# ON COPING WITH CENSORING IN SURVIVAL DATA USING FUZZY THEORY

Jaisankar. R[1*], Haripriya. H[2]

•

[1*] Professor, Department of Statistics, Bharathiar University,
Coimbatore 641 046, Tamil Nadu, INDIA
[2]Research Scholar, Department of Statistics, Bharathiar University,
Coimbatore 641 046, Tamil Nadu, INDIA
[1*] r_jaisankar@buc.edu.in, [2] haripriyah1795@gmail.com

**Abstract**

*Aim: The problem of censoring is the main concern in Survival Analysis which makes the analysis complex and may lead to unreliable results like less precise parameter/survival/hazard estimates, wider confidence intervals, reduced power to detect differences between groups, sometimes especially when the proportion of censored observations is high. In particular, if the censoring is informative that is reacted to the event, serious bias may be expected to present in the estimates. Omitting censored observations prices with loss of information and adding them is often embedded with a lack of information. It is proposed to implement fuzzy methodology for dealing with censored observations in Survival data so as to minimize their impact on the survival/hazard estimates. This paper suggests a modified approach based on the Fuzzy Theory, in which, the censored observations could be incorporated effectively. Methods: Many statistical procedures are advised to tackle censoring, in both parametric and non-parametric realms. However, their application mainly depend on the nature of censoring, whether right or left or interval censored. Even when their applied, still some lacuna may persist which could be completely eradicated when there is no censoring. Different types of fuzzy numbers are applied both to the censored and uncensored observations according to their nature and the results of the same were tested in case of the non-parametric procedure, the Kaplan-Meier estimator, using log rank test. Results: On the application of the proposed methodology one may obtain the survival probabilities which are reasonably calculated for even censored observations. The median survival time observed by the application of this methodology gives lesser value than the median obtained by the classical Kaplan eier method. The survival curve obtained by the proposed method is different from the survival curve of the classical Kaplan eier method and since the proposed methodology addresses all points of time and hence may give reliable estimates of survival probability. Conclusion: When compared with the classical Kaplan-Meier method, the proposal method gives survival probabilities which are slightly differ from the estimate obtained from non-fuzzy methodology. Since the survival probabilities can be calculated even for censored observations using fuzzy numbers, it is expected that the proposed modelling may be better than the classical Kaplan-Meier estimator.*

**Keywords:** Censoring, Non- Parametric Estimator, Vagueness, Fuzzy Sets, Fuzzification; Survival Time.

## 1. Introduction

Censoring is the primary problem that may persist in every survival data. The presence of censoring may pollute the results of the survival analysis and when the proportion of censored

observations is higher the results of the analysis would adversely be affected. The use of Non-parametric survival methods may be advised in such scenarios but even the estimates obtained from them also be distributed when too much censoring is present. Even though, some distributions like Weibull, Exponential, and Gompertz are preferred to mitigate with censoring, bias in the estimates could not be avoided significantly. Also, parametric models are capable of dealing with right censoring but left and interval censoring would be more complex. Similar is the case of semi-parametric survival models, violation of proportional hazard assumption may be exhibited when censoring occurs more frequently in one group and if censoring rates differ across covariates the estimates may become unreliable which would question the validity of the model fit. The presence of censoring leads to different consequences in non-parametric survival models when compared with parametric models. In addition to the reduction in statistical power and bias in survival estimates, if heavy censoring is present at early time points the reliability of later survival estimates decreases, and also the comparison between the estimates of two groups may be biased. In this article, a new methodology has been suggested to tackle censoring based on fuzzy methodology.

Apart from the usage of several parametric models based on the distributions of non-negative random variables, Edward L. Kaplan and Paul Meier [7] pioneered the non-parametric methods for estimating the survival probabilities and hence the survival and hazard curves.

The Kaplan-Meier estimator of the survivor function at times, for $t_{(k)} \leq \mathrm{t} \leq t_{(k+1)}$ is given by,

$$\hat{S}(t) = \prod_{j=1}^{k} \left( \frac{n_j - d_j}{n_j} \right) \tag{1}$$

where, $t_1$, $t_2$, $t_3$,... denote the actual times of death of the n individuals in the cohort, $d_1$, $d_2$, $d_3$,... denote the number of deaths that occur at that actual times, and $n_1$, $n_2$, $n_3$,... be the corresponding number of patients remaining in the cohort under study.

