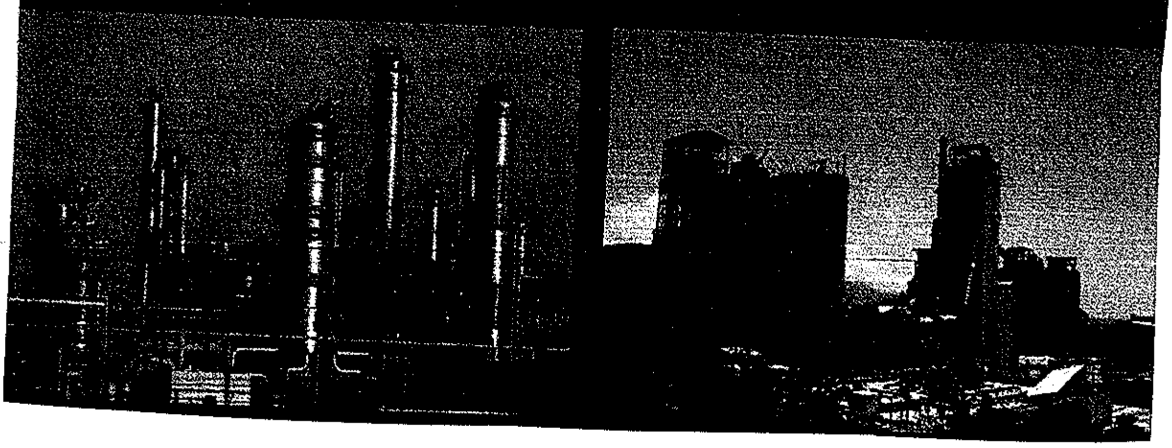
Journal of Maintenance Engineering

Volume 1 - August 2016

Dr Jyoti K. Sinha
Editor-in-Chief



Shield Crest

ISBN 978-1-911090-39-7

MMXVI

A CIP catalogue record for this book
is available from the British Library

Published by
ShieldCrest Publishing Limited
Aylesbury, Buckinghamshire,
HP22 5RR
England

Application of Mixture Models to Survival Data

Sílvia Madeira[1], Paulo Infante[2], Filipe Didelet[3]

[1]PhD in Mathematics, IIFA of University of Évora, Portugal.

[2]CIMA/IIFA,DMAT/ECT of University of Évora, Portugal.

[3]Mechanical Department of EST/IPS of Setúbal, Portugal.

Email: sparreira@gmail.com; pinfante@uevora.pt; filipe.didelet@estsetubal.ips.pt

Abstract Survival models are being widely applied to the engineering field to model time-to-event data once censored data is here a common issue. Using parametric models or not, for the case of heterogeneous data, they may not always represent a good fit. The present study relays on critical pumps survival data where traditional parametric regression might be improved in order to obtain better approaches. Considering censored data and using an empiric method to split the data into two subgroups to give the possibility to fit separated models to our censored data, we've mixture two distinct distributions according a mixture-models approach. We have concluded that it is a good method to fit data which does not fit to a usual parametric distribution and achieve reliable parameters. A constant cumulative hazard rate policy was used as well to exemplify optimum inspection times using the mixture-model, which could be an added value, when comparing with the actual maintenance policies, to check whether changes should be introduced or not.

Key words: Reliability, Mixture-models, Censored data, Survival models, Inspection Policies.

1.0 Introduction

The present study was made on behalf of Galp Energia Company, where much material was provided to apply our models as it is an oil and gas company and is now underneath the scope of Reliability Centered Maintenance (RCM). Following the tendency of efficiency improvement, Galp has been doing an effort to achieve higher reliability parameters. Our study has tried to meet their goal and contribute to a reliable and feasible assets analysis. [3] has already covered the subject which we'll improve. The load pumps for the FCC reactor work on a two-out-of-three scheme as they are critical equipment. 100% reliability for the system is required and, for now, it has been achieved, however, a better understanding on their behavior is enhanced and the present study attends as well to check both reliability times and inspection policies. Survival analysis was applied, grouping time-to-event data in sets of two equipment once they work two by two. Thus, three groups for possible combinations where made and survival analysis was applied as shown in Figure 1. First tested with non-parametric approaches as [4], and then in section 2 with Accelerated Failure Time models (as described in [5]), the three groups were subject of study to check wheatear the reliability curves were distinct or not. As equal pumps, they should have similar behavior, and apparently they have not, however, statistical hypothesis helps us to make a decision. In this paper we'll discuss if differences between survival curves are significant and if we should reject the possibility to use data from the three equipment in a global model. Mixture-models, in section 3, have revealed to be a good approach on heterogeneous data, and so, better models can be used for instance, to apply inspection policies with constant cumulative hazard rate, as discussed in section 4.

