

The Long-Term Exponentiated Complementary Exponential Geometric Distribution under a Latent Complementary Causes Framework

F. LOUZADA^{1*}, C.Y. YAMACHI², V.A.A. MARCHI² and M.A.P. FRANCO²

Received on January 13, 2013 / Accepted on March 11, 2014

ABSTRACT. In this paper we proposed a new long-term distribution derived from the exponentiated complementary exponential geometric distribution (LECEG). The LECEG distribution is obtained straightforwardly from the exponentiated complementary exponential geometric (ECEG) and accommodates decreasing and unimodal hazard functions in a latent complementary causes scenario, where only the maximum lifetime among all causes is observed. We derive the density, quantile, survival and failure rate functions for the proposed distribution, as well as some proprieties such as the characteristic function, mean, variance and r -th order statistics. The estimation is based on maximum likelihood approach. A simulation study is performed in order to assess the performance of the maximum likelihood estimates. The practical importance of the new distribution was demonstrated in three real datasets.

Keywords: exponentiated complementary exponential geometric distribution, latent competing risks, long-term survivals.

1 INTRODUCTION

Survival analysis are usually considered in areas such as public health, actuarial science and industrial reliability having with initial approach the exponential distribution. The use of the exponential distribution requires that the failure rate function be constant. However, sometimes the constant failure rate is not satisfied and so distributions which non-constant failure rate are necessary. The literature is vast and grows constantly and interested readers can refer to Adamidis & Loukas (1998), M. Chahkandi & Ganjali (2009), Louzada et al. (2011), Morais & Barreto-Souza (2010), Cancho et al. (2011), F. Hemmati et al. (2011), Bakouch et al. (2011), Louzada et al. (2013) amongst others.

Adamidis & Loukas (1998) proposed the exponential geometric (EG) distribution. The EG distribution is a compound of the geometric with the exponential distribution. The EG distribution

*Corresponding author: Francisco Louzada

¹Departamento de Matemática Aplicada e Estatística, ICMC, USP, 13566-590 São Carlos, SP, Brazil.
E-mail: louzada@icmc.usp.br

²Departamento de Estatística, UFSCar, 13565-905 São Carlos, SP, Brazil.

E-mails: cintiayurie@yahoo.com.br; vitor.marchi@gmail.com; mapfranco@ufscar.br

is characterized in a latent competing risks scenarios (Louzada-Neto, 1999) where the competing causes are unknown and only the minimum failure lifetime is observed. It also accommodates decreasing failure rates. Louzada et al. (2011) proposed the complementary exponential geometric (CEG) distribution. The CEG distribution is characterized in a latent complementary risks scenarios and accommodate increasing failure rate. The latent complementary scenario is characterized where the competing causes are unknown and only the maximum failure lifetime is observed. Yamachi et al. (2013) proposed the exponentiated complementary exponential geometric (ECEG) distribution which is a generalization of the CEG distribution. The ECEG distribution accommodate increasing and decreasing failure rates.

In some cases one part of the population is not susceptible to the event of interest and according to R. Maller & X. Zhou (1996), it seems adequate to consider the two components mixture model, where a component represents the survival time of susceptible individuals to a certain event (in risk – IR), while the other component represents the survival times of the not susceptible individuals to the event (out of risk – OR), allowing infinite survival times. An individual belongs to one group or another with certain probability.

The two components mixture model has been used in medicine, especially for data analysis of cancer clinical trials where we observe the time to occurrence of death, or the time until the outbreak of a disease, but in the presence of a significant proportion of cured or immune patients (J.W. Boag, 1949; J. Berkson & R.P. Gage, 1952). Farewell (1982) worked with some distributions with long term survivors, (Maller & S. Zhou, 1995) discuss test for models with presence of people immune to event, (Chen & Ibrahim, 2001) analyse maximum likelihood method to estimate models with cure fraction and missing covariates, (Cancho et al., 2009) introduced the Log-exponentiated-Weibull Regression Models with cure rate, (Perdona & Louzada, 2011) with a failure rate model in the presence of immune patients.

The long-time model formulation using the components mixture model is described as following. Let Y be a random variable that represents the time until the occurrence of a event of interest, and p be the probability of an individual belongs to the group OR. Considering that exist the possibility of the individual be not susceptible to the event of interest, the improper population survival function is given by (R. Maller & X. Zhou, 1996), $S(y) = pS_{OR}(y) + (1 - p)S_{IR}(y)$, where $S_{OR}(y)$ and $S_{IR}(y)$ are the survival functions of the individuals OR and IR, respectively. The individual in OR not present the time of the event of interest, so $S_{OR}(y) = 1$. Then, $S(y)$ can be rewritten as,

$$S(y) = p + (1 - p)S_{IR}(y). \quad (1.1)$$

From (1.1), $\lim_{y \rightarrow \infty} S(y) = p$, and therefore the survival function is improper and the limit correspond to the OR proportion.

Considering that the event of interest may be caused by unknown number of competing/complementary causes (Louzada-Neto, 1999), we have a scenario with latent competing/complementary failure causes in presence of long-term survivals. We can use for $S_{IR}(y)$ the EG or CEG distributions, amongst others. For example, (Roman et al., 2012) presented the LEG distribution, that uses the EG distribution in $S_{IR}(y)$.

Assuming that the $S_{IR}(y)$ is given by the ECEG distribution, we have

$$S_{IR}(y) = 1 - \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha. \tag{1.2}$$

Using (1.2) in (1.1) we have the survival function of the long-term exponentiated complementary exponential geometric (LECEG) distribution.

The paper is organized as follows. Section 2 presents the density, quantile, survival and failure rate functions for the proposed distribution and some proprieties such the characteristic function, mean, variance and k -th order statistics. Section 3 presents the inference based on maximum likelihood approach. Section 4 presents the results of a simulation study to assess the performance of the maximum likelihood estimates. Section 5 presents the LECEG distribution in presence of covariates. Section 6 illustrates the application of the proposed distribution in three real datasets. Section 7 provides some concluding remarks.

2 MODEL FORMULATION

Let Y be a nonnegative random variable denoting the lifetime of a component. The random variable Y is said to have a LECEG distribution with parameters $\alpha > 0$, $\lambda > 0$, $0 < \theta < 1$ and $0 < p < 1$ if its probability density function (pdf) is given by,

$$f(y) = \frac{(1 - p)\alpha}{(1 - e^{-\lambda y})} \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha \left(\frac{\lambda e^{-\lambda y}}{e^{-\lambda y}(1 - \theta) + \theta} \right), \tag{2.1}$$

where λ is the scale parameter, α and θ are shape parameters and p is the long-term parameter. Figure 1 shows the LECEG pdf and survival function for $p = 0.1$, $\alpha = 0.1, 0.4, 0.7, 1, 2, 3.5$, $\theta = 0.2, 0.7, 0.8$ and $\lambda = 2, 3$.

