated before the estimation. In fact as many integrals to estimate as they are individuals in each stage.

Remark 1. *We can notice that if the density $f_{U_{i,k}}(.)$ can be integrated explicitly, the estimators of $\underline{\theta}_k$ can be obtained at this step by maximizing this partial likelihood. In our case, due to the fact the density of $V_{i,k}$ knowing $U_{i,k}$ is not easy to compute, we will use a MCEM-algorithm for estimation.*

3.3 MCEM Algorithm approach

We are interested in the estimation of $\underline{\theta}_k$ that corresponds to the parameter of the lifetime distribution in the stage k.

3.3.1 The complete data likelihood

Each individual i entered in the stage k at u_i (with $u_{i,k} \in I_{i,k}^{\text{entry}}$) and exit from that stage at $v_{i,k}$ (with $v_{i,k} \in I_{i,k}^{\text{exit}}$). Using the hypothesis of independence between individuals, the complete data likelihood is given by

$$L^c\big(\theta_k \mid u_k, v_k, \underline{x}\big) = \prod_{i=1}^{n_k} \varphi_k(v_{i,k} \mid u_{i,k}, \underline{x}) \tag{12}$$

where $\underline{u}_k = (u_{1,k}, ..., u_{n_k,k})$, $\underline{v}_k = (v_{1,k}, ..., v_{n_k,k})$, $\underline{\theta}_k = (\alpha_{k,o}, \mu_{k,o})_{1 \leq o \leq r}$ and $\varphi_k(.)$ is given by the formula (5).

3.3.2 Monte Carlo EM

If we denote by $\underline{\theta}_k^{(r)}$ the parameter of the r^{th} iteration of the algorithm in the current stage k, the expectation of the complete data likelihood knowing observation is given by the proposition 2. In this proposition, k has been omitted on u_i, v_i, y_i, Δ_i and $\underline{\theta}^{(r)}$.

Proposition 2. *If we denote by*

- $p\left(u_i \mid y, \underline{\theta}^{(r)}\right)$ *the probability distribution function of u_i, knowing the data (y) and the parameters $\underline{\theta}^{(r)}$*
- $p_j\left(v_i \mid u_i, y, \underline{\theta}^{(r)}\right)$ *(for $j = 1, 2, 3, 4$) the probability distribution function of v_i knowing u_i, the data (y) and the parameters $\underline{\theta}^{(r)}$, the conditional expectation of the complete data likelihood is given by*

$$Q(\underline{\theta} \mid \underline{\theta}^{(r)}) = \sum_{i=1}^{n_k} \int_{I_i^j} Q^i(\underline{\theta}, \underline{\theta}^{(r)}) p(u_i \mid y_i, \underline{\theta}^{(r)}) du_i \tag{13}$$

where

$$Q^i(\underline{\theta}, \underline{\theta}^{(r)}) = \sum_{j=1}^{4} 1_{\{\Delta_i = j\}} \int_{I_i^j} \log\left(L^c\big(\underline{\theta} \mid u_i, v_i, y_i\big)\right) p_j\big(v_i \mid u_i, \underline{\theta}^{(r)}, y_i\big) dv_i \tag{14}$$

with

$$p_1\big(v \mid u, \underline{\theta}^{(m)}, y\big) = \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{\mathcal{S}_k(t_p \mid u)} \tag{15}$$

$$p_2\big(v \mid u, \underline{\theta}^{(m)}, y\big) = \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{\mathcal{S}_k(e'_{i,k+1} \mid u, \underline{x}) - \mathcal{S}_k(e_{i,k+1} \mid u, y)}. \tag{16}$$

$$p_3\big(v \mid u, \underline{\theta}^{(r)}, y\big) = \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{L} \cdot \int_{e^+_{i,k}-L}^{e^+_{i,k}} \frac{1}{\mathcal{S}_k(z \mid u, y)} dz \tag{17}$$

$$p_4\big(v \mid u, \underline{\theta}^{(r)}, y\big) = \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{L} \cdot \int_{e^+_{i,k}-L}^{e^+_{i,k}} \frac{1}{\mathcal{S}_k(e_{i,k+1} - L \mid u, y) - \mathcal{S}_k(z \mid u, y)} dz \tag{18}$$