2.0 Parametric and Non-parametric Approaches

Let S(t) be the probability that a member from a given population will have a lifetime exceeding t. For a sample of size *N* from the list of observations, let the observed times until the failure of the *N* sample observations be $t_1 \le t_2 \le t_3 \dots \le$. Corresponding to each t_i is n_i, the number "at risk" just prior to time t_i, and d_i the number of failures at time t_i. The Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of *S*(*t*). It is a product of the form:

$$S(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}. \qquad (1)$$

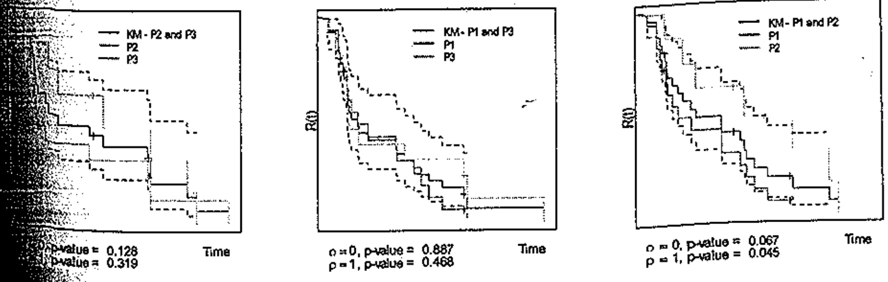


Figure 1 Models for the three groups of pumps

the non-parametric approach, we've found that, except for the third group, 1 and 2, the statistic test indicates that we clearly should not reject the null hesis in which times could possibly be identical. Log-rank and Peto & Peto gave us p-values, for the comparison of the survival curves of the three of about 0.193 and 0.197, respectively. However, a parametric approach ch group was made and reliability inference is then possible. Figure 2 illustrates the tested parametric approaches with three distributions, Exponen Weibull and Loglogistic, using equations (2), (3) and (4) for the three groups W can see from Table 1 that exponential has the best approach according an AIC terion as well as for the three individual pumps already studied in [3].

Exponential $R(t) = \exp[-(t/\lambda)], \lambda > 0,$ (2)

Weibull $R(t) = \exp[-(t/\lambda)^{\rho}], \lambda, \rho > 0$ (3)

Loglogistic $R(t) = \frac{1}{\left[1+\left(\frac{t}{\lambda}\right)^{\rho}\right]}, \ \lambda, \rho > 0$ (4)

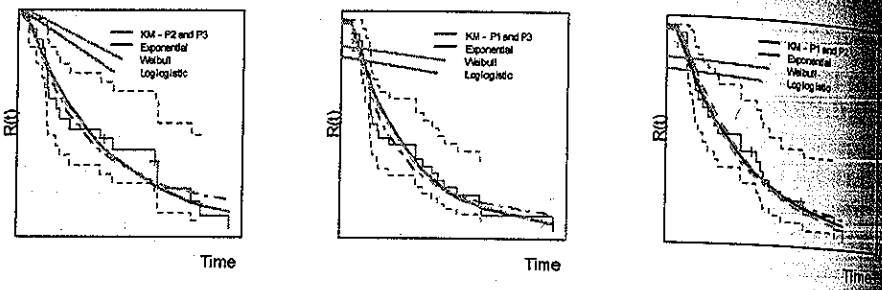


Figure 2 Parametric approaches for the three groups

Table 1 AIC criterion and parameters' estimators

Pumps			SE	LL(95%)	UL(95%)	AIC
			Exponential			
2/3	$\log\hat{\lambda}$	8.597	0.229	8.147	9.047	**366.672**
1/3	$\log\hat{\lambda}$	8.276	0.189	7.905	8.646	**521.433**
1/2	$\log\hat{\lambda}$	8.427	0.209	8.019	8.836	**435.658**
			Weibull			
2/3	$\log\hat{\lambda}$	8.592	0.239	8.123	9.060	**368.663**
	$\hat{\rho}$	0.982	0.828	-0.641	2.605	
1/3	$\log\hat{\lambda}$	8.263	0.202	7.867	8.659	**523.364**
	$\hat{\rho}$	0.962	0.861	-0.726	2.649	
1/2	$\log\hat{\lambda}$	8.459	0.185	8.096	8.823	**436.973**
	$\hat{\rho}$	1.153	0.846	-0.506	2.812	
			Loglogistic			
2/3	$\log\hat{\lambda}$	8.099	0.297	7.518	8.681	**371.388**
	$\hat{\rho}$	1.293	0.829	-0.331	2.917	
1/3	$\log\hat{\lambda}$	7.731	0:228	7.284	8.178	**523.956**
	$\hat{\rho}$	1.387	0.856	-0.291	3.064	
1/2	$\log\hat{\lambda}$	8.020	0.225	7.579	8.461	**438.946**
	$\hat{\rho}$	1.568	0.845	-0.089	3.224	

