Imputation of the Response Variable in Survival Analysis with Interval-Censored Data

Gustavo Soutinho[1(✉)] and Luís Meira-Machado[2]

[1] Research on Economics, Management and Information Technologies (REMIT), Portucalense University, Porto, Portugal
gustavo.soutinho@upt.pt
[2] Centre of Mathematics, University of Minho, Braga, Portugal

Abstract. Handling interval-censored data in survival analysis presents significant challenges, as the exact time to the event is only known to fall within predefined intervals. Common imputation strategies, such as those that use the lower bound, upper bound, or midpoint of the interval, often fail to capture the inherent uncertainty in the data, leading to biased or imprecise estimates. Prior studies have demonstrated the limitations of these approaches, particularly in accurately estimating survival probabilities and hazard ratios.

To tackle these issues, we propose the Scaled Linear Redistribution Method, a new imputation technique aimed at overcoming the limitations of existing methods. The method redistributes imputed values within the interval, keeping their variation and basic statistical behavior. While our approach has not yet been implemented, it represents a promising direction for future research. We plan to evaluate its performance through a comprehensive simulation study, comparing its performance to that of traditional imputation methods and the Turnbull estimator, a widely used nonparametric method for interval-censored data.

Keywords: Interval-censored data · Survival analysis · Imputation methods

1 Introduction

In survival analysis, the estimation of the survival function is important for understanding the time-to-event process under study. However, the presence of interval-censored data, where the exact timing of the event is unknown but known to fall within a specific interval $[L, R]$, poses significant analytical challenges. Interval censoring often arises in studies where individuals are assessed at predetermined intervals, leading to incomplete observations that are a particular case of interval-censored data. Properly handling this type of censoring is essential to avoid bias and obtain reliable estimates.

Despite the availability of specialized statistical methods for interval-censored data, many researchers continue to rely on the Kaplan-Meier estimator [1], which

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15899, pp. 78–88, 2026.
https://doi.org/10.1007/978-3-031-97663-6_6

is designed for right-censored data. In this context, to deal with interval censoring, a typical approach is to assign a specific value within the interval, such as the lower limit, the upper limit, or the midpoint of the interval. While easy to apply, these methods ignore the uncertainty of interval-censored data, which may cause biased results, particularly when event rates vary between intervals [2,4].

To overcome these challenges, nonparametric techniques like the Turnbull estimator [5] have been introduced as an extension of the Kaplan-Meier estimator for interval-censored data. This estimator employs an iterative self-consistency algorithm for survival function using maximum likelihood, which allows to effectively deal with interval censoring in the data. Despite its robustness and flexibility, the Turnbull estimator can be computationally intensive in cases with complex censoring schemes or large datasets [6]. In addition, it does not include covariates or offer hazard ratio estimates, which restricts its application in studies involving more comprehensive inferential insights.

Parametric and semiparametric regression models can be valid alternative approaches, enabling the inclusion of covariates and the modeling of the survival distribution. Among these, the parametric Weibull regression model and the Cox proportional hazards model [7] have been adapted to accommodate interval-censored data [2]. These techniques yield efficient estimates when their underlying assumptions hold. However, if these assumptions are violated or the model is misspecified, their performance can be poor.

Recently, imputation-based methods have become popular due to their adaptability and ability to work with a wide variety of models. As an example, we may refer multiple imputation, which uses covariate data to estimate missing or censored values. The smcfcs R package [8] (Substantive Model Compatible Fully Conditional Specification) highlights its ability to generate imputations that fit well with the survival model being studied. This capacity ensures that the imputed data set is in agreement with the structure of the survival process, becoming an effective tool for handling complex censoring situations [4].

Besides the conventional imputation approaches, we can also mention other new methods that were introduced to overcome the drawbacks of imputing a single value. Strapasson [2] conducted a simulation study to compare various imputation strategies, concluding that methods which explicitly account for interval censoring (such as the Turnbull estimator) tend to be more accurate. Recently, [3] proposed and validated through simulation extensions to SIMEX (Simulation-Extrapolation) that account for uncertainty in both event times and covariate trajectories. This involves modeling the joint distribution of the event and time-varying variables, or aligning imputed event times with the corresponding covariate values. However, when direct approaches are not feasible, the imputation of the midpoint can serve as a reasonable alternative, especially for narrow intervals or when event rates remain relatively stable.