Proof. We have

$$\begin{aligned}
Q(\underline{\theta} \mid \underline{\theta}^{(r)}) &= \sum_{i=1}^{n_k} \mathbb{E}_{(u_i,v_i)|y_i,\underline{\theta}^{(r)}} \log\big(L^c\big(\underline{\theta} \mid u_i, v_i, \underline{x}\big)\big) \\
&= \sum_{i=1}^{n_k} \mathbb{E}_{u_i|y_i,\underline{\theta}^{(r)}} \Big(\mathbb{E}_{v_i|y_i,u_i,\underline{\theta}^{(r)}} \log\big(L^c\big(\theta \mid u_i, v_i, y\big)\big)\Big) \\
&= \sum_{i=1}^{n_k} \int_{I^{\text{entry}}_{i,k}} \Big(\mathbb{E}_{v_i|y_i,u_i,\underline{\theta}^{(r)}} \log\big(L^c\big(\underline{\theta} \mid u_i, v_i, y\big)\big)\, p(u_i \mid y_i, \underline{\theta}^{(r)})\Big)\, \mathrm{d}u_i \\
&= \sum_{i=1}^{n_k} \int_{I^{\text{entry}}_{i,k}} Q^i(\underline{\theta}, \underline{\theta}^{(r)}) p(u_i \mid y_i, \underline{\theta}^{(r)}) \mathrm{d}u_i
\end{aligned}$$

where

$$Q^i(\underline{\theta}, \underline{\theta}^{(r)}) = \sum_{j=0}^{3} \mathbb{1}_{\{\Delta_i = j\}} \int_{I_i^j} \log\big(L^c\big(\theta \mid u_i, v_i, y\big)\big)\, p_j\big(v_i \mid u_i, \underline{\theta}^{(m)}, y_i\big) dv_i$$

with

$$p_1\big(v \mid u, \underline{\theta}^{(m)}, y\big) = \frac{\varphi_k(v \mid u, \underline{x})}{\displaystyle\int_{t_p}^{+\infty} \varphi_k(v \mid u, \underline{x})\ \mathrm{d}u} = \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{\mathcal{S}_k(t_p \mid u)},$$

$$p_2\big(v \mid u, \underline{\theta}^{(m)}, y\big) = \frac{\varphi_k(v \mid u, \underline{x})}{\displaystyle\int_{e'_{i,k+1}}^{e_{i,k+1}} \varphi_k(v \mid u, \underline{x})\ \mathrm{d}u} = \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{\mathcal{S}_k(e'_{i,k+1} \mid u, \underline{x}) - \mathcal{S}_k(e_{i,k+1} \mid u, y)},$$

$$\begin{aligned}
p_3\big(v \mid u, \underline{\theta}^{(r)}, y\big) &= \frac{1}{L} \int_{e^+_{i,k}-L}^{e^+_{i,k}} p_3\big(v \mid u, \underline{\theta}^{(r)}, y, z\big) \mathrm{d}z \text{ because of } \mathcal{A}_2 \\
&= \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{L} \cdot \int_{e^+_{i,k}-L}^{e^+_{i,k}} \frac{1}{\mathcal{S}_k(z \mid u, y)} dz,
\end{aligned}$$

$$
\begin{aligned}
p_4\big(v \mid u, \underline{\theta}^{(r)}, y\big) &= \frac{1}{L}\int_{e^+_{i,k}-L}^{e^+_{i,k}} p_3\big(v \mid u, \underline{\theta}^{(r)}, y, z\big)\mathrm{d}z \text{ because of } \mathcal{A}_2 \\
&= \frac{\varphi_{\underline{\theta}}(v \mid u, \underline{x})}{L} \cdot \int_{e^+_{i,k}-L}^{e^+_{i,k}} \frac{1}{\mathcal{S}_k(e_{i,k+1} - L \mid u, y) - \mathcal{S}_k(z \mid u, y)} dz \text{ because of } \mathcal{A}_3.
\end{aligned}
$$