The improper survival function of the LECEG distribution is given by,

$$S(y) = 1 - (1 - p) \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha, \tag{2.2}$$

where $y > 0$, $\theta \in (0, 1)$, $\alpha, \lambda > 0$, $p \in (0, 1)$.

The quantile function of (2.1) is given by

$$Q(u) = \log \left(\frac{\theta \left(1 - \left(\frac{u}{1-p} \right)^{\frac{1}{\alpha}} \right)}{\left(\frac{u}{1-p} \right)^{\frac{1}{\alpha}} (1 - \theta) + \theta} \right) \frac{1}{\lambda}, \tag{2.3}$$

where $u \in (0, 1 - p)$.

From (2.1) and (2.2) it is easy verify that the failure rate function for the LECEG distribution is given by

$$h(y) = \frac{(1 - p)\alpha\lambda e^{-\lambda y}\theta^\alpha(1 - e^{-\lambda y})^{\alpha-1}}{(e^{-\lambda y}(1 - \theta) + \theta)^{\alpha+1} - (1 - p)\theta^\alpha(1 - e^{-\lambda y})^\alpha(e^{-\lambda y}(1 - \theta) + \theta)} \tag{2.4}$$

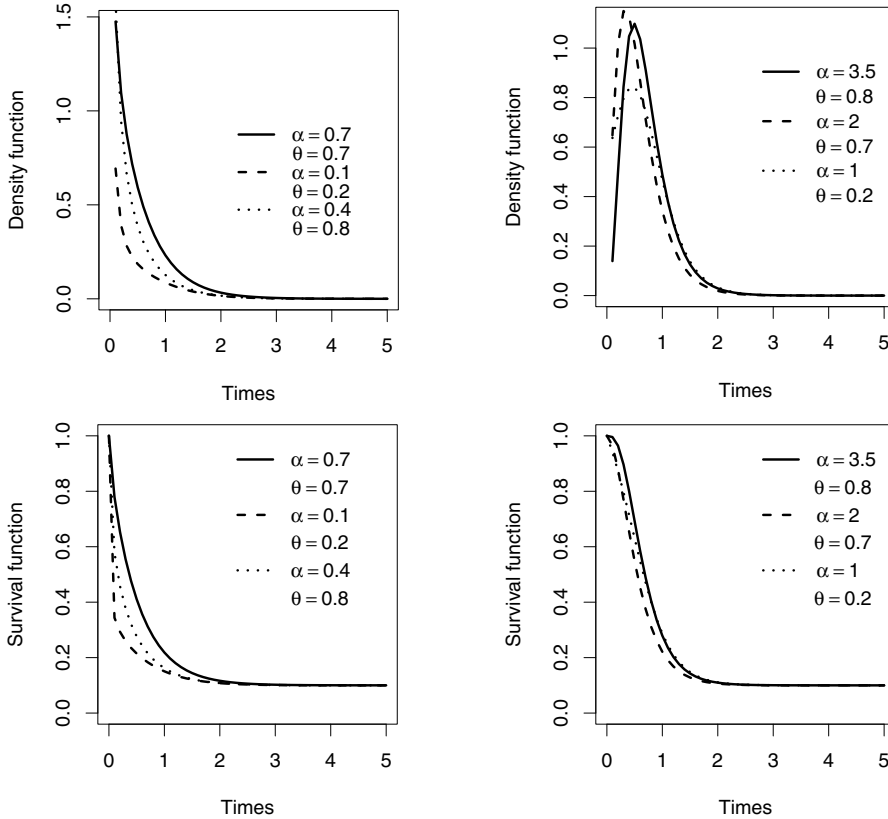


Figure 1: Upper panels: Probability density function. Lower Panels: Survival function, with $p = 0.1$ and $\lambda = 2$ and $\lambda = 3$ from left to right.

Figure 2 shows the failure rate function for $\alpha = 0.1, 0.4, 0.7, 1, 2, 3.5, \lambda = 2, 3, p = 0.1$ and $\theta = 0.2, 0.7, 0.8$. It is noted that the failure rate function (2.4) is decreasing or unimodal. Assuming some values to the parameters α, λ, θ and p , it is noted that: If $0 < \alpha < 1$ we have decreasing failure function and if $\alpha > 1$ we have unimodal failure rate function.

From Figure 2 it is noted that the failure function decreases as the proportion of cured patients increases.

Proposition 2.1. *The characteristic function of a random variable Y with LECEG distribution is given by*

$$\Phi_Y(t) = (1 - p)\alpha\theta^\alpha \Gamma\left(\frac{-iy}{\lambda} + 1\right) \sum_{k=0}^{\infty} \binom{-\alpha - 1}{k} \frac{(\theta - 1)^k \Gamma(\alpha + k)}{\Gamma(\alpha + 1 - \frac{iy}{\lambda} + k)}. \tag{2.5}$$