In this study, we build upon these advancements by introducing the Scaled Linear Redistribution Method, a novel imputation approach that adjusts

imputed values to fit within the interval bounds proportionally, preserving the variability of the original distribution.

Thus, this work represents a preliminary step, laying the groundwork for further research that extends beyond its current scope. In future work, we plan to conduct a comprehensive simulation study comparing the performance of this method against three traditional imputation strategies (lower bound, upper bound, and midpoint) as well as the Turnbull estimator. This forthcoming analysis will assess each method's ability to estimate survival probabilities and hazard ratios, offering insights into their strengths and limitations when handling interval-censored survival data. The results of these simulations, along with applications to real data, will be reported in a separate publication.

2 Notation

To facilitate the discussion of interval-censored data and the proposed methods, we introduce the following notation. Let T denote the true (unobserved) time of the event of interest. The event time is known to occur within a specific interval $[L, R]$, where L represents the lower bound and R represents the upper bound of the interval. If the event is only known to have occurred after a specific time L, we set $R = \infty$, indicating right-censoring. In cases where the event time is precisely observed, the interval reduces to a single point, such that $L = R$.

We define the censoring indicator δ to capture the type of observation, where: $\delta = 1$ corresponds to the situation in which the event time is observed exactly, meaning $L = R$; $\delta = 2$ if the observation is interval-censored, such that $L < R$; and $\delta = 0$ if the observation is right-censored, with $R = \infty$.

The estimated survival function, denoted by $\hat{S}(t)$, represents the probability that the event occurs after time t. This function plays a key role in evaluating the performance of the proposed methods and comparing them with existing approaches, such as the Turnbull estimator and the Kaplan–Meier estimator constructed using traditional imputation techniques. This notation will be used consistently throughout the study to describe the interval-censored data and the Scaled Linear Redistribution Method, and to facilitate comparisons among the methods under investigation.

3 Estimation of Survival

3.1 Kaplan-Meier Estimator

The Kaplan-Meier estimator [1], or product-limit estimator, is a well-known nonparametric technique designed to estimate the survival function $S(t)$. This method is especially useful for handling right-censored data. Accordingly, assume now that we have the event time is observed exactly $L = R$ or it is right-censored ($R = \infty$ or $\delta = 0$).

Let $L_1 < L_2 < \cdots < L_k$ denote the ordered, distinct event times of a sample with n individuals. The Kaplan-Meier product-limit estimator of the survival function can also be expressed using the Kaplan-Meier weights as follows [9]:

$$S(t) = 1 - \sum_{i=1}^{n} W_i I(L_{(i)} \leq t). \tag{1}$$

where W_i is the weight attached to $L_{(i)}$ that is defined as:

$$W_i = \frac{\delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{[j]}}{n-j+1}\right), \tag{2}$$

Although the Kaplan-Meier estimator is widely used, it cannot directly accommodate interval-censored data. For such cases, the Turnbull estimator provides a more general approach, as described in the next section.

3.2 Turnbull Estimator

The Turnbull estimator [5] generalizes the Kaplan-Meier estimator to handle interval-censored data, where the exact event time T is unknown but lies within a specified interval $[L_i, R_i]$. Interval censoring commonly arises in longitudinal studies, when the occurrence of an event can only be established retrospectively, at the time of periodic clinical visits or assessments, so that its exact timing remains unknown.