□

The expression of $Q(\underline{\theta}, \underline{\theta}^{(m)})$ given by the proposition (2) can be considered as a general form of the expectation in the **E**-step of **EM**-Algorithm for the parametric estimation of lifetime distribution in presence of: (i) double interval censured observation and (ii) in presence of a censoring variable that can occur independently and mask the observation of the transition to the next stage. We can observe that for the cases where all individuals were observed from the time $t = 0$ and whose there is not a censoring variable, $Q(\underline{\theta}, \underline{\theta}^{(r)})$ is reduced to the ones proposed by Turnbull (1976). We can also notice that for the case where the conditional density of v knowing u is a standard one (egg. exponential distribution,...) $Q(\underline{\theta}, \underline{\theta}^{(r)})$ could be obtained explicitly and the **E**-step could be easily handled. In our case due to the nonstandard form of density, it is not easy to get an explicit form of that expectation. We then used the Monte Carlo simulations (Robert and Casella (2004)) in order to estimate the quantity $Q(\underline{\theta}, \underline{\theta}^{(r)})$. In fact, the simulated values of (U_k, V_k) are obtained as follows. In each stage k, we first simulate the values of $U_k = (u_{1,k}, ..., u_{n_k,k})$ for the n_k individuals. For each value of $u_{i,k}$, we simulate $N.sim$ values of $\{v^{(r)}_{i,k,}\}_{1\leq t\leq N.sim}$ knowing the parameter $\theta^{(r)}$ and the observation y_k. We then obtain the sample

$$
\Big\{(u_{1,k}, v^{(r)}_{1,k,t}), ..., (u_{n_k,k}, v^{(r)}_{n_k,k,t})\Big\}_{1\leq t\leq N.sim}. \tag{19}
$$

The quantity $Q(\underline{\theta}, \underline{\theta}^{(r)})$ is estimated by the Monte Carlo sum

$$
\begin{aligned}
\widehat{Q}_{N.sim}(\theta, \theta^{(r)}) &= \frac{1}{N.sim}\sum_{t=1}^{N.sim} \log\big(L^c(\theta \mid \underline{u}^{(m)}, \underline{v}^{(r)}_t, \underline{x})\big) \\
&= \frac{1}{N.sim}\sum_{t=1}^{N.sim}\sum_{i=1}^{n_k} \log\big(\varphi_\theta(v^{(r)}_{i,k,t} \mid u^{(r)}_{i,k}, \underline{x})\big), \qquad (20)
\end{aligned}
$$

where the subscript r denotes the dependance of this estimator on the MC sample size. We recall here that by the law of large number, the estimator in (20) converges to the theoretical expectation $Q(\theta, \theta^{(r)})$. The standard E-step of the EM-Algorithm is thus modified here into an MCEM whereby the E-step is replaced by the estimated quantity from (20).

3.3.3 MC sample

To obtain the sample $\Big(u^{(r)}_i, v^{(r)}_{i,t}\Big)$ for a fixed value of i and r, we used the Metropolis-Hastings algorithm (Robert and Casella (2004)). More precisely, we used this algorithm in order to obtain the value of $v^{(m)}_{i,t}$ knowing the value of $u^{(m)}_i$ and of $\underline{\theta}^{(r)}$. In fact, in the first stage ($k = 1$), $u^{(r)}_i$ is

obtained uniformly in the interval $(-L, 0]$ and in the stage k $(k > 1)$, $u_i^{(r)}$ is obtained exactly as $v_{i,t}^{(m)}$. In fact, it can either be considered as the entering day in the stage k or as the exit day from the stage $k-1$. The algorithm is proposed as follows.