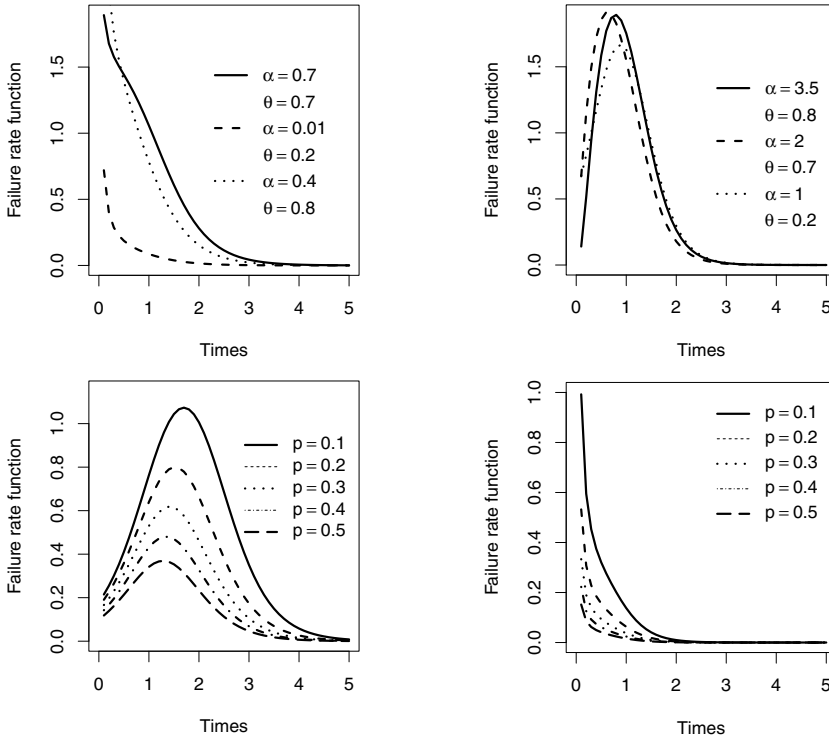


Figure 2: Upper panels: Failure rate function with $p = 0.1$, $\lambda = 2$ and $\lambda = 3$ from left to right. Lower panels: Failure rate function with $\alpha = 1$, $\lambda = 2$ (left) and $\lambda = 3$ (right) and $\theta = 0.1$ to different proportions, p .

Proof. Using the known formula $\Phi_Y(y) = \int_0^\infty e^{ity} f(t)dt$,

$$\begin{aligned} \Phi_Y(t) &= \int_0^\infty e^{ity} \frac{(1-p)\alpha}{(1-e^{-\lambda t})} \left(\frac{\theta(1-e^{-\lambda t})}{e^{-\lambda t}(1-\theta) + \theta} \right)^\alpha \left(\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}(1-\theta) + \theta} \right) dt \\ &= (1-p)\alpha\theta^\alpha \int_0^1 u^{\alpha-1} (1-u)^{\frac{-iy}{\lambda}} (1+u(\theta-1))^{-\alpha-1} du, \end{aligned} \tag{2.6}$$

where $u = 1 - e^{-\lambda t}$. Using (7.2) for solve the integral in (2.6) the characteristic function is obtained. □

Proposition 2.2. *The mean and variance from a random variable Y with LECEG distribution are given by*

$$\begin{aligned} E(Y) &= -\frac{i\alpha(1-p)\theta^\alpha \Gamma(\alpha) 2F1^{(0,0,1,0)}(\alpha, \alpha+1, \alpha+1, 1-\theta)}{\lambda \Gamma\left(\frac{\alpha\lambda+\lambda}{\lambda}\right)} + \frac{i\gamma\alpha(1-p)\Gamma(\alpha)}{\lambda \Gamma\left(\frac{\alpha\lambda+\lambda}{\lambda}\right)} \\ &+ \frac{i\alpha(1-p)\Gamma(\alpha)\psi^{(0)}\left(\frac{\alpha\lambda+\lambda}{\lambda}\right)}{\lambda \Gamma\left(\frac{\alpha\lambda+\lambda}{\lambda}\right)}, \text{ and} \end{aligned}$$

$$\begin{aligned}
 V(Y) = & \frac{\alpha(p-1)\theta^\alpha \Gamma(\alpha) \left(-2\gamma \Gamma(\alpha+1) {}_2F_1^{(0,0,1,0)}(\alpha, \alpha+1, \alpha+1, 1-\theta) \right. \\
 & \left. + \Gamma(\alpha+1) {}_2F_1^{(0,0,2,0)}(\alpha, \alpha+1, \alpha+1, 1-\theta) \right)}{\lambda^2 \Gamma(\alpha+1)^2} \\
 & - \frac{2\Gamma(\alpha+1) \psi^{(0)}(\alpha+1) {}_2F_1^{(0,0,1,0)}(\alpha, \alpha+1, \alpha+1, 1-\theta) \\
 & + \alpha(p-1)\theta^\alpha \Gamma(\alpha) \left(2F_1^{(0,0,1,0)}(\alpha, \alpha+1, \alpha+1, 1-\theta) \right)}{\lambda^2 \Gamma(\alpha+1)^2} \\
 & + \frac{\theta^{-\alpha} (-\psi^{(0)}(\alpha+1) - \gamma)^2 + \frac{1}{6} \pi^2 \theta^{-\alpha} \Gamma(\alpha+1) + \gamma^2 \theta^{-\alpha} \Gamma(\alpha+1) \\
 & + \theta^{-\alpha} \Gamma(\alpha+1) \psi^{(0)}(\alpha+1)^2 + 2\gamma \theta^{-\alpha} \Gamma(\alpha+1)}{\lambda^2 \Gamma(\alpha+1)^2} \\
 & \times \frac{\psi^{(0)}(\alpha+1) - \theta^{-\alpha} \Gamma(\alpha+1) \psi^{(1)}(\alpha+1)}{\lambda^2 \Gamma(\alpha+1)^2},
 \end{aligned}$$

respectively, where ${}_2F_1^{(0,0,1,0)}(a, b, c; z) = -\sum_{k=0}^{\infty} \frac{(a)_k (b)_k \psi(c+k) z^k}{k! (c)_k} + \psi(c) {}_2F_1(a, b, c; z)$, γ is the Euler's constant $\gamma \cong 0.5772$, $\psi^{(0)}$ is the digamma function, $\psi^{(1)}$ is the first derivative of the digamma function, $(x)_k$ is the pochhammer symbol $\Gamma(x+k)/\Gamma(x)$, $\psi(x) = \Gamma'(x)/\Gamma(x)$ and ${}_2F_1^{(0,0,2,0)}(a, b, c; z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{k!} \frac{\partial^2}{\partial c^2} \frac{1}{(c)_k} z^k$.

Proof. Considering that $E(Y^r) = \Phi_Y(0)^{(r)}/i^r$, we can obtain the first and second moments, $E(Y)$ and $E(Y^2)$, respectively. Using the first and second moments, the variance is obtained by $V(Y) = E(Y^2) - [E(Y)]^2$. The results were founded using the Mathematica Software (2010). \square

Order statistics play an important role in quality control testing and reliability, where it is necessary predict the failure of a future item based on the times of a few early failure. These predictors are often based on moments of order statistics.

Proposition 2.3. Let Y_1, Y_2, \dots, Y_n be iid random variable such that Y_j follows a LECEG distribution for $j = 1, 2, \dots, n$. The pdf of the k -th order statistic, say $Y_{k:n}$, is given by (for $y > 0$)

$$f_{k:n}(y) = \frac{(1-p)^k \alpha \theta^{\alpha k} \lambda (1 - e^{-\lambda y})^{\alpha k - 1} e^{-\lambda y}}{B(k, n-k+1) (e^{-\lambda y} (1-\theta) + \theta)^{\alpha k + 1}} \left(1 - (1-p) \left(\frac{\theta (1 - e^{-\lambda y})}{e^{-\lambda y} (1-\theta) + \theta} \right)^\alpha \right)^{n-k},$$

where $B(k, n-k+1) = \frac{(n-k)!(k-1)!}{n!}$.