Turnbull's method introduces the concept of *Turnbull intervals*, defined as the smallest set of disjoint intervals that contain all observed interval endpoints. For each interval $(u_j, v_j]$, the survival probability p_j is defined as:

$$p_j = P(u_j < T \leq v_j) = S(u_j) - S(v_j). \tag{3}$$

The likelihood function for interval-censored data is given by:

$$L(p_1, \ldots, p_m) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} p_j \right), \tag{4}$$

where:

- $\alpha_{ij} = I\{(u_j, v_j] \subseteq [L_i, R_i]\}$ indicates whether the j-th interval is contained within the observation window $[L_i, R_i]$,
- $p_j, j = 1, \ldots, m$, are the survival probabilities to be estimated.

To maximize the likelihood, Turnbull's iterative self-consistency algorithm updates the estimates of p_j as follows:

$$p_j^{(r)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij} p_j^{(r-1)}}{\sum_{k=1}^{m} \alpha_{ik} p_k^{(r-1)}}, \tag{5}$$

where r denotes the iteration step. The algorithm continues until successive updates become negligible.

The estimated survival function is then defined as:

$$\hat{S}(t) = \begin{cases} 1, & t \leq u_1, \\ 1 - \sum_{j=1}^{k} p_j, & v_j < t \leq u_{j+1}, \\ 0, & t > v_m. \end{cases} \tag{6}$$

The Turnbull estimator has several notable characteristics. It accommodates exact, interval-censored, and right-censored observations within the same framework, making it highly versatile for a variety of survival data scenarios. The estimator converges to the nonparametric maximum likelihood estimate (NPMLE) of the survival function, ensuring statistical consistency and robustness. However, the iterative nature of the algorithm makes it computationally intensive, particularly when applied to large datasets with complex censoring patterns.

While the Kaplan-Meier estimator is suited for right-censored data, the Turnbull estimator extends its applicability to interval-censored scenarios by leveraging iterative maximum likelihood methods. The Kaplan-Meier estimator with (left, right or midpoint) imputation and the Turnbull estimator will serve as foundational methods in our comparative analysis of survival estimation approaches.

3.3 Illustrative Example

To illustrate the application of the Turnbull estimator and compare it with the Kaplan-Meier estimator using midpoint imputation, we consider a dataset of ten individuals evaluated at fixed time points $t = 2, 4, 6, 8$. The small dataset is presented in Table 1.

Table 1. Dataset of ten individuals evaluated at fixed time points (2, 4, 6, and 8). All subjects were assessed at each visit.

Id	Left	Right	Midpoint	Censoring Indicator (cens)
1	4	6	5	1
2	2	4	3	1
3	6	NA	6	0
4	4	6	5	1
5	2	NA	2	0
6	6	8	7	1
7	6	8	7	1
8	6	NA	6	0
9	4	6	4	1
10	6	8	7	1

Individuals represented in rows 3, 5, and 8 contribute with right-censored observations, as their event times are unknown beyond their last recorded assessment. These cases correspond to instances where the right endpoint (R_i) is missing (denoted as NA), and therefore, the individual does not experience the event

by the end of the study's follow-up period. In contrast, the remaining individuals exhibit interval-censored data, which means, although having experienced the event, their corresponding event times are known only to lie within a specific observation window $[L_i, R_i]$. The Turnbull estimator effectively incorporates both right-censored and interval-censored observations within the same framework, making it a versatile approach for estimating the survival function in this dataset.

The Turnbull estimator requires determining a set of disjoint intervals $(u_j, v_j]$ where survival probability estimates are updated. These Turnbull intervals are constructed using all unique left and right endpoints from Table 1, ensuring that:

> *The intervals are obtained from the set of all left and right interval endpoints in such a way that u_j is a left endpoint, v_j is a right endpoint, and there is no other left or right endpoint between them.*

By applying this rule to the dataset, we identify the following Turnbull intervals:

$$(2, 4], (4, 6], (6, 8]$$

These intervals represent the time periods where events may occur, and the survival function is estimated within them.

The probability mass associated with each of these Turnbull intervals is estimated iteratively using the self-consistency algorithm. For this dataset, the final estimated probabilities p_j corresponding to the intervals $(2, 4]$, $(4, 6]$, and $(6, 8]$ are 0.1481, 0.2963, and 0.5556, respectively. These values represent the proportion of total failures distributed across the intervals, ensuring that the sum of all interval probabilities equals one. The survival function may then be constructed stepwise using these probabilities, preserving the interval-censored nature of the data.