- step 1. Specify the initial value of the parameter $\underline{\theta}_k^{(0)}$.
- step 2. Select starting point $v^{(t)}$ satisfying $\varphi_k(v^{(t)} \mid u, \underline{x}) > 0$.
- step 3. Generate ν_t from the proposal distribution, which is here the uniform distribution in $I_{i,k}^j$.
- step 4. Take $v^{(t+1)} = \nu_t$ with probability $\rho(v^{(t)}, \nu_t)$ or otherwise take $v^{(t+1)} = v^{(t)}$ with probability $1 - \rho(v^{(t)}, \nu_t)$, where

$$\rho(z, z') = \min\left(\frac{\varphi_k(z \mid u, \underline{\theta}^{(0)}, y)}{\varphi_k(z' \mid u, \underline{\theta}^{(0)}, y)}, 1\right).$$

Remark 2. *Due to the presence of integral on the expression of p_j $(j = 3, 4)$, drawing the MCMC sample each iteration of the EM algorithm could be prohibitively costly particulary for large $N.sim$. As it is discussed in Robert and Casella (2004), the computational expense of the MCMC based MCEM algorithm can be substantially improved through an application of importance sampling.*

3.3.4 Importance sampling

In this procedure, at each iteration r, rather than obtaining a new sample with the most recent iterate $\theta^{(r)}$, we importance weight in the original sample through the update distribution of v. That is

$$\widehat{Q}_{N.sim}(\underline{\theta}, \underline{\theta}^{(r)}) = \sum_{t=1}^{N.sim} \sum_{i=1}^{n_k} w_t \log\left(\varphi_k(v_{i,t} \mid u_i, y)\right) \Big/ \sum_{t=1}^{N.sim} w_t, \tag{21}$$

where the sample $(u_i, v_{i,t})$ was obtained using the initial point $\underline{\theta}_k{}^{(0)}$ and this sample is corrected for the new information we have at iteration r through the weights

$$w_t' = w_t \Big/ \sum_{t=1}^{N.sim} w_t \text{ with } w_t = \frac{p(v_t \mid u, \underline{\theta}^{(r)}, y)}{p(v_t \mid u, \underline{\theta}^{(0)}, y)}. \tag{22}$$

The importance sampling is discussed in [See Robert an Casella (1999,chap. 3)]. In practice, the objective is to write w_t as a function of the density distribution $\varphi_k(.)$. The following lemma gives the expression of w_t as a function of φ_k.

Lemma 4. *The weight w_t is proportional to the ratio of density, that is*

$$w_t = C \cdot \frac{\varphi_k(v_t \mid u, \underline{\theta}^{(r)}, y)}{\varphi_k(v_t \mid u, \underline{\theta}^{(0)}, y)}. \tag{23}$$

where C is a constant which does not depend on t.

Proof. In fact, we have

$$p(v_t \mid u, \theta, y) = \frac{p(v_t, u, \theta \mid y)}{p(u, \theta \mid y)} = \frac{p(\theta \mid v_t, u, y) \cdot p(v_t \mid u, y)}{p(\theta \mid u, y)} \tag{24}$$

and thus

$$w_t = \frac{p(v_t \mid u, \underline{\theta}^{(r)}, y)}{p(v_t \mid u, \underline{\theta}^{(0)}, y)} = \frac{p(\underline{\theta}^{(r)} \mid v_t, u, y)}{p(\underline{\theta}^{(0)} \mid v_t, u, y)} \cdot \frac{p(\underline{\theta}^{(r)} \mid u, y)}{p(\underline{\theta}^{(0)} \mid u, y)} = C \cdot \frac{\varphi_k(v_t \mid u, \underline{\theta}^{(r)}, y)}{\varphi_k(v_t \mid u, \underline{\theta}^{(0)}, y)} \tag{25}$$

since

$$p(\underline{\theta} \mid v_t, u, y) = \varphi_k(v_t \mid u, \underline{\theta}, y).$$

□

The preceding lemma allows to take as the weight, the ratio $\frac{\varphi_k(v_t|u,\underline{\theta}^{(r)},y)}{\varphi_k(v_t|u,\underline{\theta}^{(0)},y)}$

3.3.5 The Algorithm

1. Initialize $N.sim, \underline{\theta}^{(0)}$.
2. Generate (u, v_t) via the MCMC algorithm described in section 3.3.3.
   At iteration $r + 1$:
3. Compute the importance weights

$$w_t = \frac{\varphi_k(v_t \mid u, \underline{\theta}^{(m)}, y)}{k}(v_t \mid u, \varphi_{\underline{\theta}^{(0)}, y)}. \tag{26}$$

