Journal of Modern Applied Statistical Methods

Volume 21 | Issue 1 Article 2

# An EM-Algorithm for Estimating the Lifetime Distribution with Long Terms Survival Data

Patrice Takam Soh
*Department of Mathematics, University of Yaounde, Cameroon*

Ulrich Florian Simo
*Department of Mathematics and Physical Sciences, National Advanced School of Engineering, University of Yaoundé I, Cameroon*

Ibrahim Moukouop Nguena
*Department of Mathematics and Physical Sciences, National Advanced School of Engineering, University of Yaoundé I, Cameroon*

Eugene Kouassi
*Department of Economics, University of the Western Cape, South Africa*

Recommended Citation

Patrice Takam Soh, Ulrich Florian Simo, Ibrahim Moukouop Nguena, Eugene Kouassi (2022). An EM-Algorithm for Estimating the Lifetime Distribution with Long Terms Survival Data. Journal of Modern Applied Statistical Methods, 21(1), https://doi.org/10.56801/Jmasm.V21.i1.2

Journal of Modern Applied Statistical Methods
2022, Vol. 21, No. 1,
Doi: 10.56801/Jmasm.V21.i1.2

ISSN 1538 – 9472

# An EM-Algorithm for Estimating the Lifetime Distribution with Long Terms Survival Data

**Patrice Takam Soh**
Department of Mathematics, University of Yaounde, Cameroon

**Ulrich Florian Simo**
National Advanced School of Engineering, University of Yaounde I, Cameroon

**Ibrahim Moukouop Nguena**
National Advanced School of Engineering, University of Yaounde I, Cameroon

**Eugene Kouassi**
Department of Economics, University of the Western Cape, South Africa

We are interested here in the estimation of the lifetime distribution by using the right censored observations taking into account the probability p that a censored item might contract the interest event after the right censoring. This type of censorship is known in the literature as long-term survival item. We assume that the lifetime follows a log-normal distribution with parameters μ and $\sigma^2$. An EM-algorithm is developed so as to estimate the parameters μ, σ and p. Simulations indicate good accurate and robustness results. The EM-

algorithm is then applied to the Cameroonian custom services data set in order to estimate the distribution of the return delays of the Global Position System (GPS). The EM-algorithm leads to accurate results of the parameters involved with smaller mean squared error.

*Keywords:* EM-algorithm; Estimation; lifetime distribution; right censoring; simulation; application.

## 1. Introduction

The problem to be addressed here is related to the estimation of the return delay distribution of the GPS (Global Position System) in the Customs context in Cameroon. In fact, the Port of Douala in Cameroon which provides 95% of national port traffic is also the first port of Economic and Monetary Community of Central Africa (CEMAC), serving Tchad and Central African Republic with preferential rates, e.g., see Pibasso (2010). To carry out its daily activities, especially in the efficient routing of the goods towards the stations borders (with its neighbors), Cameroon Customs established a control and monitoring goods system. The system is based on GPS that are connected on each truck in transit so as to enable the follow-up from an Information Technology (IT, hereafter) platform. But they are some challenges. The first challenge is that GPS is very expensive and customs need a specific tool to optimize the amount needed per day for the different travels. The second challenge is that the Cameroonian Customs need to manage the quantity of GPS in their routine activities. Also, the number of GPS necessary for the trucks transit development on a given day depends on the ones arrived the days before and on their supply. The third main challenge is that not all the GPS launched will return and this is due to the fact that some of them might be either lost or stolen. The Customs therefore need a forecasting tool to have an idea of the number of GPS that will eventually return after a given period.

Our attention here is then on the estimation of the lifetime distribution denoted here by the random variable $T$, that is the time elapsed from the launching day of a given GPS until its return day. The idea to keep in mind is that a GPS may not be returned by the deadline $L$ and this due to the reasons mentioned earlier. We therefore have two types of observations: (i) observations for which GPS is returned on or before the deadline $L$ (in which case the duration is completely observed); and (ii) observations for which the GPS is not returned to the date $L$ (it is therefore a data right censored in the classic sense). But the additional information here is the fact that among these GPS not returned at time period $L$, there is a proportion $p$ (unknown) which could have been returned if we had increased the value of $L$ and there is another $1 - p$ proportion that will never be returned. These particulars items are referred in the literature as *long-term survival items* or *cured items*. Of course, absent any other problems, we could manage the first three challenges (complete and censored data) with a MLE estimator. The fourth more serious challenge described as long-term survival posed other issues for which an EM algorithm is needed.

Our objective here is then not only to estimate the law of the time of return of the GPS but also this proportion $p$. We estimate $T$ in a parametric way. More precisely, we assume that $T$ follows a log-normal law (inspired by the ‘motivation problem’). Some examples of lifetime duration estimated under log-normal hypothesis can be found in Dube et al. (2011); Fan and Hsu (2014) and Hemmati and Khorram (2013). In the absence of the ‘long-term survival items’, the estimation problem is reduced to a standard case of estimation of the duration law. In the presence of right censoreship the problem is dealt with in a nonparametric way (see Kaplan and Meier (1958) or in a parametric way, see Cox (1984). By taking into account the long-term survival item, but assuming that the data follow a discrete law (that of Geometric for example), the problem has already been analysed in the literature by Carrasco and Ponce-Cueto (2009); Goh and Varaprasad (1986); Kelle and Silver (1989); Toktay et al. (2000) and Toktay et al. (2003). In our case, we are interested in taking this type of censorship into account but assuming a continuous law (in this case that of the log-normal law). This makes the problem more appealing and perhaps close to the reality on the ground.