A common approach for handling interval-censored data is to apply the Kaplan-Meier estimator with midpoint imputation, where each interval $[L_i, R_i]$ is replaced by its midpoint, as shown in Table 1: $T_i = \frac{L_i + R_i}{2}$. The corresponding censoring indicator is provided in the last column of the table, where a value of 0 denotes a right-censored observation, while a value of 1 indicates that an event is assumed to occur at the midpoint of the interval.

Applying the Kaplan-Meier estimator to these imputed values provides an approximate survival curve, but at the cost of assuming that all events occur exactly at the midpoint rather than being distributed across the Turnbull intervals.

In this specific case, applying the `survfit` function from the `survival` package in R for interval-censored data, using the `interval2` method within the `Surv` function, produces the same survival estimates as those obtained through midpoint imputation. This equivalence occurs because the study design assumes fixed evaluation times for all individuals, with no missed assessments at any evaluation point. When all individuals are evaluated at identical, predetermined

time points, the Turnbull estimator effectively distributes event probabilities in a manner that aligns with the midpoint imputation approach. For instance, considering the event times 2, 3, 5, 6, and 7, the survival estimates produced by both estimators are 1, 0.889, 0.556, 0.556, and 0.

However, this equivalence does not hold in more general cases. If the study includes individuals with variable assessment times or cases where participants miss scheduled evaluations, the two methods can yield different survival estimates. For instance, if we modify the first individual's observation by setting the lower bound to 2 instead of 4, this change would reflect a scenario where this individual missed the evaluation at time 4. Under this modification, the Turnbull estimator and the midpoint imputation method would produce distinct survival estimates, as the interval-censored model appropriately accounts for the uncertainty introduced by the missing assessment. Thus, in this case, the Turnbull estimator yields survival probabilities of 1, 0.852, 0.556, 0.556, and 0, while the midpoint imputation estimator gives 1, 0.889, 0.556, 0.556, and 0.

4 Imputation Methods

4.1 Usual Imputation Methods

For interval-censored observations, where $L_i < R_i$ and R_i is finite, imputation presents a unique challenge. Conventional methods for handling such observations often involve imputing a single representative value within the interval. Typical methods are:

- Left Imputation: Assigning the event time to the lower (L_i) limit of the interval, under the assumption that it occurred at the earliest possible time within the given range.
- Right Imputation: Assigning the event time to the upper bound (R_i) of the interval, assuming it occurred at the latest possible moment within the observed range.
- Midpoint Imputation: Assigning the event time to the midpoint interval ($(L_i + R_i)/2$) of the interval, assuming the event time is uniformly distributed within the interval.

Although these methods are simple, they fail when taking into consideration the variability and uncertainty within the interval. Furthermore, they may exhibit systematic bias, particularly when the event times do not follow a uniform distribution or are strongly influenced by covariates.

Another possible approach considers the response variable as missing and estimates its value based on the connections between the covariates and the survival outcome. This method treats the interval-censored response as a missing data point, which can be imputed using appropriate techniques. However, the difficulty lies in ensuring that the imputed value remains within the bounds of the observed interval $[L_i, R_i]$.

4.2 The Scaled Linear Redistribution Method

In this section, we introduce a new proposal that we call the Scaled Linear Redistribution Method. This new method ensures that imputed values stay within the interval and keeps the natural variation in the data.

Step-by-Step Procedure

The Scaled Linear Redistribution Method comprises the following step-by-step procedures:

Step 1: For each individual with interval-censored data, perform an initial random imputation by simulating a value $T_i^{(0)}$ from a uniform distribution:

$$T_i^{(0)} \sim U(L_i, R_i),$$

where L_i and R_i are the lower and upper bounds of the censoring interval.