4. **E**-step: Estimate $\widehat{Q}_m(\underline{\theta}, \underline{\theta}^{(m)})$ by

$$\widehat{Q}_{N.sim}(\underline{\theta}, \underline{\theta}^{(m)}) = \frac{1}{N.sim} \sum_{t=1}^{N.sim} \sum_{i=1}^{n_k} w_t \log\left(\varphi_{\underline{\theta}}(v_{i,t} \mid u_{i,t}, \underline{x})\right). \tag{27}$$

5. **M**-step: Maximize $\widehat{Q}_{N.sim}(\underline{\theta}, \underline{\theta}^{(r)})$ under constrains $\theta > 0$ to obtain $\underline{\theta}^{(r+1)}$.
6. Repeat Step 2 through 5 until convergence.

3.4 Estimation strategy

The algorithm we proposed in the previous section will be used to estimate the parameters $\underline{\theta}_k$ and we will start by the first stage. The only tricky point to highlight here is the distribution of U_i that corresponds to the entry day in the current stage. In the first stage, U_i will be obtained using the uniform distribution and in the other stages ($k > 1$), U_i will be obtained using the lifetime distribution estimated in the stage $k - 1$.

The main difficulty in the likelihood estimations resides in the presence of integrals in the expressions of $\tilde{x}_o$ and in the different probabilities p_j (for $j = 1, 2, 3, 4$). We will use the the trapezium formula to approximate these integrals (See Appendices **A**).

To analyze the variability of our estimations, we will use the nonparametric bootstrap which can be described as follows.We resampling the data and for each re-sampled data, we used the MCEM-algorithm described in the previous sections to obtain the parameters $\underline{\theta}_k^\ell$ at the ℓ^{th} resampling. The procedure will execute B times (for $B = 1000$ for example) in order to deduce the estimated bias and variance of $\widehat{\theta}_k$.

3.4.1 Initialization of the parameters θ_k

To initialize the parameters θ_k, we will use the OLS method by minimizing the distance between $\widehat{H}_k(\tau_j)$ and $\sum_{o=1}^{r} \alpha_{k,o}[\bar{x}_o(t)]^{\mu_{k,o}}$ which ic defined by

$$\underset{\theta_k \in \Theta_k}{\operatorname{argmin}} \sum_{j=1}^{q_k} \left(\widehat{H}_k(\tau_j) - \sum_{o=1}^{r} \alpha_{k,o}[\bar{x}_o(\tau_j)]^{\mu_{k,o}} \right)^2. \tag{28}$$

under constraint $\underline{\theta}_k > 0$ where $\bar{x}_o(\tau_j)$ represents the cumulative climate submitted to $i \in J_k(\tau_j)$ ($J_k(\tau_j)$ is the number of individuals that had been observed at least once in the stage k and spent at least τ_j days). It is given by

$$\overline{x}_{i,o}(\tau_j) = \sum_{t=e_{i,k}}^{\tau_j} x_o(t). \tag{29}$$

To obtain $H_k(\tau_j)$, we proceed as follows:

- We estimate the discrete lifetime by the Kaplan Meier estimator, that is

$$\widehat{\mathcal{S}}_k^{\text{KM}}(\tau_j) = \prod_{\ell=1}^{j} \left(1 - \frac{d_{\ell,k}}{N_{\ell,k}} \right) \text{ avec } j = 1, ..., q_k \tag{30}$$

  where

  - $d_{\ell,k}$ is the number of individual whose the observed discrete duration Z_k^{obs} equals to τ_ℓ. This duration is defined by $Z^{\text{obs}} = \min(Z_k^d, Z_k^c)$ where $Z_{i,k}^d = e_{i,k+1} - e_{i,k}$ and $Z_{i,k}^c = e_{i,k}^+ - e_{i,k}$.
  - $N_{\ell,k}$ is the number of individual whose the discrete duration is greater or equal to τ_ℓ.