Proof. Considering a random sample of size n from the LECEG distribution. It is well known that

$$f_{k:n}(y) = \frac{1}{B(k, n-k+1)} f(y) (F(y))^{k-1} [S(y)]^{n-k}, \tag{2.7}$$

where $B(k, n - k + 1) = \frac{(n - k)!(k - 1)!}{n!}$.

Using (2.7), we have,

$$\begin{aligned} f_{k:n}(y) &= \frac{1}{B(k, n - k + 1)} f(y)(F(y))^{k-1} (S(y))^{n-k} \\ &= \frac{(1 - p)\alpha}{B(k, n - k + 1)(1 - e^{-\lambda y})} \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha \left(\frac{\lambda e^{-\lambda y}}{e^{-\lambda y}(1 - \theta) + \theta} \right) \\ &\quad \times \left((1 - p) \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha \right)^{k-1} \left(1 - (1 - p) \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha \right)^{n-k} \\ &= \frac{(1 - p)^k \alpha \theta^{\alpha k} \lambda (1 - e^{-\lambda y})^{\alpha k - 1} e^{-\lambda y}}{B(k, n - k + 1)(e^{-\lambda y}(1 - \theta) + \theta)^{\alpha k + 1}} \left(1 - (1 - p) \left(\frac{\theta(1 - e^{-\lambda y})}{e^{-\lambda y}(1 - \theta) + \theta} \right)^\alpha \right)^{n-k}. \end{aligned}$$

□

Proposition 2.4. *The characteristic function of the k-th order statistic from a random variable Y with LECEG distribution is given by*

$$\begin{aligned} \Phi_{Y_{k:n}}(t) &= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{(-n + k)_r}{r!} \frac{(1 - p)^{r+k} \theta^{\alpha(k+r)} \alpha}{B(k, n - k + 1)} \\ &\quad \times \left(-\alpha(k + r) - 1s \right) \frac{(\theta - 1)^s \Gamma(\alpha(k + r) + s)}{\Gamma(\alpha(k + r) + 1 + s \frac{-iy}{\lambda})} \Gamma\left(\frac{-iy}{\lambda} + 1\right), \end{aligned} \tag{2.8}$$

where $B(k, n - k + 1) = \frac{(n - k)!(k - 1)!}{n!}$.

Proof. Using the known formula $\Phi_{Y_{k:n}}(t) = \int_0^\infty e^{ity} f_{k:n}(t) dt$ and (7.1) from Appendix,

$$\begin{aligned} \Phi_{Y_{k:n}}(t) &= \frac{(1 - p)^k \alpha \theta^{\alpha k} \lambda}{B(k, n - k + 1)} \int_0^\infty \frac{e^{ity - \lambda} \left((1 - e^{-\lambda y})^{\alpha k - 1} \right)}{(e^{-\lambda y}(1 - \theta) + \theta)^{\alpha k + 1}} \left(1 - (1 - p) \left(\frac{\theta(1 - e^{-\lambda t})}{e^{-\lambda t}(1 - \theta) + \theta} \right)^\alpha \right)^{n-k} dt \\ &= \frac{(1 - p)^{k+r} \alpha \theta^{\alpha k + \alpha r}}{B(k, n - k + 1)} \sum_{r=0}^{\infty} \frac{(-n + k)_r}{r!} \int_0^1 (1 - u)^{\frac{-iy}{\lambda}} u^{\alpha k + \alpha r - 1} (1 - u(1 - \theta))^{-\alpha k - \alpha r - 1} du \\ &= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{(-n + k)_r}{r!} \frac{(1 - p)^{r+k} \theta^{\alpha r + \alpha k} \alpha}{B(k, n - k + 1)} \binom{-\alpha k - \alpha r - 1}{s} \\ &\quad \times \frac{(\theta - 1)^s \Gamma(\alpha k + \alpha r + s)}{\Gamma(\alpha k + \alpha r + 1 + s \frac{-iy}{\lambda})} \Gamma\left(\frac{-iy}{\lambda} + 1\right), \end{aligned}$$

where $B(k, n - k + 1) = \frac{(n - k)!(k - 1)!}{n!}$.

□

Proposition 2.5. *The mean of k -th order statistic from a LECEG distribution is given by*

$$E_{k:n}(Y) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{1}{A} (1-p)^{k+r} s \alpha (-1+\theta)^s \theta^{\alpha(k+r)} \Gamma(1+n) \times \Gamma(-\alpha(k+r)) \Gamma(1+s+\alpha(k+r)) (k-n)_r \psi^0(1+\alpha(k+r)),$$

where $A = \lambda \Gamma(k) \Gamma(1+r) \Gamma(1-k+n) \Gamma(1+s) \Gamma(-s-\alpha(k+r)) \Gamma(1+\alpha(k+r))$.

Proof. Considering that $E(Y_{k:n}) = \Phi_{Y_{k:n}}(0)' / i^r$, we have the first moment. □

In reliability, the ratio of two consecutive moments of reversed residual life characterize the distribution uniquely. The reversed failure rate function is given by

$$h(y) = \frac{f(y)}{F(y)} = \frac{\alpha}{(1-e^{-\lambda y})} \left(\frac{\lambda e^{-\lambda y}}{e^{-\lambda y}(1-\theta) + \theta} \right).$$