Step 2: Estimate a regression model using a parametric Accelerated Failure Time (AFT) model [10–12] or a machine learning model such as XGBoost [13] using the initial imputed survival times. Then, use the trained model to predict new survival times for individuals with interval-censored data.

Predicted times may include values outside the interval bounds. For instance, if an individual's event is known to occur between $L = 4$ and $R = 5$, imputations may be observed over a broader range (e.g., $[2, 12]$).

Step 3: Iteratively perform the simulation and prediction process by repeating Step 1 and Step 2 M times (e.g., $M = 100$), each time generating new simulated values and refitting the predictive model. This results in a distribution of predicted survival times for each interval-censored observation.

Step 4: For each individual i, obtain the empirical distribution of predicted times $\hat{T}_i^{(1)}, \ldots, \hat{T}_i^{(M)}$ from the previous step. To ensure compatibility with the original censoring information, these predicted values are rescaled to fit within the interval $[L_i, R_i]$ using the following transformation:

$$T_i^{\text{scaled}} = L_i + \frac{\hat{T}_i - \min(\hat{T}_i)}{\max(\hat{T}_i) - \min(\hat{T}_i)}(R_i - L_i). \tag{7}$$

That way, all values stay within the interval while keeping their variability.

Step 5: Using the rescaled imputed times, re-run the regression model, using a parametric model or machine learning method on the updated dataset. The final imputed values for interval-censored observations are then obtained as the median of the predicted, scaled survival distributions:

$$T_i^{\text{final}} = \text{median}(\hat{T}_i^{(1)}, \ldots, \hat{T}_i^{(M)}). \tag{8}$$

To assess the uncertainty of this median, calculate the standard deviation (SD) of the M imputations for each individual. Finally, the SE for each individual

is computed by dividing the SD by the square root of M, reflecting the variability and uncertainty in the estimated median across the replicates. This process is repeated for each individual, and the resulting SE values provide a measure of the precision of the imputations for each subject.

By redistributing values proportionally, this approach ensures that the variability of the initial imputed distribution is retained, while all values are constrained within the interval $[L, R]$. This avoids the artificial narrowing of the data range associated with imputing a single point such as midpoint while maintaining the stochastic nature of the imputations.

One of the advantages of the proposed Scaled Linear Redistribution Method lies in its ability to incorporate covariate information directly into the imputation process. Rather than treating censored intervals in isolation, the method used the relationship between covariates and the survival outcome to obtain informed imputations.

This ensures the model stays aligned with the way the data was imputed. This contrasts with simpler strategies, such as using the interval midpoint or one of its bounds (lower or upper), which can introduce bias or overlook essential information in the data.

The method also preserves the natural variability in the data. By redistributing values proportionally within the interval, it captures the available randomness of the process without flattening the distribution, resulting in more realistic and robust imputations.

Since the imputations rely on a survival model (Step 2), the method stays aligned with the later analysis, ensuring consistency between assumptions and imputations.

Table 2 provides a comparison of commonly used imputation approaches, including the well-known lower bound, upper bound, and midpoint imputations, as well as the proposed Scaled Redistribution method. While conventional methods such as lower and upper bound imputations are simple, they often lead to systematic biases by either underestimating or overestimating event times. The midpoint imputation assumes a uniform distribution within the interval, which may be unrealistic in many scenarios. In contrast, the Scaled Redistribution

Table 2. Comparison of different imputation methods for interval-censored data.

Method	Description	Limitations
Lower Bound Imputation	Impute L, the start of the interval.	Underestimates event times and ignores variability
Upper Bound Imputation	Impute R, the end of the interval.	Overestimates event times and ignores variability
Midpoint Imputation	Impute $(L + R)/2$, the midpoint of the interval.	Assumes uniform distribution within the interval; unrealistic
Scaled Redistribution	Adjusts imputed values to fit within $[L, R]$ while preserving variability.	Computationally intensive for very narrow intervals

method ensures that imputed values respect the interval bounds while preserving variability, although it can be computationally demanding for narrow intervals. Table 2 also summarizes these methods and highlights some of their advantages and limitations.