Nelson and Aalen [8] developed an estimator for estimating the cumulative hazard function which is directly related to the Kaplan-Meier estimator obtained by maximizing the empirical likelihood.

The estimator is of the form,

$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{2}$$

where $d_i$ is the number of events at time $t_i$, $n_i$ is the total number of individuals at risk at time $t_i$. Many other non-parametric estimators for estimating the survival/hazard function in the presence of censoring such as, Breslow [1], and Kalbfleisch-Prentice [6], are available in the literature.

Nowadays, survival analysis makes use of fuzzy algebra and logic to deal with uncertain failure times, because there may undoubtedly be some fuzziness involved in survival times that occur in nature. Viertl [13] expanded the statistical estimation of the reliability characterizing function, which is typically more or less fuzzy when lifetime data are observed as fuzzy integers.

Grzegorzewski and Hryniewicz [2] used fuzzy sets for modeling the impreciseness of the lifetimes or censored times in the generalization of the exponential model, but the precise information of the observed failures is required. Grzegorzewski and Hryniewicz [3] calculated the mean time to failure, which is also seen as a fuzzy number, and calculated the associated confidence interval using fuzzy numbers.

Viertl [12] discussed the problems of statistical inference relating to the fuzziness present in the data. Shafiq and Viertl [9] stated that the lifetime observations are not precise but more or less fuzzy and based upon this fact they found non parametric estimates for reliability functions based on smoothed empirical distribution function adapted cumulative sum and fuzzy valued empirical distribution function. The interval valued estimation of the reliability function is also suggested. Shafiq. M and Viertl. R [10] have proposed a generalized Kaplan-Meier estimator when the survival times are fuzzy.

Jaisankar et.al [4] introduced a Fuzzy Kaplan Meier estimator and compared it with the traditional approach, which portrayed that the Fuzzified approach gives more reliable results than the Classical one. Jaisankar. R. and Varshini, K. S. P. [5] suggested a fuzzy based procedure through which the censored observations can be tackled to be included in the non-parametric survival analysis. The Kaplan - Meier's procedure was taken for the application of the proposed methodology.

## 2. Methods

Since survival durations are inherently continuous, and treating them as exact is somewhat meaningless. The methodology proposed is rely on this fact and every observed survival time is converted in to fuzzy by means of Triangular Fuzzy number. Due to the nature of survival times that the event might have been occurred prior to the observed survival time, in particular for death like events, Left preference Triangular fuzzy numbers are assumed, which are of the form

$$\tilde{A} = (a, b_w, c) \tag{3}$$

Where

- $a$ is the lower bound where more weightage is given,
- $b_w$ is the middle value generally can be calculated using a weighted mean (adjusted central value), ensuring that the membership function skews towards left, where $b_w = \frac{w_1 a + w_2 c}{w_1 + w_2}$, $w_1 > w_2$, where $w_1$ and $w_2$ are weights
- $c$ is the upper bound here it is considered as the original survival time

The censored subjects are known to have survived up to the time of censoring but the status about the occurrence of their event is not known till the end of the study. If it is known that the event has occurred after they have censored in the later period of time it won't be precise. In both cases the preciseness is not possible and hence assuming fuzziness for the survival times with give reliable results. Here it is assumed that the survival times of censored subjects as Interval valued fuzzy numbers in which the interval is taken to be from the censored time to the fuzzified survival times of the last subject observed. The form of the interval valued fuzzy number is given by

$$\tilde{A} = \left\{ \left( x, \left[ \vartheta_{\tilde{A}}^{L}(x), \vartheta_{\tilde{A}}^{U}(x) \right] \right) \;\middle|\; x \in X \right\} \tag{4}$$

where

- $\vartheta_{\tilde{A}}^{L}(x)$ is the lower bound of the Interval valued Fuzzy set to the actual survival time
- $\vartheta_{\tilde{A}}^{L}(u)$ is the upper bound of the Interval valued Fuzzy set to the maximum survival time in the data set

After these fuzzification procedures the Kaplan Meier estimator has been taken for substantiating and establishing the results.