3 INFERENCE

In this section we consider maximum likelihood estimation (MLE). Assuming the lifetimes Y_i , $i = 1, \dots, n$ from the LECEG distribution independently and identically distributed and independent from the censoring mechanism c_i , $i = 1, \dots, n$, the maximum likelihood estimates (MLEs) of the parameters are obtained by direct maximization of the log-likelihood function given by,

$$\begin{aligned} \ell(\alpha, \lambda, \theta, p) &= \sum_{i=1}^n c_i \left[\log \left(\frac{(1-p)\alpha}{(1-e^{-\lambda y_i})} \left(\frac{\theta(1-e^{-\lambda y_i})}{e^{-\lambda y_i}(1-\theta) + \theta} \right)^\alpha \left(\frac{\lambda e^{-\lambda y_i}}{e^{-\lambda y_i}(1-\theta) + \theta} \right) \right) \right] \\ &+ \sum_{i=1}^n (1-c_i) \log \left(1 - (1-p) \left(\frac{\theta(1-e^{-\lambda y_i})}{e^{-\lambda y_i}(1-\theta) + \theta} \right)^\alpha \right) \\ &= \sum_{i=1}^n c_i \log((1-p)\alpha\lambda\theta^\alpha) - \lambda \sum_{i=1}^n c_i y_i + \sum_{i=1}^n c_i (\alpha-1) \log(1-e^{-\lambda y_i}) \\ &- \sum_{i=1}^n c_i (\alpha+1) \log(e^{-\lambda y_i}(1-\theta) + \theta) + \sum_{i=1}^n (1-c_i) \\ &\times \log \left(1 - (1-p) \left(\frac{\theta(1-e^{-\lambda y_i})}{e^{-\lambda y_i}(1-\theta) + \theta} \right)^\alpha \right), \end{aligned} \tag{3.1}$$

where c_i is a censoring indicator which is equal to 0 or 1, if the data is censored or observed, respectively.

The advantage of the MLE procedure is that it runs immediately using existing statistical packages. We have considered the *optim* routine of the R (Team, 2008). Large-sample inference for

the parameters are based on the *MLEs* and their estimated standard errors in the asymptotic normality.

The asymptotic normality is also useful for testing goodness of fit of the parameters in the distribution, also it is useful for comparing the LECEG distribution with some of its special sub-models, the long-term complementary exponential geometric (LCEG) and the long-term exponential (LE) distributions, via the likelihood ratio (LR) statistic.

4 SIMULATION STUDY

In this section is realized a simulation study to assess the asymptotic performance of MLEs in the LECEG model. Was generated lifetimes samples from the LECEG distribution with parameters values $\alpha = 1.3$, $\theta = 0.9$, $\lambda = 1.1$ and a cured fraction $p = 0.10$ through rejection algorithm using the weibull distribution with scale parameter 1.2 and shape 1.1 as auxiliary function. Also the samples was generated with censorship levels of 20%, 30% and 40% for sample sizes of $n = 30, 50, 100, 200$ and 500. For each sample size, we conducted 1000 simulations and then calculated the average bias (AB) and the mean square error (MSE) of the MLEs. Table 1 shows the bias and MSE of the MLEs for sample sizes $n = 30, 50, 100, 200$ and 500. From Table 1 was observed that the AB and MSE are close to zero and the MSEs decrease as sample size increases and both AB and MSE increases according to censorship level increases in the sample. Similar results are observed for others combinations of α, λ, θ and p .

5 THE LECEG REGRESSION MODEL

In this section the LECEG is presented when the location parameter is affected by covariates. Models that the location parameter is affected by covariates are considered when the survival time is affected by characteristics (covariates) of the individual.

Considering covariates for the location parameter λ in the LECEG model, the survival function (2.2) for the individual $y_i, i = 1, \dots, n$ is given by

$$S(y_i|\beta, \alpha, \theta, p) = 1 - (1 - p) \left(\frac{\theta(1 - e^{-\lambda_i y_i})}{e^{-\lambda_i y_i}(1 - \theta) + \theta} \right)^\alpha, \tag{5.1}$$

where $\lambda_i = e^{x_i' \beta}$, $x_i' = (1, x_{i1}, \dots, x_{ip})$ is the covariate vector and $\beta' = (\beta_0, \dots, \beta_p)$ is the respective coefficient vector.

From (3.1), the log-likelihood is given by

$$\begin{aligned} \ell(\alpha, \beta, \theta, p|y, x, c) &= \sum_{i=1}^n c_i \log((1 - p)\alpha\theta^\alpha \lambda_i) + (\alpha - 1) \sum_{i=1}^n c_i \log(1 - e^{-\lambda_i y_i}) \\ &\quad - \sum_{i=1}^n c_i \lambda_i y_i - (1 + \alpha) \sum_{i=1}^n c_i \log(e^{-\lambda_i y_i}(1 - \theta) + \theta) \\ &\quad + \sum_{i=1}^n (1 - c_i) \log \left(1 - (1 - p) \left(\frac{\theta(1 - e^{-\lambda_i y_i})}{e^{-\lambda_i y_i}(1 - \theta) + \theta} \right)^\alpha \right), \end{aligned}$$

Table 1: Results from simulations.

		Scenario					
		$\alpha = 1.3$		$\lambda = 1.1$		$\theta = 0.9$	
		20%		30%		40%	
n	Parameters	AB	MSE	AB	MSE	AB	MSE
30	α	-0.008	0.252	-0.013	0.297	-0.037	0.458
	λ	-0.091	0.192	-0.061	0.254	-0.100	1.461
	θ	0.201	0.162	0.229	0.184	0.256	0.209
	p	0.090	0.008	0.076	0.008	0.054	0.146
50	α	0.023	0.138	0.018	0.165	0.021	0.181
	λ	0.005	0.099	0.077	0.120	0.128	0.170
	θ	0.170	0.127	0.185	0.141	0.210	0.162
	p	0.093	0.009	0.086	0.008	0.073	0.123
100	α	0.039	0.066	0.047	0.071	0.054	0.081
	λ	0.087	0.045	0.172	0.073	0.238	0.127
	θ	0.127	0.087	0.140	0.095	0.164	0.116
	p	0.096	0.009	0.091	0.008	0.081	0.096
200	α	0.037	0.032	0.044	0.035	0.052	0.039
	λ	0.141	0.037	0.230	0.072	0.317	0.123
	θ	0.078	0.051	0.091	0.059	0.106	0.068
	p	0.098	0.009	0.095	0.009	0.088	0.065
500	α	0.025	0.014	0.030	0.015	0.035	0.016
	λ	0.173	0.036	0.268	0.078	0.359	0.137
	θ	0.037	0.024	0.039	0.028	0.049	0.033
	p	0.098	0.009	0.099	0.009	0.093	0.039

*AB: Average Bias – *MSE: Mean Square Error

where $\lambda_i = e^{x_i' \beta}$, $x' = (1, x_{i1}, \dots, x_{ip})$ is the covariate vector and $\beta' = (\beta_0, \dots, \beta_p)$ is the respective coefficient vector.

In modeling, to determine which potential covariates are influencing the response variable, processes of covariate selection such forward and backward stepwise are commonly used. Based on the covariate selection process described in Collet (1994), we consider the following steps:

Step 1: Fit the model with only the intercept and record the log likelihood, then adjust all possible models with one covariate and perform the likelihood ratio test to see which were significant adjustments. If we have more than one model remain significantly with the one that has the lowest $-l(\cdot)$ and AIC, BIC criteria.