5 Conclusion

Interval-censored data is challenging because we don't know exactly when events happen. In this context, imputation techniques such as lower-bound, upper-bound, and midpoint methods, even though they provide straightforward and computationally efficient solutions, often overlook the natural variability within observed intervals, which can lead to biases and reduce the accuracy of survival estimates. To address these problems, we propose the Scaled Linear Redistribution Method, a new approach that keeps the variation in the data and respects the original interval boundaries.

The Scaled Linear Redistribution Method derives from imputations produced through the `smcfcs` library, which accounts for covariate information and ensures consistency with the survival model. This approach adjusts imputed values to stay within the observed interval, which helps reduce the reliance on arbitrary assumptions commonly found in simpler methods, such as the uniform distribution used in midpoint imputation.

The proposed methods are currently under investigation, meaning that the presented formulation has not yet been implemented or validated. Based on these early ideas, the method seems to offer advantages in flexibility and how it handles the data. However, no conclusion can yet be drawn regarding performance relative to existing imputation techniques. The proportional adjustment mechanism could also increase computational complexity, particularly when dealing with datasets that have narrow intervals or a high volume of observations. Additionally, since the method relies on initial imputations from `smcfcs`, its overall effectiveness is dependent on the accuracy of those imputations.

This remains an ongoing area of research. Future efforts will concentrate on implementing the method, improving its computational efficiency, and performing comprehensive simulation studies to evaluate its accuracy and robustness across various censoring scenarios, including left-censoring and mixed-censoring contexts. To this end, future work will include a thorough comparison of different methodologies, making use of both real-world datasets and simulation studies.

In summary, the Scaled Linear Redistribution Method holds potential as an innovative approach for addressing interval-censored survival data. While it aims to improve on existing methods, the approach still needs to be fully tested. If proven effective, this method could become a useful tool in survival analysis. Ongoing work and future findings will be presented elsewhere.

Acknowledgments. This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

1. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**(282), 457–481 (1958)
2. Strapasson, D.R.: Modelos para análise de sobrevivência com censura intervalar: Estudo de simulação e aplicação. Master's thesis, Universidade de São Paulo (2007)
3. Abrahamowicz, M., Beauchamp, M.-E., Moura, C.S., Bernatsky, S., Guerra, S.F., Danieli, C.: Adapting SIMEX to correct for bias due to interval-censored outcomes in survival analysis with time-varying exposure. Biom. J. **64**(8), 1467–1485 (2022)
4. Gomes, G., Giolo, S.R., Colosimo, E.A.: Survival analysis for interval-censored data: an R tutorial. Stat. Model. **9**(3), 269–287 (2009)
5. Turnbull, B.W.: The empirical distribution function with arbitrarily grouped, censored, and truncated data. J. Roy. Stat. Soc.: Ser. B (Methodol.) **38**(3), 290–295 (1976)
6. Lin, T.I., Lee, J.C., Yen, S.Y.: Finite mixture modelling for interval-censored survival data. Stat. Med. **29**(1), 7–20 (2010)
7. Cox, D.R.: Regression models and life-tables (with discussion). J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**(2), 187–202 (1972)
8. Bartlett, J.W., Keogh, R., Bonneville, E.F., Ekstrøm, C.T.: smcfcs: substantive model compatible fully conditional specification. R package version 1.9.1 (2024). https://cran.r-project.org/package=smcfcs
9. Meira-machado, L.: The Kaplan-Meier estimator: new insights and applications in multi-state survival analysis. In: Lecture Notes in Computer Science, pp. 129–139 (2023)
10. Lawless, J.F.: Statistical Models and Methods for Lifetime Data, 2nd edn. Wiley, Hoboken (2003)
11. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data, 2nd edn. Wiley, Hoboken (2002)
12. Wei, L.J.: The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat. Med. **11**(14–15), 1871–1879 (1992)
13. Wang, H., Li, B., Ghosh, D.: A gradient boosting approach for censored survival data. Stat. Med. **38**(23), 4834–4846 (2019)