Is then arguable to compare the curves for the three groups with survival times of all three pumps and that we'll from now call *global model* (Figure 3). Visually speaking, we see that the groups are in between the confidence interval for the global model. Furthermore, as parameters does not differ as much from each other's, maybe for a rough approach, we shall not reject to use the global model when we have one of those three pairs of pumps working. The parametric study for the global model is discussed in Section 3.0.

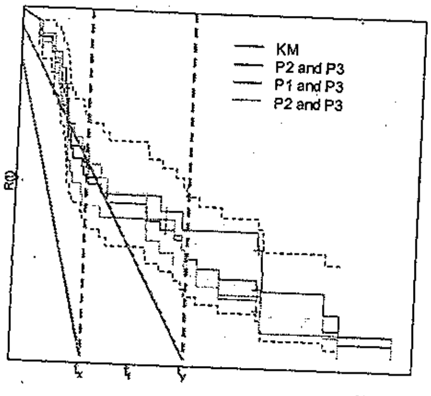


Figure 3 Comparison between the model for groups and the global model

3.0 Mixture-Models

Following the approach of [1], we've tried to adjust a mixture-model to our global model as it seems visually to have heterogeneous data. In our particular case, censored data hindered us to use the model as a whole and thus, we've split data into two subgroups and both were first modeled separately. An empiric split could be intuitively made for time t_i - which could probably be a point between t_x and t_y - where it seems to have an inflection point. However, a method to find an optimal cutting point (CP) in an analytical way still in progress. An alternative way to select the CP was using R software, whit which we've looped through all possible time values as candidate cutting points and applied them to the tested mixture distributions. Time values which have minimized MSE (10) and AIC (11) for the models were chosen to be tested as CP. Of course we must bear in mind that different CP's might be found for distinct used pairs of distributions. We found relevant to give consistency to the chosen CP, testing its goodness of fit as we will further see. Four mixture models were tested for several distributions (Exponential-Weibull, Exponential-Loglogistic, Weibull-Weibull and Weibull-Loglogistic) regarding equation (5) and according their respective density-functions. The maximum likelihood estimators of the parameters for the tested mixture-models are shown in Table 2 according the optimal CP found for each model. The value for the π parameter, which is the mixture weight for the distributions ($\pi \in (0,1)$) was estimated according an MSE approach optimizing the value for π to minimize the squared errors. As two CP candidates were found, t_A and t_B, AIC criterion held us to arguably select one of the models model with CP = t_1. Which one to choose might be debatable, so using the MSE criterion and a graphical visualization, a Weibull-Loglogistic model was chosen traduced by equation (6).

$$f_{XY}(t;\psi) = \pi f_X(t;\psi_X) + (1-\pi) f_Y(t;\psi_Y) \tag{5}$$

where $f_X(t;\psi_X)$ and $f_Y(t;\psi_Y)$ are the distributions of the mixture model with respective set of parameters ψ_X and ψ_Y. And so, for our chosen model, we'll have:

$$f_{XY}(t;\lambda,\rho) = \pi f_X(t;\lambda_X,\rho_X) + (1-\pi) f_Y(t;\lambda_Y,\rho_Y) \tag{6}$$

being $f_X(t;\lambda_X,\rho_X)$ and $f_Y(t;\lambda_Y,\rho_Y)$ the Weibull and Loglogistic distributions. From (6) we can easily derive equation (7), to achieve the reliability function (8) and (9):

$$R_{XY}(t;\lambda,\rho) = \int_t^{\infty} (\pi f_X(u;\lambda_X,\rho_X) + (1-\pi) f_Y(u;\lambda_Y,\rho_Y)) \quad du \tag{7}$$

Simplifying, $R_{XY}(t) = \pi R(t;\lambda_X,\rho_X) + (1-\pi) R(t;\lambda_Y,\rho_Y)$ (8)

$$R_{XY}(t) = \pi \exp\{(-t/\lambda_X)^{\rho_X}\} + (1-\pi)/(1+(t/\lambda_Y)^{\rho_Y}) \tag{9}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\big(S(t) - R_{XY}(t)\big)^2 \tag{10}$$

$$AIC = 2k - 2\ln\big(\mathcal{L}(\psi_{XY};t_i)\big) \tag{11}$$

with k the number of parameters and $\mathcal{L}$ the maximum value of the likelihood function.