Step 2: Fit the model with two covariates considering the covariate model selected in Step 1, and noting the one with the lowest $-l(\cdot)$ and AIC, BIC criteria.

Step 3 and so on: Fit models with the possible addition of another covariate always considering the covariates selected in the previous step until there remain no further covariate, remembering to stay with the one with the lowest $-\ell(\cdot)$ and AIC, BIC criteria.

6 APPLICATIONS

In this section, we compare the LECEG distribution with its particular case (the LCEG distribution and LE distribution), as well as with the long-term exponentiated gumbel (LEGU) and long-term Lindley (LLI) distribution, on three datasets.

LEGU is the distribution from (Lindley, 1958) added the parameter p of long-term and LLI the distribution from (Nadarajah & Kotz, 2006) with the p long-term parameter added.

The datasets are related to the medical field, which is related to the genesis and establishment of the proposed model. The main idea is show the applicability of the new distribution and the direct possibility of choosing between it or its particular case, as well as its competitiveness in terms of fitting related to an usual survival distribution.

As our first dataset, we consider 25 lifetimes with approximately 32% of censoring from P. Allison (1995). The second dataset, consists of 40 lifetimes of patients undergoing treatment with 7.5% of censoring Prentice (1973).

Firstly, in order to verify the shape of the failure rate function, we follow a standard graphical methodology for data analysis, the total time on test (TTT) plot, described by R. Maller & X. Zhou (1983). According to Aarset (1987), in its empirical version the TTT plot is given by $G(r/n) = [(\sum_{i=1}^r Y_{i:n}) - (n - r)Y_{r:n}] / (\sum_{i=1}^n Y_{i:n})$, where $r = 1, \dots, n$ and $Y_{i:n}$ represent the order statistics of the sample. It has been shown that the failure function is increasing (decreasing) if the TTT plot is concave (convex). Figure 3 show the TTT plot for the considered datasets, implying in decreasing failure rate functions. Therefore was fitted the five distributions for the datasets, the LECEG, LCEG, LE, LEGU and LLI. Table 2 shows the MLEs and their variances in parentheses for the five fitted distributions.

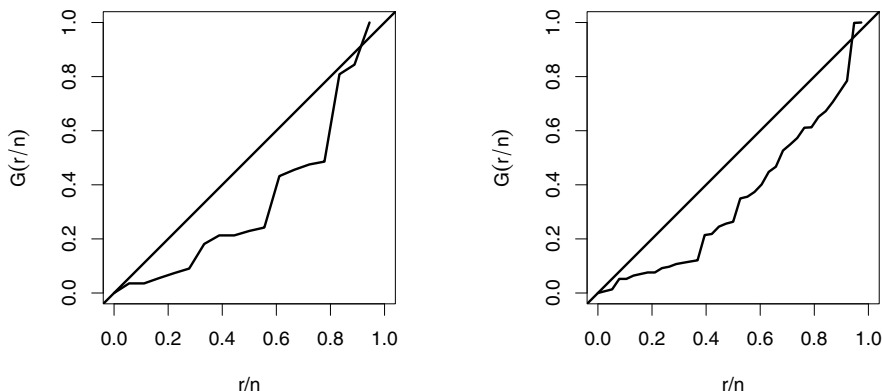


Figure 3: TTT plots for the datasets P. Allison (1995) (left panel) and Prentice (1973) (right panel).

Table 2: MLEs of the fitted distributions for the datasets P. Allison (1995) and Prentice (1973).

Dataset	Dist	α	λ	θ
P. Allison (1995)	LECEG	0.4776(0.0375)	$6.0 \times 10^{-4}(5.9 \times 10^{-7})$	0.6910(0.0010)
	LCEG	–	0.0010(2.6×10^{-7})	0.9990(0.0001)
	LE	–	0.0011(2.5×10^{-7})	–
	LEGU	0.0647(0.0002)	–	–
	LLI	–	0.0030(4.5×10^{-7})	–
Prentice (1973)	LECEG	0.6300(0.0154)	0.0040(1.3×10^{-6})	0.9990(0.0001)
	LCEG	–	0.0060(1.1×10^{-6})	0.9990(1.3×10^{-5})
	LE	–	0.0063(1.1×10^{-6})	–
	LEGU	0.1720(0.0009)	–	–
	LLI	–	0.0120(2.2×10^{-6})	–

Dataset	Dist	σ	μ	p
P. Allison (1995)	LECEG	–	–	0.0005(0.0001)
	LCEG	–	–	0.0230(0.0283)
	LE	–	–	0.0240(0.0272)
	LEGU	62.470(0.0238)	–30.375(930.51)	$8.7 \times 10^{-7}(5.1 \times 10^{-8})$
	LLI	–	–	0.1340(0.0093)
Prentice (1973)	LECEG	–	–	$1.7 \times 10^{-5}(5.1 \times 10^{-7})$
	LCEG	–	–	$1.1 \times 10^{-5}(3.3 \times 10^{-6})$
	LE	–	–	$4.5 \times 10^{-5}(1.3 \times 10^{-6})$
	LEGU	27.177(0.0051)	2.812(110.58)	$1.0 \times 10^{-4}(3.6 \times 10^{-6})$
	LLI	–	–	$2.3 \times 10^{-5}(6.7 \times 10^{-7})$

We compare the fitted distribution by the $-\ell(\hat{\psi})$ (where $\hat{\psi}' = (\hat{\alpha}, \hat{\lambda}, \hat{\theta}, \hat{p})$), the AIC (Akaike’s information criterion, $-2l(\hat{\psi}) + 2k$, where k is the number of parameters in the model) and BIC (Schawartz’s Bayesian information criterion, $-2l(\hat{\psi}) + 2 \log(n)$, where n is the size sample). The preferred model is the one with the smaller value on each criterion. Table 3 shows the values of the $-\ell(\hat{\psi})$, AIC and BIC, the Kolmogorov-Sminorv (KS) statistic and its p -value, jointly with the criterion values of the long-term exponentiated gumbel (LEGU) distribution and the long-term Lindley (LLI) distribution. The LECEG distribution outperforms its concurrent distributions in all considered criteria for the two datasets. These results are corroborated by the plots in Figure 4, which shows the fitted survival functions for all fitted distributions and the empirical Kaplan-Meier survival function.