- $\widehat{H}_k(\tau_j)$ is obtained by the relationship between the survival function and the cumulative risk, that is

$$\widehat{H}_k(\tau_j) = -\ln\left(\widehat{\mathcal{S}}_k^{\text{KM}}(\tau_j)\right) \tag{31}$$

The minimization will be done by using the command *ConsOptim* available on R software. We recall here that the principle has been motivated by the main idea that helped us to find empirically the family model (See Appendices **C** for details.).

3.5 Statistics Inference of parameters

After estimating the parameters $\underline{\theta}$, we used the likelihood ratio test to test the hypothesis $H_0 : \alpha_{k,j} = \alpha_{k,j'}$ for $j \neq j'$ against the alternative one, that is $H_1 : \alpha_{k,j} \neq \alpha_{k,j'}$. In fact, this test allowed to compare the effects of corresponding time varying covariates x_j and $x_{j'}$ for $(1 \leq j \neq j' \leq r)$. If we denote by $\widetilde{\theta}_k$ (resp. $\widehat{\theta}_k$) the estimators of θ_k obtained from the null hypothesis (resp. alternative hypothesis), the statistic of test is given by

$$L_n = 2\big(\mathcal{L}_k(\widehat{\theta}_k) - \mathcal{L}_k(\widetilde{\theta}_k)\big) \tag{32}$$

which is a classical likelihood ratio test (Dacunha et al., 1992) and this statistic is approximated by a chi-square distribution with $(2r-1)$ as degree of freedom. We recall here that $\mathcal{L}_k$ is the logarithm of the likelihood of observations which is obtained at a given $\underline{\theta}$ by approximating the integrals (simple and double ones) present in the partial likelihood given by the proposition 2.

In practice, the null hypothesis is rejected if the statistic L_n is more large, that is $L_n > \chi^2_{2r-1,1-\alpha}$.

3.6 Potential Methodological Limitations

The method proposed here is based on many parameters. In fact the number of parameters in the model is $2 \times r \times K$ with depend on the number of time varying covariate and the number of stages. The larger the number of variables the more numerous the parameters. This can then complicate the estimation. In this work the simulation approach has been chosen because in our motivation example we had six parameters per developmental stage.

The parametric model proposed here is not easy to manage in inferential statistic. In fact, if we needed to test the effect of one time varying covriate using our model, for example if we wanted to test the null hypothesis $\alpha_{k,\ell} = 0$, the likelihood would have been degenerated because the space parameters should not be an open set and the standard ratio test could not be applied. This problem is being done by the authors of this work. But here the ratio test has been used only to compare the effect of different covariates.

The strategy of estimation proposed here in each stage has at least two levels of approximation. The first level comes from the fact we used estimation provided by the previous stage and the second one provided in the likelihood estimation.

Appendices

Appendices A: Approximation of $\widetilde{x}_o$

Computation of $\widetilde{x}_o$

We first recall the trapezoid formula that will be used in order to approximate the integral, that is

$$\overline{x}_o(b \mid a) = \int_a^b x_o(t)dt \simeq \frac{b-a}{n}\left(\sum_{j=1}^{n-1} x_o(t_j) + \frac{1}{2}(x_o(b) + x_o(a))\right)$$

The function $\widetilde{x}(v \mid u)$ is given by

$$\widetilde{x}(v \mid u) = \frac{\overline{x}(v \mid u)}{\overline{x}(t_{\max} \mid 1)}.$$

Remark 3. *We need to evaluate $\overline{x}(v \mid u)$ where u and v are continuous and do not fall directly on the observation days. We are then going to consider the cumulative values of climate as the cumulatives values that fall between u and v. More precisely, we will have $t_0 = [u] + 1$ and $t_n = [v]$ and for $1 \leq j \leq n-1$, the t_j represent the observed values between the days $[u] + 2$ and $[v] - 1$. We then add the condition that, if $[v] - [u] - 1 \leq 1$ the cumulative goes to 0 else it is given by the following integral*

$$\overline{x}(v \mid u) = \int_a^b x_o(t)dt \simeq \sum_{j=1}^{n-1} x_o(t_j) + \frac{1}{2}\big(x_o([u]+1) + x_o([v])\big)$$

with $n = [v] - [u] - 1$ et $h = 1$ and since, $\frac{(b-a)}{n} = 1$.