The rest of the paper is organized as follows: Section 2 deals with the description of the GPS issue. Section 3 presents the GPS modeling process and related issues. Section 4 is related to the estimation procedure. Section 5 deals with the simulation study. Section 6 considers a real data application. Section 7 concludes the paper.

## 2. Description of the Problem

Douala is the second largest city of Cameroon with a renowned Port Authority. This port is used by almost all CEMAC countries including the landlocked countries, Central African Country, Chad etc. The port is facing a number of challenges including:

(i) Lost of goods;

(ii) Lost of competitiveness;

(iii) Delay in delivery of goods and services;

(iv) Stolen of goods etc

(v) And reputation

As a response to these issues, the Douala port authorities have developed an IT platform with the introduction of the GPS (Global Pointing System) for tracking, monitoring and assisting the port users. The Douala port authorities would like to provide answers to the following questions:

Q1: What is the distribution of the time elapsed from the launching day of a given GPS until its return day?

Q2: What is the maximum, the minimum, the mean, the mode, the median and the variance of this distribution?

Q3: What proportion $p$ of the GPS units will eventually return?

Indeed, the answer to these questions would allow Customs to predict the amount of GPS available in stock at any given time and this will allow them to know the amount of GPS to be available per day for trucks.

To answer the above questions, we follow a cohort of $N$ = 3645 trucks with GPS from March 3, 2014 to May 12, 2014. For each GPS, we record daily: its launch date and its eventual return date. The GPS units whose return date is not observed until the last observation date are therefore considered in ‘survival analysis’ as the right censored GPS. To take into account the fact that some of the GPS units censored on the right may have returned after the last observation date, we denote by $p$ this unobserved proportion which will be estimated from the observations. These are in fact the GPS that were not stolen and are not lost.

## 3. GPS Survival Modeling and Related Issues

### 3.1 Continuous-time modelling

We aim at estimating the time elapsed from initial time to the occurrence of an interest event. We consider a sample of $n$ items which are observed during a given period of time, from an origin date to an end date $L$. For each item $i$, we denote by $S_i$ its starting time and $R_i$ its end date, that is the moment where the interest event occurs. We assume that in the sample, there are exactly $K$ items of which $R_i < +\infty$ and $n - K$ of which $R_i = +\infty$, $K$ and $n - K$ representing respectively the cardinals of $J_1$ and $J_2$ where $J_1$ and $J_2$ are defined by

$$J_1 = \{i = 1, ..., n \mid R_i < +\infty\} \text{ and } J_2 = \{i = 1, ..., n \mid R_i = +\infty\}.$$

The time elapsed from the starting time to the end date is denoted by $T_i = R_i$-$S_i$ and $T_1$,...,$T_n$ are then considered as *i.i.d* continuous random variables. In the following, we assume that for each $i$, knowing $R_i < +\infty$, $T_i$ follows a log-normal distribution (motivated from empirical data), that is

$$\forall i \in J_1, \mathcal{L}\left(T_i \mid R_i < \infty\right) \sim \text{Log} - \mathcal{N}(\mu, \sigma^2) \quad \text{ with } \mu \in \mathbb{R}, \sigma > 0 \tag{1}$$

where $\mathcal{L}\left(T_i \mid R_i < \infty\right)$ is red as ’the distribution law of $T_i$ knowing that $R_i < \infty$’ and ~ is red as ’follows’. The probability density function of $T_i$ knowing that $R_i < +\infty$ can then be written as:

$$\mathbb{P}\left(T_i \in [t, t + dt) \mid R_i < \infty\right) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right) dt. \tag{2}$$

Since the distribution of $T$ is known in a parametric way, its mean, median and variance are given respectively by

- Mean: $\mathbb{E}(T) = \exp(\mu + \dfrac{\sigma^2}{2})$
- Median: $Me(T) = exp(\mu)$
- Mode: $M_0(T) = exp\ (\mu - \sigma^2)$
- Variance: $Var(T) = (exp\ (\sigma^2) - 1)\ exp\ (2\mu + \sigma^2)$

To answer questions Q1 and Q2, traditionally it suffices to estimate the parameters $\mu$ and $\sigma$. But in the present case there are some difficulties in terms of estimation; difficulties inherent to the nature of the data. The next section is more specific.

Discrete-time modelling

The observed data for a fixed $L$ (limit data of observation) is represented by $Y_i =$

$(X_i,\delta_i)$ where $X_i = \min(T_i, L_i)$ with $L_i = L - S_i$ and $\delta_i = 1_{\{R_{i \leq L}\}}$.

Setting $m_L = \sum\limits_i^n \delta_i = \sum\limits_i^n 1_{\{R_i \leq L\}}$, the $n$ items may be organized as follows

$$\begin{cases} X_i = T_i & \text{for } i = 1, \ldots, m_L \\ X_i = L_i & \text{for } i = m_L + 1, \ldots, n. \end{cases}$$

Let us denote by $\mathbb{P}(R_i < \infty) = p$ and $\mathbb{P}\ (T_i = +\infty, R_i = \infty) = \mathbb{P}\ (R_i = \infty) = 1 - p$. We can write that $\forall t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}\left(T_i \in [t,t+dt), R_i < \infty\right) &= \mathbb{P}(R_i < \infty) \times \mathbb{P}\left(T_i \in [t,t+dt) \mid R_i < \infty\right) \\ &= p \times \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right) dt.\end{aligned}$$

In the following we denote by $f_{T_i}(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t-\mu)^2}{2\sigma^2}\right)$ the probability density function of $T_i$ and $F_{T_i}$ its cumulative density function.

In the following, we consider two estimation methods: (i) the classical maximum likelihood estimation approach which allows to estimate $\mu$ and $\sigma$ and (ii) an EM algorithm type approach which allows to estimate $p$, $\mu$ and $\sigma$.