Table 2 Estimated parameters for the tested mixture-models

Weibull-Loglogistic $CP = t_A$	**Exponential-Weibull** $CP = t_B$	**Weibull-Weibull** $CP = t_A$	**Exponential-Loglogistic** $CP = t_B$
Weibull $\log\hat{\lambda} = 7.182$ $\hat{\rho} = 2.106$ **Loglogistic** $\log\hat{\lambda} = 8.861$ $\hat{\rho} = 3.5733$	**Exponential** $\log\hat{\lambda} = 7.595$ **Weibull** $\log\hat{\lambda} = 9.198$ $\hat{\rho} = 3.721$	**Weibull** $\log\hat{\lambda} = 7.182$ $\hat{\rho} = 2.106$ **Weibull** $\log\hat{\lambda} = 9.057$ $\hat{\rho} = 2.563$	**Exponential** $\log\hat{\lambda} = 7.595$ **Loglogistic** $\log\hat{\lambda} = 9.053$ $\hat{\rho} = 5.449$
$\hat{\pi} = 0.508$	$\hat{\pi} = 0.651$	$\hat{\pi} = 0.509$	$\hat{\pi} = 0.651$
AIC= 706.777 **MSE=0.0007**	**AIC=709.586** **MSE=0.003**	**AIC=705.839** **MSE=0.0008**	**AIC=709.996** **MSE=0.003**

Reliability curves for the tested models are represented in Figure 4. We can see that a good improvement was achieved with this methodology. A range of values for variations of chosen CP (t_1) were tested and percentage of variation of the AIC, MSE and R(t) results are shown in Table 3. We can see that there are no relevant variations in AIC, MSE and Reliability values for the presented range. With the gathered information for the reliability function, we are now in conditions to use a wide range of tools to cook reliability information to held maintenance management.

Table 3 Impact (%) for CP variations

Range (t_A)	AIC (706.8)	MSE (<0.01)	R(t) (0.49)
$0.6t_A$	7.31	71.18	18.10
$0.7t_A$	4.46	50.9	16.13
$0.8t_A$	0.81	10.34	11.10
$0.9t_A$	~0	~0	8.91
$0.05t_A$	~0	~0	8.91
$1.05t_A$	~0	~0	8.91
$1.1t_A$	0.19	2.71	8.91
$1.2t_A$	1.11	1.01	7.93
$1.3t_A$	1.11	1.01	7.93
$1.4t_A$	1.11	1.01	7.93

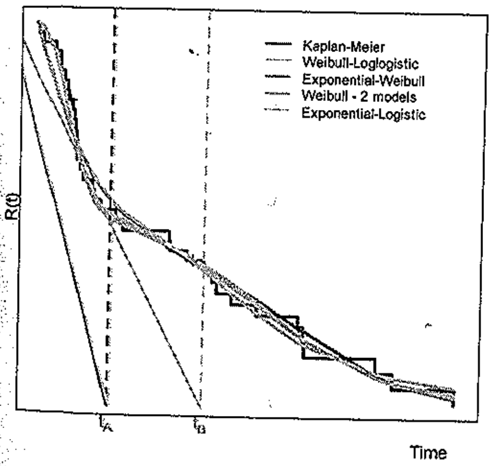


Figure 4 Mixture models for CP = t_A and t_B

4.0 Constant Hazard Inspection Policy

Finding the better fit for our survival times would make no sense if we didn't make something effective with it. We found interesting the approach of [2] following the work developed in [6] when determining the optimum points of inspection keeping the cumulative hazard constant. This is particular useful if we think in environments where the hazard ratio is not constant. A constant cumulative hazard H(t) means that we have.

$$H(t) = -\ln R(t) \tag{12}$$

$$H(t_{i+1}) - H(t_i) = \Delta H \tag{13}$$

$$R(t_i) = \exp(-i\Delta H) \tag{14}$$

thus

$$t_i = R^{-1}\{\exp(-i\Delta H)\}$$
$$t_0 = 0 \qquad (15)$$

As equation (6) has not a closed form in order to easy calculate its inverse function, a numeric method was used. In practice, if we fix the cumulative hazard according the relation obtained in [6] for perfect inspection context we'll have $\Delta H \cong \frac{P}{E(T)}$, where p is the periodic inspection period and E(T) is the mean time to failure. For our case, we know that the periodic time interval of inspection defined as being optimal, is P=535. If we have a $E(T)$=3571 we'll obtain ΔH=0.15 and it's now possible to know the intervention periods keeping that cumulative hazard. The interventions times are shown in Table 4. Figure 5 depicts four reliability functions derived from the mixture model, and we see that when the hazard rate is increasing, periods of intervention get smaller, and when it's decreasing they get larger. We'll have then a constant intervention periodicity schedule when a constant hazard rate is verified. We can see that when comparing it with the periodic policy, fewer interventions are made using this methodology. Further analysis would be interesting to do in a next work to compare it as well concerning costs.

