Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27841-cd

Cross-Validation vs. Information Criteria: Weibull Model Selection under Incomplete and Uncertain Failure Data

David Westermann

Machine Protection and Electrical Integrity Group, European Organization for Nuclear Research (CERN), Switzerland. Institute of Machine Components (IMA), University of Stuttgart, Germany.
E-mail: david.westermann@cern.ch

Martin Dazer

Institute of Machine Components (IMA), University of Stuttgart, Germany.
E-mail: martin.dazer@ima.uni-stuttgart.de

Lukas Felsberger, Jan Uythoven and Daniel Wollmann

Machine Protection and Electrical Integrity Group, European Organization for Nuclear Research (CERN), Switzerland. E-mail: lukas.felsberger@cern.ch, jan.uythoven@cern.ch and daniel.wollmann@cern.ch

Weibull analysis is a central tool in reliability engineering, to model system failures and support maintenance planning. Failure data are rare and expensive to obtain, practically always limited, censored, and the number or type of underlying failure mechanisms uncertain. Consequently, it can be unclear which Weibull model accurately describes the data. Traditionally, candidate Weibull models are fitted to all available data, with information criteria (ICs) such as Akaike's information criterion (AIC) or Bayesian information criterion (BIC) used to identify the best model by balancing fit and model complexity. However, fitting to the entire dataset without assessing predictive performance on unseen data can lead to suboptimal model selection. Cross-validation (CV) offers an alternative by evaluating a model's predictive accuracy, which can improve model selection and failure forecasts. Although prior work indicates that CV can outperform ICs in some modeling tasks, the circumstances in which it yields superior performance for single and multiple Weibull models across different censoring regimes remain insufficiently understood.
To address this gap, synthetic data are generated from single and multiple two- and three-parameter Weibull models to assess model selection across realistic sparsity levels. CV exhibits the lowest predictive errors in small-sample, high-censoring scenarios. ICs yield comparable predictive performance only for larger, less censored samples. Performance in identifying the true model depends on the Weibull model type, with BIC performing best for parsimonious two-parameter models, CV for three-parameter models, and AICc for competing or mixture models. The methods are then applied to long-term failure data from two systems at the European Organization for Nuclear Research (CERN) to validate the findings, revealing additional effects not fully modeled by the synthetic setup. The findings can be translated to improved reliability modeling and cost savings in scenarios involving incomplete and uncertain failure data, but should be refined using additional real-world examples.

Keywords: Weibull analysis, model selection, cross-validation, information criteria, AICc, BIC, incomplete data, uncertain data, predictive accuracy, true underlying model.

1. Introduction

Failure data of a system can provide insights into failure mechanisms and hazard rates, enabling product improvement and predictive maintenance planning. To exploit this information, the model that best represents the underlying failure behavior and yields the highest predictive accuracy must be identified. Accurate model selection is particularly crucial for small sample size and high censoring scenarios, whether for fielded systems requiring timely replacement or cost-effective lifetime tests requiring low sample sizes or test times. Typically, Weibull analysis is conducted by fitting multiple models and selecting the model with the lowest information criterion (IC) value. However, it does not explicitly assess generalization. Cross-validation (CV) is established in machine

learning but rarely applied in reliability engineering. It explicitly evaluates generalization and may improve model selection.

1.1. *Background*

The relevant background is organized as follows: first, an overview of Weibull analysis and Maximum Likelihood Estimation (MLE) is given, followed by a review of model selection techniques, including AIC, AICc, BIC, and cross-validation.

1.1.1. *Weibull analysis*

In a traditional Weibull analysis, the number of underlying failure mechanisms should be known. In practice this is often difficult or expensive to establish a priori and multiple Weibull models are fitted to all available time-to-event data, including both failures and suspensions, representing not yet failed units. To model data originating from a single failure mechanism, a two-parameter (2P) or three-parameter (3P) Weibull model is used. The failure probability equation of the 3P-Weibull model (Eq. (1)) is defined by the shape parameter β, scale parameter λ, and location parameter γ, which is equal to zero for 2P (Weibull (1951)),

$$\mathrm{F(t)} = 1 - \exp\left(-\left(\frac{t-\gamma}{\lambda}\right)^{\beta}\right). \qquad (1)$$

Systems are typically exposed to competing risks in the form of multiple failure mechanisms acting simultaneously (Cox (1959)). For example, an electronic system may fail due to a short or open circuit of one or several components. Such systems are modeled as series configurations, where each failure mechanism contributes a failure probability F_i, $F = 1 - \prod_i (1 - F_i)$. In contrast, mixture models apply to populations comprising multiple subpopulations, e.g., systems from different manufacturing lots with varying quality levels. For two subpopulations, the total failure probability equals $F = p \cdot F_1 + (1-p) \cdot F_2$, where p represents the proportion (Mendenhall and Hader (1958)).