## 4. Estimation Procedure

We present two estimation techniques: the traditional maximum likelihood estimation and the EM-algorithm procedure.

### 4.1 Estimation of μ and σ² by Maximum-Likelihood Method

Under the classical assumptions of non-informative censorship and the independence between items, the likelihood for our sample containing complete observations and right censored observations can be written as

$$L(\mu,\sigma) = \prod_{i=1}^{n} \left\{\frac{1}{\sigma}\phi_{\mathbf{nor}}\left[\frac{\log(t_i)-\mu}{\sigma}\right]\right\}^{\delta_i} \left\{1 - \Phi_{\mathbf{nor}}\left[\frac{\log(t_i)-\mu}{\sigma}\right]\right\}^{1-\delta_i} \tag{3}$$

where $\varphi_{\mathbf{nor}}(z) = (1/\sqrt{2\pi})\exp(-z^2/2)$ is the probability density function for a standardized normal and $\Phi_{\mathbf{nor}}(z) = \int_{-\infty}^{z} \phi_{\mathbf{nor}}(w)dw$ the cumulative density function for a standardized normal.

The logarithm of the likelihood of observations is given by

$$\begin{aligned}\log L(\mu,\sigma) &= \sum_{i=1}^{n} \left\{-\delta_i\left[\log(\sigma) + \frac{1}{2}\log(2\pi) + \frac{(\log(t_i)-\mu)^2}{2\sigma^2}\right] + (1-\delta_i)\right. \\ &+ \left.(1-\delta_i)\log\left[1 - (1/\sqrt{2\pi})\int_0^{t_i} \exp\left(\frac{-(\log(u_i)-\mu)^2}{2\sigma^2}\right)du_i\right]\right\}\end{aligned}$$

and the gradient of log $L$ is given by $\left(\frac{\partial \log L}{\partial \mu}, \frac{\partial \log L}{\partial \sigma}\right)$ where

$$\begin{aligned}\frac{\partial \log L}{\partial \mu} &= \sum_{i=1}^{n} \left\{\delta_i\left[\frac{(\log(t_i)-\mu)}{\sigma^2}\right]\right. \\ &\left. +(1-\delta_i)\frac{\frac{-(1/\sqrt{2\pi})}{\sigma^2}\int_0^{t_i}(\log(u_i)-\mu)\exp\left(-\frac{(\log(u_i)-\mu)^2}{2\sigma^2}\right)du_i}{1-(1/\sqrt{2\pi})\int_0^{t_i}\exp\left(-\frac{(\log(u_i)-\mu)^2}{2\sigma^2}\right)du_i}\right\}\end{aligned}$$

and

$$\frac{\partial \log L}{\partial \sigma} = \sum_{i=1}^{n}\left\{-\delta_i\left[\frac{1}{\sigma}-\frac{(\log(t_i)-\mu)^2}{\sigma^3}\right] + (1-\delta_i)\frac{\frac{-(1/\sqrt{2\pi})}{\sigma^3}\int_0^{t_i}(\log(u_i)-\mu)^2\exp\left(-\frac{(\log(u_i)-\mu)^2}{2\sigma^2}\right)du_i}{1-(1/\sqrt{2\pi})\int_0^{t_i}\exp\left(-\frac{(\log(u_i)-\mu)^2}{2\sigma^2}\right)du_i}\right\}.$$

The estimators $(\mu\hat{},\ \sigma\hat{})$ are obtained by maximizing the log-likelihood under constraints $\mu \geq 0$ and $\sigma > 0$ using the command '*constrOptim*' (in R software) by integrating the gradient function as input in order to increase the precision of estimates.

This first method allows to estimate the parameters $\mu$ and $\sigma^2$ but it does not take into account the estimation of the probability $p$ which is our main concern. The next subsection is then devoted to the EM-algorithm which is formulated in order to take into account the probability $p$.

## 4.2 EM-Algorithm for Parameters Estimation

We recall here that $T = (T_1, \ldots ,T_n)$ is a random vector with a joint density $f_\theta(t) = f(t,\theta)$ where $\theta = (p,\mu,\sigma^2)$. As some of the complete-data vector $T$ is observed, we will denote by expressing $T$ as $(T^{\mathbf{obs}}, T^{\mathbf{uobs}})$ where $T^{\mathbf{obs}}$ denotes the observed and $T^{\mathbf{uobs}}$ the unobserved or missing data. More precisely, $T_i^{\mathbf{obs}} = T_i$ represents the elapsed time until the interest event when the end date is before $L$ and $T_i^{\mathbf{uobs}}(T_i^{\mathbf{uobs}} > L_i)$ represents the elapsed time until the interest event when that event occurs after $L$. The algorithm is presented in three steps: *(i)* complete likelihood, *(ii)* E-step and *(iii)* M-step.

In the following, for convenience, we note random variables by upper-case letters and their realizations by lower-case letters (for example $t_i$ and $r_i$ are used as a realization of random variable $T_i$ and $R_i$ respectively).

### 4.2.1. The Complete Likelihood

To write the complete likelihood, we proceed as follows: for a given individual $i$, its contribution to the likelihood is

$p\cdot\frac{1}{t_i\sigma\sqrt{2\pi}}e^{-\frac{(\log t_i-\mu)^2}{2\sigma^2}}$ if $i \in J_1$ or $(1 - p)$ if $i \in J_2$.

The complete likelihood of data is obtained by the product

$$\prod_{i=1}^{n}\left[\left(p\cdot\frac{1}{t_i\sigma\sqrt{2\pi}}e^{-\frac{(\log t_i-\mu)^2}{2\sigma^2}}\right)^{1_{\{i\in J_1\}}}(1-p)^{1_{\{i\in J_2\}}}\right]. \tag{4}$$

since individuals are supposed to be independent.