The analysis was performed after the defuzzification of fuzzy numbers. The fuzzy numbers assumed for the uncensored survival times are defuzzified with reference to a value of alpha cut ($\alpha$ = 0.5) using the following method. For a left preference triangular fuzzy number, the centroid formula for defuzzifying is

$$D = \frac{L + 2M + U}{4} \tag{5}$$

where

- $L$ is the lower bound
- $M$ is the middle bound
- $U$ is the upper bound

The fuzzy numbers proposed for censored observations are defuzzified using centroid method proposed by with reference to a value of alpha cut ($\alpha$ = 0.5) using the following method. The centroid method calculates the average of lower and upper bounds. The formula is:

$$\text{Defuzzified Value} = \frac{\text{Lower Bound} + \text{Upper Bound}}{2} \tag{6}$$

Since it takes in to account both the lower and upper membership degrees, it provides a balanced crisp value after the defuzzification procedure.

The survival times obtained so was applied to the Kaplan - Meier method, and the corresponding results are compared with the Classical Kaplan - Meier method. The survival curves obtained from the proposed method is compared with the survival curve obtained from the Classical Kaplan - Meier method using the log rank test.

## 3. Numerical Illustration

The remission survival periods of 42 leukemia patients, half of whom receive a specific novel treatment therapy and the other half receive standard therapy, are included in the data set for demonstration purposes. Treatment status is the exposure variable of interest ($R_x = 0$, for new treatment, $R_x = 1$ for standard treatment). Sex and log white blood cell count, or logWBC, are two additional variables for control that could be confounding factors. The relapse variable determines failure status (0 if censoring, 1 if failure). ** (Source: Data provided courtesy of David. G. Kleinbaum and Mitchel Klein in the textbook - Survival Analysis - A Self learning text- Third Edition (Page no: 89). URL:http://www.uop.edu.pk/ocontents/survival-analysis-self-learning-book.pdf

## 4. Results

The survival and hazard function curves for the Classical Kaplan-Meier estimator and the Fuzzified Kaplan-Meier estimator, along with their median survival time representation, are depicted above for visual comparison. Figures 1 and 2 illustrate a significant difference among the survival and hazard curves both prior to and subsequent to fuzzification. The median survival time determined by the Classical Kaplan-Meier estimate was 12.00, exceeding the 10.60 observed with the Fuzzified Kaplan-Meier estimator.

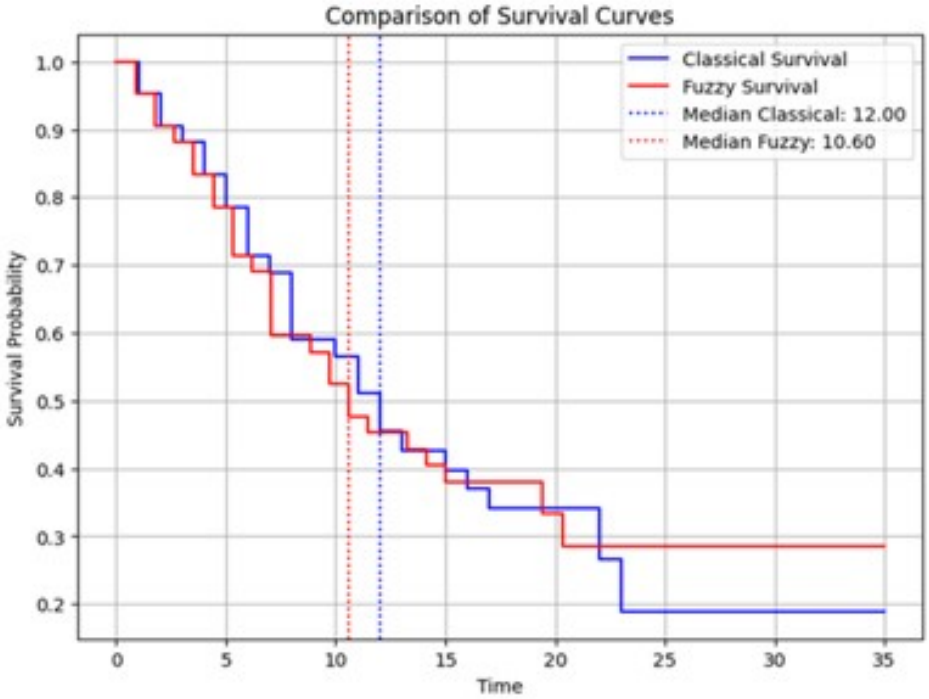


**Figure 1:** *Comparison of Classical and Fuzzified Survival curves along with the representation of Median Survival time in both Classical and Fuzzified cases*