To fit a Weibull model, its parameters must be estimated. In practice, maximum likelihood estimation (MLE) is one of the most common used methods, determining the parameters by maximizing the likelihood function, which is the probability of observing the given data with the specified distribution. This involves setting the derivatives of the log-likelihood (LL) function,

$$\ln L = \sum_{i=1}^{n} \left[\delta_i \ln f(t_i) + (1-\delta_i) \ln R(t_i)\right], \qquad (2)$$

to zero, yielding nonlinear equations for the Weibull distribution that require numerical solution via optimizers. The LL is calculated as the sum of the log-densities for failures $f(t_i)$ and log-reliability values $R(t_i)$, multiplied with the indicator δ_i to distinguish between the failed and suspended states. (Harter and Moore (1965); Lai (2014))

1.1.2. *Model selection*

Model selection aims to identify the best model from a set of candidates. The primary goal is typically the model's ability to generalize, i.e. its capacity to accurately predict unseen data. For this the true underlying model needs to be identified to avoid overfitting by selecting models with too many parameters. Common model selection methods are information criteria (e.g., AIC, AICc, BIC) and resampling methods, such as cross-validation. Model selection with **information criteria**, such as Akaike's information criterion (Akaike (1974)),

$$\mathrm{AIC} = -2\ln(\hat{L}) + 2k, \qquad (3)$$

and the Bayesian information criterion (Schwarz (1978)),

$$\mathrm{BIC} = -2\ln(\hat{L}) + k\ln(n), \qquad (4)$$

favors the model with the lowest IC value. Both are based on rewarding the goodness-of-fit, by taking the likelihood term ($-2\ln(\hat{L})$, $\hat{L}$ = maximized likelihood) into account while penalizing model complexity based on the number of model parameters k. AIC penalizes model parameters by multiplying them by two, whereas BIC multiplies them by the natural logarithm of the sample size n, resulting in harsher penalization than AIC for sample sizes larger than seven. Due to its penalty construction, BIC tends to select less complex models. Another commonly used IC is the corrected AIC (Hurvich and Tsai (1989)),

$$\mathrm{AICc} = -2\ln(\hat{L}) + 2k + \frac{2k(k+1)}{n-k-1}. \qquad (5)$$

It is recommended when the sample size to model parameter ratio, n/k, is smaller than 40, as AIC does not penalize model complexity sufficiently in these scenarios. To avoid this in practice, Burnham and Anderson (2004) recommend using AICc regardless of the ratio, as both criteria yield nearly identical values for larger ratios. Following Burnham and Anderson (2004), candidate models are assessed using ΔAIC, the difference in score relative to the top-ranked model, i.e. the model with the lowest IC score. A ΔAIC less than two indicates substantial support, values from four to seven indicate considerably less support, and differences greater than ten suggest the model can be discounted. The same applies for AICc. A similar rule holds for BIC, introduced by Raftery (1995). Strong support exists when a candidate is within a range of two to the top-ranked model.

Cross-validation evaluates model generalization on unseen data by repeatedly splitting into training and validation sets. A common CV method is k-fold CV, which divides n observations into k equal folds (typically $k = 5$ or 10). In each of k iterations, the model is trained on $k - 1$ folds and validated on the remaining fold. Standard k-fold CV is unsuitable for Weibull analysis of highly censored data, as folds may contain only suspensions. Stratified k-fold CV instead preserves the failure:suspension ratio (e.g., 20:80 from 20 failures, 80 suspensions) per fold (Kohavi (1995)). Generalization error is measured with metrics like negative log-likelihood. (James et al. (2021))

1.2. *Related work and research gaps*

According to Arlot and Celisse (2010), it is a natural question in model selection whether criteria, such as AIC, should be preferred over cross-validation. However, this question has been rarely addressed in the context of reliability engineering and Weibull analysis. To date, ICs have been widely used in Weibull analysis and associated computing libraries such as the Python *reliability* library (Reid (2025)). In contrast, in other disciplines, where event forecasting is of major interest, CV is regularly applied, and corresponding comparisons with ICs have been investigated.