The logarithm of the complete likelihood is given by

$$\begin{aligned}
\ell_c\left(\theta, \underline{t}, \underline{r}\right) &= \sum_{i=1}^{n}\left\{1_{\{r_i<\infty\}}\log\left[p\cdot\frac{1}{t_i\sigma\sqrt{2\pi}}e^{-\frac{(\log t_i-\mu)^2}{2\sigma^2}}\right]+1_{\{r_i=\infty\}}\log(1-p)\right\}\\
&= K\log(p)-K\log(\sigma)-K\log(\sqrt{2\pi})+(n-K)\log(1-p)\\
&- \sum_{i=1}^{n}1_{\{r_i<\infty\}}\log(t_i)-\sum_{i=1}^{n}1_{\{r_i<\infty\}}\frac{(\log t_i-\mu)^2}{2\sigma^2}\\
&= K\log(p)-K\log(\sigma)+(n-K)\log(1-p)-K\frac{\mu^2}{2\sigma^2}+\frac{\mu}{\sigma^2}\sum_{i=1}^{n}1_{\{r_i<\infty\}}\log t_i\\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{n}1_{\{r_i<\infty\}}(\log t_i)^2+cste.
\end{aligned}$$

where $\underline{t}=(t_1,...,t_n), \underline{r}=(r_1,...,r_n)$, $cste=-K\log(\sqrt{2\pi})-\sum_{i=1}^{n}1_{\{r_i<\infty\}}\log(t_i)$.

The logarithm of the complete likelihood is then

$$\ell_c\left(\theta, \underline{t}, \underline{r}\right)=K\log(p)-K\log(\sigma)+(n-K)\log(1-p)-K\frac{\mu^2}{2\sigma^2}+\frac{\mu}{\sigma^2}K'-\frac{1}{2\sigma^2}K''+cste \quad (5)$$

Where $K$’ = $\sum_{i=1}^{n}1_{\{r_i<\infty\}}\log t_i$ and $K$’’ =$\sum_{i=1}^{n}1_{\{r_i<\infty\}}(\log t_i)^2$

### 4.2.2. E-step

This step consists of evaluating of the quantity $Q(\theta, \theta^{(j)})$ where $\theta^{(j)}$ represents the vector of parameters estimated at the iteration $r$ of the algorithm. Using the definition, we have $\theta = (p, \mu ,\sigma^2)$, $\theta^{(j)} = (p^{(j)},\mu^{(j)},\sigma^{2(j)})$ and

$$\begin{aligned}
Q(\theta,\theta^{(j)}) &= \mathbb{E}\left(\ell_c(\theta;\underline{t},\underline{r})\mid\theta^{(j)},Y\right)=\mathbb{E}_{\theta^{(j)}}\left(\ell_c(\theta;\underline{t},\underline{r})\mid Y\right)\\
&= \mathbb{E}_{\theta^{(j)}}\left(K\mid Y\right)\left[\log(p)-\log(\sigma)-\log(1-p)-\frac{\mu^2}{2\sigma^2}\right]+n\log(1-p)\\
&+ \frac{\mu}{\sigma^2}\mathbb{E}_{\theta^{(j)}}\left(K'\mid Y\right)-\frac{1}{2\sigma^2}\mathbb{E}_{\theta^{(j)}}\left(K''\mid Y\right).
\end{aligned}$$

where the conditional expectation is written $E_\theta(.) = E(. \mid \theta)$.

- We easily prove that

$$E\theta_{(j)}(K \mid Y) = m_L + \sum_{i=m_L+1}^{n}\alpha^i \quad (6)$$

where

$$\alpha_i=\frac{\mathbb{P}\left(R_i>L\mid R_i<\infty,\theta^{(j)}\right)p^{(j)}}{\mathbb{P}\left(R_i>L\mid\theta^{(j)}\right)} \quad (7)$$

with

$$p^{(j)}=\mathbb{P}\left(R_i<\infty\mid\theta^{(j)}\right),\ \mathbb{P}\left(R_i>L\mid R_i<\infty,\theta^{(j)}\right)=1-F_T\left(L_i;\mu^{(j)},\sigma^{(j)}\right)$$

where $F_T$ represents the cumulative density function of $T$ and

$$\mathbb{P}\left(R_i > L \mid \theta^{(j)}\right) = \mathbb{P}\left(R_i > L \mid R_i < \infty, \theta^{(j)}\right) p^{(j)} + (1 - p^{(j)}).$$

- using the same technique, we obtain

$$\mathbb{E}_{\theta^{(j)}}\left(K' \mid Y\right) = \sum_{i=1}^{m_L} \log(t_i) + \sum_{i=m_L+1}^{n} \mathbb{E}_{\theta^{(j)}}\left(\log T_i \mid T_i > L_i, R_i < \infty\right) \alpha_i \tag{8}$$

With

$$\mathbb{E}_{\theta^{(j)}}\left(\log T_i \mid T_i > L_i, R_i < \infty\right) = \frac{\mu^{(j)} - \int_0^{L_i} \log(x) f_T\left(x; \mu^{(j)}, \sigma^{(j)}\right) dx}{1 - F_T\left(L_i; \mu^{(j)}, \sigma^{(j)}\right)},$$

- In the same way, we have

$$\mathbb{E}_{\theta^{(j)}}\left(K'' \mid Y\right) = \sum_{i=1}^{m_L} [\log(t_i)]^2 + \sum_{i=m_L+1}^{n} \mathbb{E}_{\theta^{(j)}}\left([\log(t_i)]^2 \mid T_i > L_i, R_i < \infty\right) \alpha_i \tag{9}$$

where

$$\mathbb{E}_{\theta^{(j)}}\left([\log(T)]^2 \mid T_i > L_i, R_i < \infty\right) = \frac{\mathbb{E}_{\theta^{(j)}}\left([\log(T)]^2\right) - \int_0^{L_i} [\log(x)]^2 f(x; \mu^{(j)}, \sigma^{(j)}) dx}{1 - F_T\left(L_i; \mu^{(j)}, \sigma^{(j)}\right)}.$$