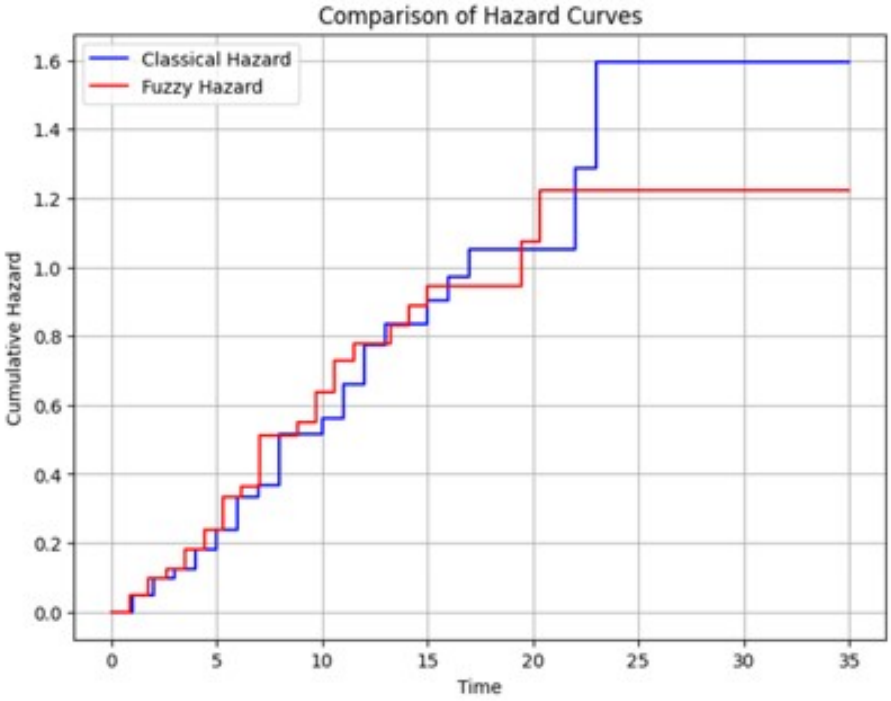


**Figure 2:** *Comparison of Classical and Fuzzified Hazard curves*

To assess if a significant difference exists between the survival curves derived from the Classical Kaplan-Meier estimator and its fuzzified counterpart, the log-rank test was used, with the null hypothesis assuming that the curves are similar. The test statistic for this is represented as follows:
Test statistic is,

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{\sum V_i} \tag{7}$$

where,

- $O_i$ is the observed number of events in Group $i$,
- $E_i$ is the expected number of events in Group $i$,
- $V_i$ is the variance of the difference between the observed and expected events.

The test statistic is calculated using the defuzzified values of survival times, and the details of the results are given below

**Table 1:** *Log rank test results*

| **Log Rank test** | |
|---|---|
| p – Value | 0.833 |

The comparison of the curves via the Log-rank test (Figure 3) indicates that no significant differences are evident in these curves. However, it is dependent upon the choice of the fuzzification factor utilized for the process of fuzzification. When performing a statistical analysis of data derived from a random sample, it is essential to consider both the randomness and the imprecision of the data, particularly when dealing with continuous data. This is particularly accurate with the survival time data. Consequently, a fuzzified variant of the Kaplan-Meier survival estimator was developed, and the associated values of the survival and hazard functions were also calculated.