Table 3: Criterion values and the *KS*-Statistics and its *p*-values to AIC, BIC the datasets from P. Allison (1995) and from Prentice (1973).

Dataset	Dist	$-\ell(\hat{\psi})$	AIC	BIC	KS-statistic	p-value
P. Allison	LECEG	128.5016	265.0032	269.8787	0.1482	0.6414
	LCEG	132.6698	271.3396	274.9962	0.2584	0.0709
	LE	132.6690	269.3380	271.7758	0.2582	0.0712
	LEGU	134.7562	277.5123	282.3878	0.2495	0.0887
	LLI	148.0361	300.0722	302.51	0.3405	0.0060
Prentice 1973	LECEG	220.9193	449.8385	456.5941	0.1116	0.7006
	LCEG	223.9223	453.8447	458.9113	0.2088	0.0610
	LE	223.9231	451.8461	455.2239	0.2088	0.0609
	LEGU	229.5509	467.1018	473.8573	0.1942	0.0976
	LLI	244.4418	492.8836	496.2614	0.3164	0.0006

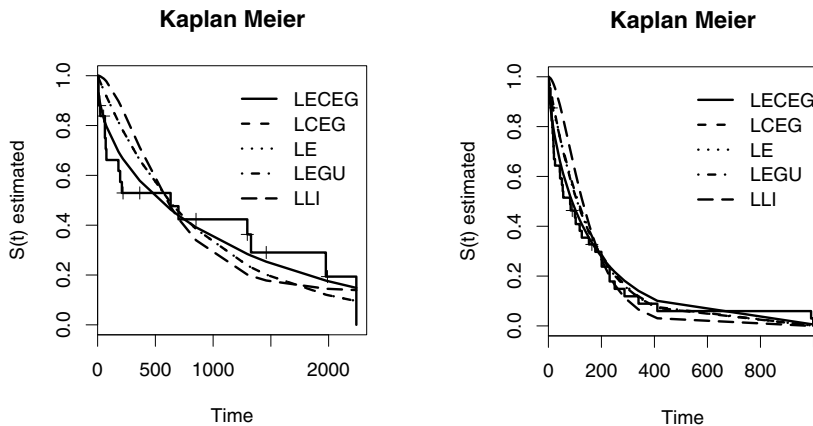


Figure 4: Kaplan Meier curve with estimated survival functions obtained via LECEG, LCEG and LE distributions for P. Allison (1995) and Prentice (1973).

As our third dataset, consider the data extracted from Kalbfleisch & Prentice (1980), An acknowledgement goes to the book Kalbfleisch & Prentice (1980). The dataset considered consist of the following variables: Acceptance years(X_1), years (X_2), status(alive or dead)(c), transplant(yes or no)(X_3), surgery(if the patient has gone through some kind of surgery before)(X_4) and the lifetime(Y). Table 4 shows the coefficient and/or interaction term, with their standard deviations in parenthesis, obtained from the MLE procedure using the variable selection describes in Section 5. Also the $-l(\cdot)$, AIC and BIC criterion values are presented. From Table 4 the LECEG model was considered as the best fitted model.

Table 4: The MLEs values to the parameters of the LECEG, LCEG and LE and their standard deviation in parenthesis, and the $-\ell(\cdot)$, AIC and BIC criterion values.

	LECEG	Term relation	LCEG	Term relation	LE	Term relation
α	0.828 (0.143)					
θ	0.697 (0.231)		0.999 (1.2×10^{-10})			
p	1.7×10^{-5} (3×10^{-6})		0.080 (0.047)		0.076 (0.054)	
β_0	8.71 (7.385)	Intercept	-5.820 (0.784)	Intercept	13.183 (6.492)	Intercept
β_1	-32.339 (9.147)	X_3	-2.283 (0.298)	X_3	-25.988 (9.826)	X_3
β_2	-1.067 (0.359)	X_4	0.046 (0.016)	X_2	0.054 (0.015)	X_2
β_3	-0.178 (0.105)	X_1	-63.942 (30.293)	X_4	-0.275 (0.093)	X_1
β_4	0.441 (0.130)	$X_1 X_3$	0.879 (0.421)	$X_1 X_4$	-1.135 (0.483)	X_4
β_5					0.336 (0.140)	$X_1 X_3$
$-\ell(\cdot)$	437.613		474.210		471.892	
AIC	891.226		962.421		957.785	
BIC	912.304		980.8641		976.228	

According Cox & Snell (1968) if the model fits the data well then the true cumulative hazard function conditional on the covariate vector has an exponential distribution with a hazard rate of one, i.e, we can verify if the residuals have an exponential distribution with parameter one. Figure 5 shows the Kaplan Meier curve of the Cox-Snell residuals for the fitted distributions superimposed by an exponential distribution with a hazard rate of one, indicating a reasonable fit for the LECEG model over the LCEG and LE models.

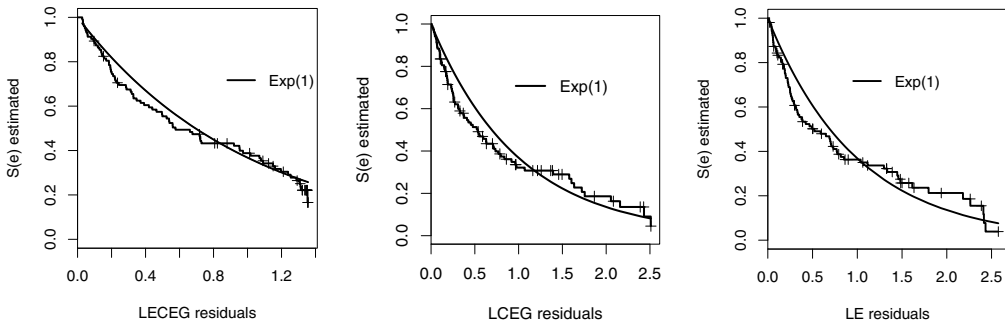


Figure 5: Kaplan Meier curves of the Cox-Snell residuals for the fitted distributions superimposed by an exponential distribution with a hazard rate of one.

Considering the LECEG model with $\lambda_i = \exp(8.71 - 32.33 * X_3 - 1.06 * X_4 - 0.17 * X_1 + 0.44 * X_1 * X_3)$, the average risk of dying for those patient who did not undergo surgery is 2.89 times of the average risk of dying for those patient already underwent surgery.