4.2.3. M-step

The M-step involves the maximization of $Q(\theta;\theta^{(j)})$, that is

$$\begin{aligned} Q(\theta; \theta^{(j)}) &= A_{\theta^{(j)}}(Y) \left[\log(p) - \log(\sigma) - \log(1-p) - \frac{\mu^2}{2\sigma^2}\right] \\ &\quad + n \log(1-p) + \frac{\mu}{\sigma^2} B_{\theta^{(j)}}(Y) - \frac{1}{2\sigma^2} C_{\theta^{(j)}}(Y) \end{aligned} \tag{10}$$

where $A_{\theta^{(j)}}(Y) = \mathbb{E}_{\theta^{(j)}}(K \mid Y)$, $B_{\theta^{(j)}}(Y) = \mathbb{E}_{\theta^{(j)}}(K' \mid Y)$ and $C_{\theta^{(j)}}(Y) = \mathbb{E}_{\theta^{(j)}}(K'' \mid Y)$

$$\begin{aligned} \frac{\partial Q}{\partial p}(\theta; \theta^{(j)}) &= A_{\theta^{(j)}}(Y) \left[\frac{1}{p} + \frac{1}{1-p}\right] - \frac{n}{1-p} = \frac{A_{\theta^{(j)}}(Y)}{p} - \frac{1}{1-p}\left(A_{\theta^{(j)}}(Y) - n\right) \\ \frac{\partial Q}{\partial \mu}(\theta; \theta^{(j)}) &= -\frac{A_{\theta^{(j)}}(Y)}{\sigma^2}\mu + \frac{B_{\theta^{(j)}}(Y)}{\sigma^2} \\ \frac{\partial Q}{\partial \sigma^2}(\theta; \theta^{(j)}) &= A_{\theta^{(j)}}(Y) \left[-\frac{1}{2\sigma^2} + \frac{\mu^2}{2\sigma^4}\right] - \frac{\mu}{\sigma^4} B_{\theta^{(j)}}(Y) + \frac{1}{2\sigma^4} C_{\theta^{(j)}}(Y) \end{aligned}$$

and by solving Equation $\nabla_\theta Q(\theta;\theta^{(j)}) = 0$, we obtain that

$$p = \frac{A_{\theta^{(j)}}(Y)}{n} \tag{11}$$

$$\mu = \frac{B_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} \tag{12}$$

$$0 = \frac{1}{\sigma^4}\left[A_{\theta^{(j)}}(Y)\left(-\frac{\sigma^2}{2} + \frac{\mu^2}{2}\right) - \mu B_{\theta^{(j)}}(Y) + \frac{1}{2} C_{\theta^{(j)}}(Y)\right] \tag{13}$$

where $\sigma^2$ is obtained by solving the last equation of the previous system. The link between the parameters at step $j$ and the ones for step $j + 1$ is then given by

$$\widehat{p}^{(j+1)} = \frac{A_{\theta^{(j)}}(Y)}{n} \tag{14}$$

$$\widehat{\mu}^{(j+1)} = \frac{B_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} \tag{15}$$

$$\widehat{\sigma}^{(j+1)2} = \left[\mu^{(j+1)}\right]^2 - 2\mu^{(j+1)}\frac{B_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} + \frac{C_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} \tag{16}$$

We can now confirm that the found values obtained are the maximum by checking the second order condition based on the Hessian Matrix of $Q$. The Hessian Matrix is then given by

$$H_{\mathbf{ess}}(\theta, \theta^{(j)}) = \left(\frac{\partial^2 Q(\theta; \theta^{(j)})}{\partial\theta_\ell, \partial\theta_k}\right)_{1\leq \ell, k \leq 3} = \begin{pmatrix} H_{pp}(\theta, \theta^{(j)}) & 0 & 0 \\ 0 & H_{\mu^2}(\theta, \theta^{(j)}) & H_{\mu\sigma^2}(\theta, \theta^{(j)}) \\ 0 & H_{\sigma^2\mu}(\theta, \theta^{(j)}) & H_{\sigma^4}(\theta, \theta^{(j)}) \end{pmatrix}$$

where $\theta_1 = p$, $\theta_2 = \mu$ and $\theta_3 = \sigma^2$. After some algebra we have

$$\begin{aligned} H_{pp}(\theta, \theta^{(j)}) &= -\frac{A_{\theta^{(j)}}}{p^2} + \frac{1}{(1-p)^2}\left(A_{\theta^{(j)}} - n\right) \\ H_{\mu^2}(\theta, \theta^{(j)}) &= -\frac{A_{\theta^{(j)}}}{\sigma^2}; \quad H_{\mu\sigma^2}(\theta, \theta^{(j)}) = \frac{A_{\theta^{(j)}}}{\sigma^4}\mu - \frac{B_{\theta^{(j)}}}{\sigma^4} \\ H_{\sigma^4}(\theta, \theta^{(j)}) &= A_{\theta^{(j)}}(Y)\left[\frac{1}{2\sigma^4} - \frac{2\mu^2}{\sigma^6}\right] + \frac{2\mu}{\sigma^6}B_{\theta^{(j)}}(Y) - \frac{1}{\sigma^6}C_{\theta^{(j)}}(Y) \end{aligned}$$

.