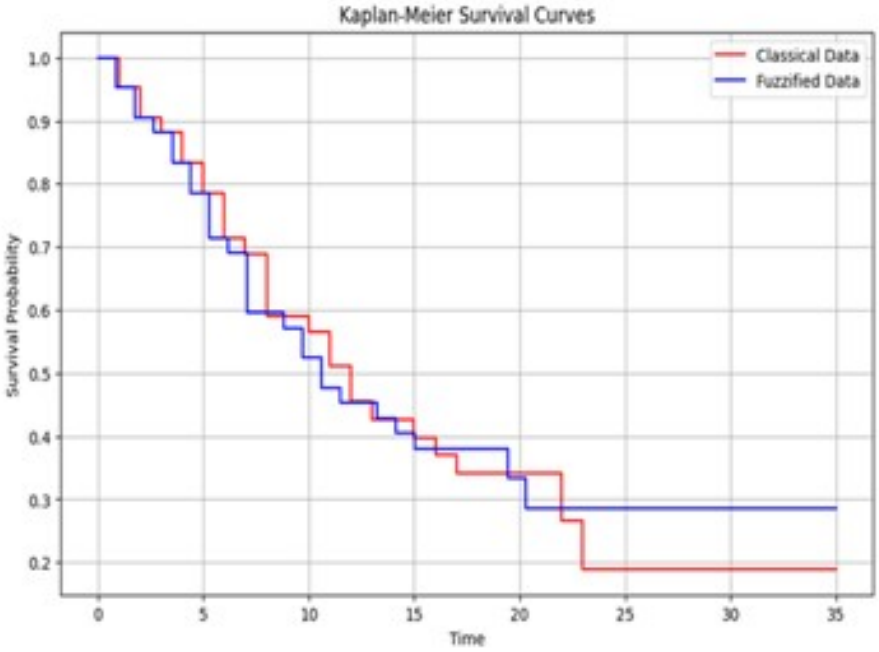


**Figure 3:** *Comparison of Survival curves for the Log rank test*

The incorporation of fuzziness may enhance the model's authenticity; hence yielding correct findings, as observations of a continuous nature often display ambiguity. Disregarding this would decrease the validity of the conclusions derived from every kind of analysis.

## 5. Discussion

Every continuous observations is subject to the presence of fuzziness and every event in survival analysis may not be measured accurately of its time of occurrence. This fact necessitates the present work and it is found that associating fuzziness in Survival analysis leads to changes in survival as well as hazard estimates when compared with the Classical methods that considers only randomness. As the incorporation of fuzziness makes the analysis more natural, one may expect the results which are reliable and applicable.

## References

[1] Breslow N. E. (1975). Analysis of survival data under the proportional hazards model *International Statistical Review/Revue Internationale de Statistique*, 43(1):45-57.

[2] Grzegorzewski, P. and Hryniewicz, O. (1999). Lifetime tests for vague data. *Computing with Words in Information/Intelligent Systems 2: Applications*, 176-193.

[3] Grzegorzewski, P. and Hryniewicz, O. (2002). Computing with words and life data. *International Journal of Applied Mathematics and Computer Science*, 12(3):337-345.

[4] Jaisankar, R., Parvatha Varshini, K. S. and Siva, M. (2022). A Fuzzy Approach to Kaplan-Meier Estimator. *Mathematical Statistician and Engineering Applications*, 71(3s2): 1107-1114.

[5] Jaisankar, R. and Varshini, K. S. P. (2024). On Addressing Censoring in Survival Data Using Fuzzy Theory. *Indian Journal of Science and Technology*, 17(4): 312-316.

[6] Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 267-278.

[7] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457-481.

[8] Nelson W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4): 945-966.
[9] Shafiq, M. and Viertl, R. (2015). Empirical reliability functions based on fuzzy life time data. *Journal of Intelligent & Fuzzy Systems*, 28(2): 707-711.
[10] Shafiq, M. and Viertl, R. (2015) . Generalized Kaplan Meier Estimator for Fuzzy Survival Times. *Śląski Przegląd Statystyczny*, 13(19):7-14.
[11] Thomas, R. Fleming and David, P. Harrington (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics - Theory and Methods*, 13(20): 2469-2486.
[12] Viertl R. Statistical Methods for Non-Precise Data. *CRC Press*, 1995.
[13] Viertl R. (2009). On reliability estimation based on fuzzy lifetime data. *Journal of Statistical Planning and Inference*, 139(5):1750-1755.