Table 4 Periods of inspection keeping the cumulative hazard constant

i	t_i
1	1369
2	2834
3	4115
4	5103
5	5896
6	6579
7	7196
8	7775
9	8331
10	8874
11	9411
12	9947

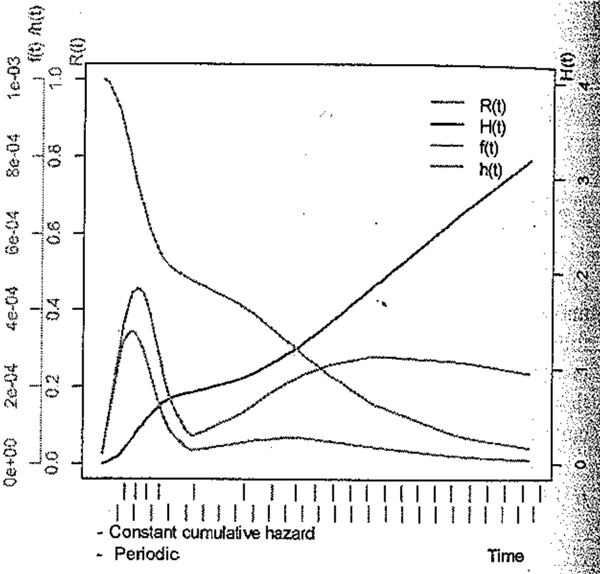


Figure 5 Reliability functions for the survival mixture model

5.0 Results

The obtained results show us that improvement for reliability curves is possible as reliable information is provided and a better approximation for real data is achieved. Mixture-models revealed to have the needed flexibility to approximate heterogeneous data and provide good parametric fits. However, there still is room for innovation concerning an ideal cutting point or cutting points, for time-to-event data and further investigation must keep ongoing so an accurate value can be obtained. In addition, much other information on maintenance periodicity can as well be obtained as comparisons between inspection policies. Interesting conclusions could certainly be obtained when comparing with a fixed periodic inspection scheme verifying if there is a cost reduction.

References

[1] Erisoglu, U., Erisoglu, M., and Erol, H. (2011). Non-parametric estimation in survival models. *International Journal of Computer, Automation, Control and Information Engineering*, (6):544–548.

[2] Rodrigues Dias, J., Infante, P. (2008). Control charts with predetermined sampling intervals.International Journal of Quality and Reliability Management, Vol. 25, N. 4, p. 423-435.

[3] Madeira, S., Infante, P., and Didelet, F. (2014). Using statistical models in industrial equipment. In *Proceedings of the 22nd European Congress and Expo on Maintenance and Asset Management*, pages 478–482.

[4] Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

[5] Rausand, M., Hoyland, A. (2004). *System Reliability Theory - Models, Statistical Methods and Applications*. Second Edition. Wiley

[6] Rodrigues Dias, J. (1987), *Systems inspection policies*, PhD thesis, University of Évora, Évora, in Portuguese.

Authors' Biography

Dr Sílvia Parreira Madeira

Dr. Sílvia Madeira was developing her PhD investigation work on Galp Energia, a Portuguese oil & gas company. She was part of Reliability department and has been involved in RCM procedures and prepared reliability analysis. She's accompanied by the Department of Mathematic of University of Évora as well as by the Department of Mechanic of the Polytechnic Institute of Setúbal.

Dr Paulo Infante

Dr. Paulo Infante is an Assistant Professor at Department of Mathematics of University of Évora and member of the Research Centre in Mathematics and Applications (CIMA-UE). Ph.D. in Mathematics, his current research interests are focused in the areas of Quality Control, Reliability, Survival Analysis, Design of Experiments, Generalized Linear Models and Demography. He has published several papers in several national and international journals. Currently he is adviser of several PhD and Masters theses.

Dr Filipe Didelet

Dr. Filipe Didelet is an Associate Professor at Technology College, Setubal Polytechnics. Hę works normally in the reliability prediction of thermal equipment by means of mathematical and functional models and in maintenance management methodologies.