If $j_c$ +1 represents the convergence step, that is $\hat{\theta} = (\hat{p}, \hat{\mu}, \hat{\sigma}^2) = \theta^{(jc+1)}$ the Hessian Matrix at $(\hat{\theta}, \theta^{(jc)}) = (\theta^{(jc+1)}, \theta^{(jc)})$ is obtained as follows:

$$\begin{aligned} H_{pp}(\hat{\theta}, \theta^{(j_c)}) &= -\frac{A_{\theta^{(j_c)}}}{\hat{p}^2} + \frac{1}{(1-\hat{p})^2}\left(A_{\theta^{(j_c)}} - n\right) \\ H_{\mu^2}(\hat{\theta}, \theta^{(j_c)}) &= -\frac{A_{\theta^{(j_c)}}}{\hat{\sigma}^2}; \quad H_{\mu\sigma^2}(\hat{\theta}, \theta^{(j_c)}) = \frac{A_{\theta^{(j_c)}}}{\hat{\sigma}^4}\hat{\mu} - \frac{B_{\theta^{(j_c)}}}{\hat{\sigma}^4} \\ H_{\sigma^4}(\hat{\theta}, \theta^{(j_c)}) &= A_{\theta^{(j_c)}}(Y)\left[\frac{1}{2\hat{\sigma}^4} - \frac{2\hat{\mu}^2}{\hat{\sigma}^6}\right] + \frac{2\hat{\mu}}{\hat{\sigma}^6}B_{\theta^{(j_c)}}(Y) - \frac{1}{\hat{\sigma}^6}C_{\theta^{(j_c)}}(Y) \end{aligned}$$

.

Using the relationship between $(\theta^{(jc+1)})$ and $\theta^{(jc)}$ given by (14) − (16) and the fact that $\theta^{(jc+1)} \approx \theta^{(jc)}$ (since $\theta^{(jc+1)}$ is the convergence value) we deduce that

$$A_{\theta^{(j)}} = \widehat{p}n; \quad B_{\theta^{(j)}}(Y) = A_{\theta^{(j)}}(Y)\widehat{\mu} \text{ and} \tag{17}$$

$$\widehat{\sigma}^2 = \hat{\mu}^2 - 2\hat{\mu}\frac{B_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} + \frac{C_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} = -\hat{\mu}^2 + \frac{C_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)} \tag{18}$$

and we have

$$H_{pp}(\hat{\theta},\theta^{(j)}) = -\frac{n}{\hat{p}(1-\hat{p})};\ H_{\mu^2}(\hat{\theta},\theta^{(j)}) = -\frac{n\hat{p}}{\hat{\sigma}^2};\ H_{\mu\sigma^2}(\hat{\theta},\theta^{(j)}) = 0$$

$$H_{\sigma^4}(\hat{\theta},\theta^{(j)}) = -\frac{n\hat{\mu}^2\hat{p}}{2\hat{\sigma}^6}\text{ since}\left(\hat{\sigma}^2+\hat{\mu}^2 = \frac{C_{\theta^{(j)}}(Y)}{A_{\theta^{(j)}}(Y)}\right).$$

The Hessian Matrix at ($\hat{\theta}$, $\theta^{(jc)}$) is then given by

$$H_{\mathbf{ess}}(\hat{\theta},\theta^{(j_c)}) \approx \begin{pmatrix} -\frac{n}{\hat{p}(1-\hat{p})} & 0 & 0 \\ 0 & -\frac{n\hat{p}}{\hat{\sigma}^2} & 0 \\ 0 & 0 & -\frac{n\hat{\mu}^2\hat{p}}{2\hat{\sigma}^6} \end{pmatrix}.$$

Since all eigenvalues are negative, $H_{\mathbf{ess}}(\hat{\theta},\ \theta^{(jc)})$ is then negative definite and we deduce that $\hat{\theta}$ is the maximum of $Q$.

### 4.3 The Steps of the EM-algorithm

In practice, the different steps of the algorithm can be summarized as follows:

- **Step 1**: We initialize the value of $\theta$ by $\theta^{(0)} = (p^{(0)},\mu^{(0)},\sigma^{(2,0)})$. In this work, $(\mu^{(0)},\sigma^{(2,0)})$ is initialized by the estimates obtained from maximum likelihood method and $p$ is initialized at 0.5.
- **Step 2**: We use the expression of $A_{\theta^{((j)}}(Y\ )$ and $B_{\theta^{(j)}}(Y\ )$ in order to deduce the expression of $\theta^{(j+1)}$ from the ones of $\theta^{(j)}$
- **Step 3**: We evaluate the distance between $\theta^{(j+1)}$ and $\theta^{(j)}$ and compare it with a given threshold (for example $\epsilon = 10^{-3}$)
- **Step 4**: When the distance is less than $\epsilon$, we stop the algorithm and the estimation is the current value of $\theta$, if not we evaluate the new values of $\theta$. We then repeat these steps until convergence.

## 5. Simulation study

In this Section, in order to examine the performance of the proposed estimators, we conduct several simulations; the accuracy and robustness of the estimators involved are then assessed.

### 5.1 The Design

To analyze the accuracy of the estimators obtained by the aforementioned algorithm, we apply the method to several samples of data obtained from the log-normal distribution with different parameter values. These parameters are choosen in order to cover the different speeds of the cumulative function of log-normal distribution, that is $\mu \in \{0.1,1\}$ and $\sigma \in \{0.12,\ 0.25,\ 0.5\}$. The limit time $L$ has been chosen according to the values of ($\mu$, $\sigma$) and the different values are presented in Table 1.