Liu and Zhou (2024) used CV on short psychological time series, typically modeled with high-parametric models that risk overfitting and poor generalization. One part of their work compared two CV methods against AIC/BIC for selecting models with better predictive accuracy. To this end, three models with 12 to 63 parameters were fitted to datasets with the first 30 to 500 observations. *One finding was that CV generally outperformed AIC/BIC in selecting the model with best predictive accuracy, while BIC often selected overly simple models leading to underfitting*. However, censored data or Weibull models were not considered. The latter was considered by Takezawa (2018), comparing for small datasets AIC and likelihood CV for selecting between 2P-Weibull ($\beta = 1$–2 and $\lambda = 3$) and exponential models, aiming for optimal predictive accuracy. Generally, *CV demonstrated a lower estimation bias in approximating the expected log-likelihood making it preferable to AIC in this respect*. Censoring was considered by Bermejo and Grimm (2024). They compared model selection using AIC/BIC against 10-fold CV with AIC/BIC as the validation metric. Using synthetic datasets of 250 patients each, and applying right censoring at median survival time, they fitted standard parametric and flexible survival models. *On average, CV selected models with lower prediction error than fitting to all data using AIC/BIC, favoring simpler models*. It remains unclear whether these findings also apply for other censoring regimes and different Weibull models.

Literature indicates that CV outperforms ICs in model selection for predictive accuracy. While CV is typically used for selecting a model with best predictive accuracy, identifying the true underlying model can be equally important, for example, when determining the number of underlying failure mechanisms. Myung et al. (2009) generated datasets from 6- and 7-parameter models with $n = 27$, to which each candidate model was fitted. *ICs and sample-based methods more frequently identified the true underlying model than when solely relying on Maximum Likelihood. ICs outperformed CV on average: AICc/BIC excelled for the simpler model and AIC for the 7P model.*

However, Weibull models or censoring were not considered.

Existing research comparing ICs and CV has not addressed multiple Weibull models, nor the impact of various censoring regimes despite studies on small sample sizes. Moreover, standard statistical or heuristic methods, such as ΔAIC/AICc thresholds, are rarely applied in this context. Consequently, their value when applying ICs and CV remains unclear.

1.3. *Study overview*

Based on the identified research gaps the following research questions are defined in the context of Weibull analysis:

- Which method, ICs or CV, is **superior in selecting models with a higher predictive accuracy and true underlying models**, in particular when failure data is sparse?
- How do **statistical significance tests and heuristic methods influence model selection** with ICs or CV?

To address these questions, synthetic data are generated from single 2P/3P-Weibull, competing risk, and mixture Weibull models, accounting for various censoring regimes, sample sizes, and Weibull parameters. AICc, BIC, and stratified k-fold CV are compared to determine when ICs or CV yield superior model selection performance. To assess the influence of significance tests and heuristic methods on model selection performance, IC and CV specific strategies are applied, such as a $\Delta\mathrm{AICc/BIC} \leq 2$ threshold.

2. Methodology

The methodology follows the workflow depicted in Fig. 1, from data generation through model selection to evaluation via a Python script. The model selection and evaluation are then applied to two CERN accelerator systems.

2.1. *Data generation*

First, data are generated from one individual 2P-, one individual 3P-, two competing 2P- or two mixed 2P-Weibull models with stated parameter values for n, β, λ, and model specific parameters. Data from a 2P or 3P model are generated from the corresponding distribution. For competing risk (Cr) the earliest generated failure time is retained as system lifetime. Only the β and λ combinations that fulfill a minimum required percentage (min) of failures (20% or 40%) from each Weibull component are considered. For mixture (Mix) models, n failure times are simulated from each 2P-Weibull model and stochastically selected for each of the samples based on the proportion p (20% or 40%). For Cr and Mix, parameter combinations with identical β and λ values are excluded. Sample sizes are selected to reflect small, medium and large (100, 1000, 10000) populations. Beta values reflect early-life (0.5) and aging failures driven by gradual or instantaneous aging onset (1.5 to 12). A β value of one is excluded as for a constant hazard rate model selection becomes indistinguishable. For λ, values from 2.5 to 40 are selected to represent common ranges across various systems. For 3P models, small to large failure free times (1 to 16) are considered, reflecting typical values. After generating failure data, Type-II censoring is applied: after a predefined number of failures, remaining units are suspended.

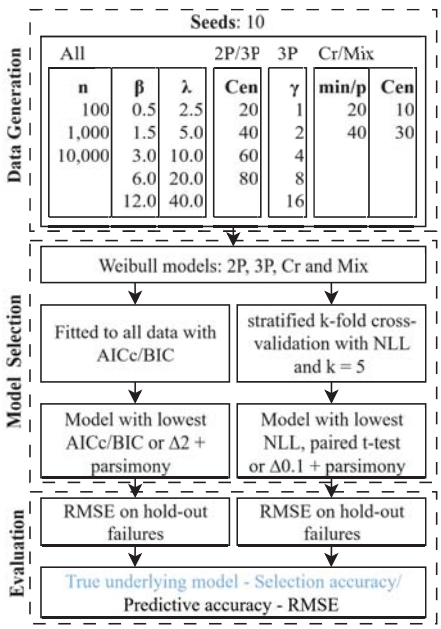

Fig. 1. Workflow for synthetic data experiments.