Appendices B: Approximation of double integrals

To compute the double integral, we used an approximation, that derived in the approximation of an integral in the triangle **??**.

$$\begin{aligned}\frac{1}{A}\int_\Omega F(\xi_1,\xi_2,\xi_3)\,\mathrm{d}u\,\mathrm{d}v \quad &\simeq \quad w^1\left[F(g_1,g_1,1-2g_1) + F(g_1,1-2g_1,g_1)\right.\\ &\qquad \left. + F(1-2g_1,g_1,g_1) + F(g_2,g_2,1-2g_2) +\right]\\ &+ \quad w^2\left[F(g_2,1-2g_2,g_2) + F(1-2g_2,g_2,g_2)\right]\end{aligned}$$

where $w^1 = 0.22338$; $w^2 = 0.10995$; $g_1 = 0.4459$; $g_2 = 0.0916$. We then used this approximation to approximate the function on a square as follows: If we denote the vertex of a given square by $A(a,c)$, $B(b,c)$, $C(b,d)$ and $D(a,d)$, we have the following approximation

$$\int_\Omega F(\xi_1,\xi_2,\xi_3)\,\mathrm{d}u\,\mathrm{d}v \quad \simeq \quad \int_{\Omega_1} F(\xi_1,\xi_2,\xi_3)\,\mathrm{d}u\,\mathrm{d}v + \int_{\Omega_2} F(\xi_1,\xi_2,\xi_3)\,\mathrm{d}u\,\mathrm{d}v$$

where $\Omega_1 = (ABD)$ and $\Omega_2 = (SDB)$ are triangles and $(ABCD)$ is a square.

Appendices C: Empirical model

We need to compare in the same figure, the evolution of the cumulative risk of changing stage and the evolution of the cumulated climatic variables.

The non-parametric estimation method used here for the cumulative risk function $\widehat{H}_k(\tau_j)$ is the one estimated in the section 3.4.1.

We then need to define climate in function of the time spent in the given stage. Fruits enter in each stage k at the different days and each of them is submitted to a cumulative climate during the time spent in the stage. To construct an estimate of this climate as a function of the age spent in the stage, we proceed as follow. In a given stage and for a given discrete age τ_j with $j = 1, ..., q_k$,

we consider the set of fruits that stays at least τ_j (denoted by $J_k(\tau_j)$). For each $i \in J_k(\tau_j)$, we consider the cumulative climate submitted to i defined by

$$\overline{x}_{i,o}(\tau_j) = \sum_{t=e_{i,k}}^{\tau_j} x_o(t). \tag{33}$$

The mean of cumulative climates by all the fruits which have been observed at least once in the stage k is defined by

$$\overline{x}_o(\tau_j) = \frac{1}{n_k(\tau_j)} \sum_{i \in \mathcal{J}_k} \overline{x}_{i,o}(\tau_j) = \frac{1}{n_k(\tau_j)} \sum_{i \in \mathcal{J}_k} \left(\sum_{t=e_{i,k}}^{\tau_j} x_o(t) \right) \tag{34}$$

where $n_k(\tau_j) = \text{card}\mathcal{J}_k(\tau_j)$ represents the total number of fruits that have been observed at least once in the stage k and such that their discrete age in the stage k was greater than τ_j.

References

Austin, P. T., Hall, A. J., Snelgar, W. P. and Currie, M., (2002). Modelling kiwifruit budbreak as a function of temperature and budbreak interaction. *Animals of Botaby*, 89, 695-706.

Cox, D. R.(1972). Regression models and life tables (with discussion). *Journal of Royal Statistic Society*, 34, 187-220.

Dennis, B., Kemp, W. P. and Beckwith, R. C., (1986). Stochastic model of insect phenology. *Environ. Entomol*, 15, 540-546.