**Table 1.** Simulation’s Different Values

| | | | | | | |
|---|---|---|---|---|---|---|
| $\mu$ | 1 | 1 | 1 | 0.1 | 0.1 | 0.1 |
| $\sigma$ | 0.12 | 0.25 | 0.5 | 0.12 | 0.25 | 0.5 |
| $L$ | 3.1 | 4 | 8 | 1.6 | 2.2 | 3.8 |

It is worth mentioning that the values of $L$ is chosen such that we have enough values of censored data in our sample. The simulations are worked out for three different values of sample size ($n$ = 100, 250, 500, 1000 and 2000). For each value of $\mu$ and $\sigma$ and for each value of sample size $n$, we simulate the lifetime for $n$ items from log-normal distribution by using the fact that if the simulated value is less than $L$ this value is considered as the observed lifetime of $i$ otherwise its observation lifetime is $L$. For each simulated dataset, we used the EM-algorithm to estimate the values of $p$, $\mu$ and $\sigma$ and the maximum likelihood method as well to estimate $\mu$ and $\sigma$; we then compare the two approaches.

To analyse the variability of the estimators obtained by the EM-algorithm, we construct bootstrap samples of estimators in a parametric way following Efron and Tibshirani (1994). To do this, we simulate the data from estimated parameters, and we re-estimate the parameters using simulated data and so on (we repeat the process $B$ = 1000 times, $B$ equals the number of replications). We then obtain a bootstrap sample of estimated parameters. The bootstrap estimator is obtained as the average of this sample. The estimated bias (Est.bias) is therefore obtained as the difference between the bootstrap estimator and the one deduced from the data. The estimated standard deviation (Est.sd) is that of the bootstrap sample. The variability of the estimators obtained from the maximum likelihood method are obtained asymptotically. These different values are presented in Table 2.

To analyse the robustness of our method, we consider its behaviour on the data using different distributions, frequently cited in the survival analysis: Poisson distribution (Pois.), Geometric distribution (Geom.), Negative binomial distribution (Neg. B.), Compound Poisson distribution (Com-P.), Log-normal distribution (Log-nor.), Gamma distribution (Gamma) and Weibull distribution (Weibull). For each of these distributions, we use the real data (presented in the following Section) to estimate by maximum likelihood method the different corresponding parameters (e.g, see Table 5). We then use these estimated parameters to simulate the data with the same sample size close to the one of our data, that is $n$ = 3000. The value of $L$ is chosen according to the distributions ($L$ = 10 for log-norm, $L$ = 12 for Poisson, nbinon, Com-pois, Gamma and Weibull, $L$ = 35 for geometric distribution) and we compare the different criteria (Est.Bias and Est.MSE =(Est.Bias)$^2$+(Est.SD)$^2$, where Est.Bias and Est.SD are obtained from nonparametric bootstrap) in order to check how our method behaves when the data come from a distribution different from the log-normal one.

### 5.2 Results and Interpretations

From Table 2, we observe that: (*i*) regardless of the estimation method, the estimates of $\mu$ and $\sigma$ do not depend on the sample size; (*ii*) the values of $\mu$ˆ estimated from Likelihood method (MLE) are almost constant whatever the sample size and look higher than the ones obtained from EM-algorithm; (*iii*) the values of ˆ$\sigma$ obtained from EM-algorithm are in general lower than the ones obtained by the MLE method regardless of the sample size. From this last observation, we can deduce that the EM-algorithm is more accuracy than the MLE method.

From Table 3 we can observe that the convergence speed is the same (number of iteration = 3) for all the distribution laws except the case where the data come from the geometric distribution. Looking at the estimated bias in absolute value and the mean square error, we observe that, except the Poisson distribution where the value is higher than the corresponding values; for other distributions we do not observe a real difference between the values of estimated bias and estimated MSE for all the distributions. From this last observation it is clear that our method outperforms the MLE method.

## 6. A Real Data Example

### 6.1 The Data

The data are daily and span the period March 3, 2014 to May 12, 2014 i.e $N = 3645$ observations or trucks with GPS. For each GPS, we observe its date of put in transit ($s_i$), its date back in stock ($r_i$), and $t_i = r_i - s_i + 1$ its return delays (in days). We recall here that these data come from a large data set generated from the Nexus platform, which is the computer system that assists the customs transit in GPS aids connected on trucks. For each transit, we have three possible itineraries or corridors (Itinerary 1: Douala-Yassa-Bonis-Bogdibo, Itinerary 2: Douala-Yassa-Bonis-Kousseri, and Itinerary 3: Douala-Yassa-Bonis-Garoua Boulais). Note that the GPS that are put in transit before and do not return before are considered as right censored GPS. For each of these three itineraries, we apply the EM-algorithm with 03 different values of $L$ (corresponding respectively to $L_1$= April 14, 2014, $L_2$= April 28, 2014 and $L_3$= May 12, 2014) and the results are compared.

In the following, the data are then organized following the three different itineraries. Some basic statistics are provided in 4. In Table 4, by considering each itinerary, the following information are obtained: the minimum of the return delay (in days) of both itineraries (1) and (2) are the same simply because the cities Kousseri and Bogdibo are geographically close. The minimum return delay of itinerary (3) is lower than the others because of its proximity with Douala compared to Bogdibo and Kousseri. In addition, when the number of items is small, the standard deviation of the different return delays is also high. In that case we observe more truck returns.

### 6.2 Distribution of the Data

In order to find the distribution followed by the data, we use the parameters obtained in Table 5 to build the estimated cumulative function and put in competition on the same Figure different candidate distributions. The purpose is to choose that best fit the data. Different cumulative density functions are presented in Figure 2 corresponding to the three itineraries. Based on results (in fact the characteristics of each distribution) in Figure 2, it turns out that return delay distribution for each itinerary is closer to a log-normal distribution. This result is motivated by the Kolmogorov-Smirnov test statistic (the lowest value of D in Table 5). It should be noticed, however, that the p-value of the Kolmogorov test is not significant and this is explained by the fact that the data contain significant ex-eaquo values.