7 CONCLUDING REMARKS

In this paper we provided the LECEG distribution that is an extension of the ECEG distribution with a long term parameter and belong in the latent complementary risks scenarios, i.e., where the lifetime associated with a particular risk is not observable but the maximum lifetime among all risks. The properties of the proposed distribution are discussed, including its probability density function, the quantile, the survival and failure rate functions, the characteristic function and r-th order statistics. Maximum likelihood inference is implemented straightforwardly. The practical importance of the LECEG distribution was demonstrated in three applications.

ACKNOWLEDGMENT

We thanks the Editorial Boarding and the Anonymous Referees for their comments and criticisms in earlier versions of the manuscript. The research is partially supported by CNPq, CAPES and FAPESP.

Appendix A

The following binomial series expansion is used in the present paper

$$(1 - x)^{-s} = \sum_{r=0}^{\infty} \frac{(s)_r}{r!} x^r, \tag{7.1}$$

where $(r)_k = r(r + 1)...(r + k - 1)$ is a Pochhammer symbol, $(-r)_k = (-1)^k (r - k + 1)_k$ and if $|x| < 1$ the series converge.

Through the paper we use the following relationship:

$$\int_0^1 z^{p-1} (1 - z)^{n-1} (1 + bz^m)^l dz = \Gamma(n) \sum_{k=0}^{\infty} \binom{l}{k} \frac{b^k \Gamma(p + km)}{\Gamma(p + n + km)}. \tag{7.2}$$

RESUMO. Neste trabalho, uma nova distribuição de longa duração derivada da distribuição geométrica exponencial complementar exponenciada é apresentada. A distribuição LECEG acomoda taxa de risco decrescente e unimodal, presente em um cenário de causas complementar latente, em que apenas o tempo máximo de vida entre todas as causas é observado. Obtivemos as funções de densidade, quantis, sobrevivência e taxa de falha para a distribuição proposta, bem como algumas propriedades, tais como a função característica, média, variância e estatísticas r-ésima ordem. A estimativa é baseada na abordagem de máxima verossimilhança. Um estudo de simulação é realizado a fim de avaliar o desempenho das estimativas de máxima verossimilhança. A importância prática da nova distribuição foi demonstrada em três conjuntos de dados reais.

Palavras-chave: distribuição geométrica exponencial complementar exponenciada, termo de longa duração, riscos complementares.

REFERENCES

- [1] K. Adamidis & S. Loukas. A lifetime distribution with decreasing failure rate. *Statistics Probability Letters*, **39** (1998), 35–42.
- [2] M. Chahkandi & M. Ganjali. On some lifetime distributions with decreasing failure rate. *Computational Statistics & Data Analysis*, **53**(1) (2009), 4433–4440.
- [3] F. Louzada, M. Roman & V.G. Cancho. The complementary exponential geometric distribution: Model, properties, and a comparison with its counterpart. *Computational Statistics & Data Analysis*, **55**(8) (2011), 2516–2524.
- [4] A.L. Morais & W. Barreto-Souza. A compound class of weibull and power series distributions. *Computational Statistics & Data Analysis*, **55**(1) (2010), 1410–1425.
- [5] V.G. Cancho, F. Louzada-Neto & G.D.C. Barriga. The poisson-exponential lifetime distribution. *Computational Statistics & Data Analysis*, **55**(1) (2011), 677–686.
- [6] F. Hemmati, E. Khorram & S. Rezakhan. A new three-parameter ageing distribution. *Journal of Statistical Planning & Inference*, **141**(1) (2011), 2266–2275.
- [7] H.S. Bakouch, B.M. Al-Zahrani, A.A. Al-Shomrani, V.A.A. Marchi & F. Louzada. An extended lindley distribution. *Journal of the Korean Statistical Society*, (2011).
- [8] F. Louzada, V. Marchi & J. Carpenter. The complementary exponentiated exponential geometric lifetime distribution. *Journal of Probability and Statistics*, **12**, (2013).
- [9] Louzada-Neto. Poly-hazard regression models for lifetime data. *Biometrics*, **55** (1999), 1121–1125.
- [10] C.Y. Yamachi, M. Roman, F. Louzada, Maria A.P. Franco & V.G. Cancho. The Exponentiated Complementary Exponential Geometric Distribution. *Journal of Modern Mathematics Frontier*, **2**(1), (2013).
- [11] R. Maller & X. Zhou. *Survival Analysis with Long-Term Survivors*. John Wiley and Sons Chichester, (1996).
- [12] J.W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, (1949), 15–44.
- [13] J. Berkson & R.P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47** (1952), 501–515.
- [14] V.T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**(4) (1982), 1041–1046.
- [15] A.R. Maller & S. Zhou. Testing for the presence of immune or cured individuals in censored survival data. *Biometrics*, **51** (1995), 1197–1205.
- [16] M.H. Chen & J.G. Ibrahim. Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, **57** (2001), 43–52.
- [17] V.G. Cancho, E.M.M. Ortega & H. Bolfarine. The log-exponentiated-weibull regression models with cure rate: Local influence and residual analysis. *Journal of Data Science*, **7** (2009), 433–458.

- [18] G.S.C. Perdoná & F. Louzada. A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics*, **38** (2011), 1395–1405.
- [19] M. Roman, F. Louzada, V.G. Cancho & J.G. Leite. A new long-term survival distribution for cancer data. *Journal of Data Science*, **10** (2012), 241–258.
- [20] Mathematica Software. Wolfram mathematica: Technical computing software. *Wolfram Research, Inc.*, (2010).
- [21] R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, (2008).
- [22] D. Collet. *Modelling Survival Data in Medical Research*. Chapman and Hall, (1994).
- [23] D.V. Lindley. Fiducial distributions and bayes theorem. *Journal of the Royal Statistical Society*, **20**(1) (1958), 102–107.
- [24] S. Nadarajah & S. Kotz. The exponentiated type distributions. *Acta Applicandae Mathematicae*, **92**(2) (2006), 97–111.
- [25] P. Allison. *Survival Analysis Using SAS a practical guide*. SAS, (1995).
- [26] R.L. Prentice. Exponential survivals with censoring and explanatory variable. *Biometrika*, **60** (1973), 279–288.
- [27] R. Maller & X. Zhou. *Graphical Methods for Data Analysis*. Chapman and Hall, (1983).
- [28] M.V. Aarset. How to identify a bathtub hazard rate. *Statistical Methods and Applications*, **36**(1) (1987), 106–108.
- [29] J.D. Kalbfleisch & R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley & Sons, (1980).
- [30] D.R. Cox & E.J. Snell. A general definition of residuals. *Royal Statistical Society*, **30**(2) (1968), 248–275.