Dupuy, J. F. and Mesbah., M. (2014). Estimation of the Asymptotic Variance of Semiparametric Maximum Likelihood Estimators in the Cox Model with a Missing Time-Dependent Covariate. *Communications in Statistics Theory and Methods*. 21, 263-275.

Garófalo Chaves, L., H. and Joaquim Pequeno, R., A., (2011). *IJEE an Official Peer Reviewed Journal of Babol Noshirvani University of Technology*, 496, ...

Génard, M., Bertin, N., Borel, C., Bussières, P., Gautier, H., Habib, R., Lechaudel, M., Lecomte, A., Lescourret, F., Lobit, P. and Lescourret, B. Q., (2007). Bathtuband Related Failure Rate Characterizations, *IJEE an Official Peer Reviewed Journal of Babol Noshirvani University of Technology*, 58, 917-928.

Glaser and Watson, (1980). Bathtuband Related Failure Rate Characterizations, *Journal of the American Statistical Association*, 75, 667-671.

Henton, S. M., Piller, G., J. and Gandar, P. W., (1999). A fruit growth model dependent on both carbon supply and inherent fruit characteristics. *Annals of Botany*, 83, 509-514.

Kemp, W. P., Dennis, B., and Beckwith, R. C., (1986). Stochastic phenology model for the westerm spruce budworm (Lepidoptera : Tortricidae). *Environ. Entomol*, 15, 547-554.

Klein, J. P. and Moeschberger M. L., (1997). Survival analysis: techniques for censored and truncated data. *Springer-Verlag*, New York, USA.

Monfort, A., (1971). Cours de statistique mathématique. Paris, France: Economica.

Ndoumbe-Nkeng, M., (2002). Incidence des facteurs agro-?cologiques sur l'?pid?miologie de la pourriture brune des fruits du cacaoyer au Cameroun : contribution ? la mise en place d'un mod?le d'avertissement agricole. Paris, France: INA-PG, PhD thesis.

Robert, C., Casella, G.(2004). Monte Carlo Statistical Methods, Springer.

Smith, R.M. and Bain, L.J., (2005). An Exponential Power Life-Testing Distribution, *Communications in Statistics*, 4, 469-481.

Takam Soh, P., Ndong Nguéma, E. P., Gwet H. and Ndoumbè-Nkeng, M., (2013). Smooth estimation of a lifetime distribution with competing risks by using regular interval observations: application to cocoa fruits growth. Journal of Royal Statistics Society.

Takashi, A. Hung Nama, L., Hietz, P. Tanaka, N., Karunaratne, S.; (2002). Seasonal fluctuations in live and dead biomass of phragmites australis as described by a growth and decomposition model: implications of duration of aerobic conditions for litter mineralization and sedimentation. *Aquatic Botany*, 73, 223-239.

Takashi, A., Rajapakse, L., Takeshi, F., (2008). Applications of organ-specific growth models; modelling of resource translocation and the role of emergent aquatic plants in element cycles. *ecological modelling*, 215, 170-179.

Takashi, A. Shiromi, K., (2000). Dynamic modeling of the growth of phragmites australis: model description. *Aquatic Botany*, 63, 301-318.

Takashi, A., Vu Kien, T., Manatunge, J., (2000). Modeling the effects of macrophyte growth and decomposition on the nutrient budget in Shallow Lakes. A*quatic Botany*, 68, 217-237.

Schubert, Christiane C and Denmark, Kent and Crandall, Beth and Grome, Anna and Pappas, James (2013). *Annals of emergency medicine*, 61, 96-109.

Turnbull, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *JRSS*, Series B, 38:, pp. 290-295.

Young Firms and Murat Karoz. (2013). Cox Regression Models with Time-Varying Covariates Applied to Survival Success of Young Firms. *Journal ofEconomic and Social Studies*, .., ...

Zuidema, P. A., Leffelaar, P. A., Gerritsma, W., Mommer, L. Anten, N. P.R.,(2005). Physiological production model for cocoa (Theobroma cacao) : model presentation, validation and application. *Agricultural Systems*, 84, 195-225.