### 6.3 Estimation of parameters Using EM-algorithm

We observe that the algorithm converges very fast, particularly just after 3 iterations as we can see in Figure 3. The estimates are reported in Table 7 where we recall that the algorithm has been initialized by the maximum Likelihood estimates. From Table 7, we observe that: (i) the estimated $p$ value is always around 0.9 regardless of the itinerary, which means that based on the current available data, a very small number of GPS are actually lost; (ii) the estimated values of $\mu$ and $\sigma$ are almost the same for the three itineraries, which means that the delay is almost the same for the three itineraries; *(iii)* the Est.MSE of $p$ and $\mu$ are very small ($< 10^{-2}$) but the ones of $\sigma^2$ is around 0.75, which means that the estimate of $\sigma^2$ is less accurate than those of the two other parameters. This estimate of $\sigma^2$ may be sensitive to the sample size. This can be illustrated by the fact that in Figure 1, in the distribution of $\sigma^2$, the EM-algorithm does not correspond to the peak of the distribution.;

Results also indicate that the mean delay is 10 days for Itinerary 1 (Table 9); 11 days for Itinerary 2 (Table 10); and 7 days for Itinerary 3 (Table 11). In each case, the median is close to the mean; and the mode is not different from the median. The above results are obtained from the formula in 3.1. The above results are obtained from formula described

### 6.4 Analysis of Variability of the Different Estimators

To determine the variability of the different estimators, we use the parametric Bootstrap with re-sampling and based on replications. The aim here is to obtain (i) the estimated bias, (ii) the estimated variance and (iii) the MSE.

In Figure 1 we report the estimated density of the different parameter distributions and for each of them, we draw the vertical line corresponding to the parameters estimation obtained from EM-algorithm. This Figure shows that the estimations always correspond almost to the peak of the curve.

## 7. Conclusions

We propose here an EM-algorithm in order to estimate the parameters of lognormal distribution using right censored observations in a particular case where the interest event might occur after the censorship. The methodology advocated is applied in order to obtain GPS return delay distribution in customs context in Cameroon.

A simulation study indicates the good behaviour of the estimators involved regardless of the sample size and the distribution used.

Comparing results from the EM-algorithm and the MLE we notice that the first one outperforms the latter from all perspectives. We then conclude that the EM-algorithm proposed is appropriate in resolving the problem under consideration.

### 7.1 Acknowledgements

We thank the African Center of Excellence in Technologies, Information and Communication (CETIC) for its assistance. We would also like to thank anonymous referees for helpful comments. The usual disclaimer applies.

## References

Aragon, Y. (2011). Séries temporelles avec R : Méthodes et cas, Université Toulouse 1 – Capitole, Springer Paris Berlin Heidelberg New York.

Box, GEP, Jenkins, GM and Reinsel, GC (1976). Time Series Analysis, Forecasting and Control, Holden-Day, Third Edition, Series G. 13

Carrasco-Gallego, R and Ponce-Cueto, E (2009). Forecasting the returns in reusable containers’ closed-loop supply chains: A case in the LPG industry, 3rd International Conference on Industrial Engineering and Industrial Management, Barcelona-Terrassa.

Cox, DR and Oakes, D (1984). Analysis of survival data (Vol. 21). CRC Press.

Dempster, AP, Laird, NM and Rubin, DB (1977) ‘Maximum likelihood from incomplete data via the EM algorithm’, Journal of the Royal Statistical Society, Series B, 39(1), 1-38.

Dube, S, Pradhan, B and Kundu, D (2011) ‘Parameter estimation of the hybrid censored lognormal distribution’, Journal of Statistical Computation and Simulation, 81(3), 275 287.

Efron, B and Tibshirani, RJ (1994). An Introduction to the Bootstrap. CRC Press.

Fan, TH and Hsu, TM (2014) ‘Statistical inference of a two-component series system with correlated log-normal lifetime distribution under multiple type-I censoring’, IEEE Transactions on Reliability, 64(1), 376-385.

Goh, TN and Varaprasad, N (1986) ‘A statistical methodology for the analysis of the Life Cycle of Reusable Containers, IEEE Transactions, 18(1), 42-47.

Hemmati, F and Khorram, E (2013) ‘Statistical analysis of the log-normal distribution under type-II progressive hybrid censoring schemes’, Communications in Statistics -Simulation and Computation, 42(1), 52-75.

Kaplan, EL and Meier, P (1958) ‘Nonparametric estimation from incomplete observations’, Journal of the American Statistical Association, 53(282), 457-481.

P. Kelle and E. A. Silver, (1939). Forecasting the Returns of Reusable Containers, Journal of Operations Management, 8(1), pp. 17-35.

Klein, JP and Moeschberger, ML (1997). Survival analysis: techniques for censored and truncated data, Springer-Verlag Vol., New York, USA.

Pankratz, A (1991). Forecasting with dynamic regression models, New York: Wiley.

Pibasso, AM (2010). Transit Cameroun-Centrafrique et Tchad, le GPS de la discorde, http:// centrafrique-presse.over-blog.com/article-transit-cameroun-centrafrique.

Thomas, AL (1982) ‘Finding the observed information matrix when using the EM-algorithm’, Journal of the Royal Statistical Society B, 44(2), 226-233.

Toktay, LB, Wein, LM and Zenios, SA (2000) ‘Inventory management of remanufacturable Products’, Management Science, 46(11), 1412-1426.

Toktay, LB and Van Der Laan, EA and De Brito, MP (2003) Managing Product Returns: The Role of Forecasting, Econometric Institute Report EI.