For example, 20% censoring in 50 units censors the last ten observations at the 40th failure time, with true failure times retained as held-out data for predictive evaluation. Censoring levels range from 20-80% for 2P/3P models and 10% or 30% for Cr/Mix models to ensure at least 10% of failures per Weibull component. All data generation is repeated with ten seeds, providing sufficient statistics for identifying model selection performance across scenarios, even after filtering individual sampled data scenarios where MLE estimation failed.

2.2. *Model selection*

After data generation, 2P/3P-Weibull, competing risk, and mixture models were fitted using the *reliability* library (Reid (2025)) via MLE with various optimizers to ensure robust convergence. Models were fitted to all data for AICc/BIC computation or to the training sets for CV. Following common practice, 5-fold CV is employed. The negative log-likelihood (NLL) is used as a validation error metric to account for censored data. Model selection metrics are minimum AICc, BIC, CV validation error (average NLL), or $\Delta\mathrm{AICc/BIC} \leq 2$, corrected resampled t-test for CV, and a CV variant with a $\Delta \leq 0.1$ threshold to exclude selecting models based on irrelevant performance differences originating from numerical effects. In the presence of multiple equally performing models, the parsimony principle is applied, selecting the model with the fewest parameters (Burnham and Anderson (2004)). Combined with the delta thresholds, this is referred to as the *delta rule* in this study. The corrected resampled t-test by Nadeau and Bengio (2003) determines statistical significance while accounting for CV fold overlap. The test statistic t,

$$t = \frac{\bar{d}}{\sqrt{S_d^2\left(\frac{1}{k} + \rho\right)}}, \tag{6}$$

uses the average of fold-wise validation error differences $\bar{d}$ and sample variance of the differences S_d^2, number of folds k, and correction factor $\rho = 1/(k-1)$. The resulting p-value from the t-distribution is compared against Bonferroni (1936)-adjusted significance level α ($\alpha_{adj} = \alpha/m$, m = number of comparisons) to control multiple comparisons.

2.3. *Evaluation*

The model selection accuracy for AICc, BIC, and CV is assessed through the percentage of correctly identified models. An accuracy of 25% equals the probability of randomly selecting the model from four equally likely candidates ($1/4 = 25\%$). The predictive accuracy of these selected models is compared using the Root Mean Squared Error (RMSE) of the failure probability,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_{\mathrm{hold}}}\sum_{i=1}^{n_{\mathrm{hold}}}\left(\hat{F}(t_i) - F(t_i)\right)^2}, \tag{7}$$

calculated on hold-out failures (i.e., failure times of censored units, n_{hold} = number of holdout failures). An RMSE of 0.1 indicates an average deviation of ten percentage points between the predicted $\hat{F}$ and the true distribution F ($F(t_i) = k_i/n$, k = rank) within hold-out data.

A scan across application scenarios and associated data-generation parameters analyses the performance variations of each method with increasingly sparse data. This is further elaborated in Sec 3.

2.4. *Use cases*

The model selection and evaluation framework is applied to right-censored data from two CERN accelerator systems. The first is the CIBF, the power supplies for the Beam Interlock System (BIS) of the Large Hadron Collider, designed to protect against accidental beam-induced damage (Puccio et al. (2010)). 145 systems with failures from two competing failure mechanisms are considered (85% censoring): the first (M1) with fifteen failures ($\beta_1 = 1.8$, $\lambda_1 = 26.9$), the second (M2) with seven failures ($\beta_2 = 8.6$, $\lambda_2 = 11.6$). The second application involves the Super Proton Synchrotron (SPS) main dipole magnets (CERN (1976)). Operational since 1981, these components require decades of future service; thus, selecting a model with peak predictive accuracy is essential for forecasting their remaining lifetime. 249 magnets and 24 observed failures (90% cen-

soring) are considered, supposedly from one failure mechanism (M1). Data are artificially Type-II censored in 2017 or 2018 to create six or five hold-out failures, respectively, for prediction error assessment on data up to 2025.

3. Results

The results section first addresses true model selection, followed by predictive accuracy, with synthetic data results presented before those from real-world applications. Table 1 summarizes findings for synthetic data, displaying the selection accuracy of the true underlying model on the left and the predictive accuracy, expressed as the averaged RMSE, on the right. Performance is evaluated for each method across best, moderate, and worst-case scenarios. Associated sample sizes and censoring levels are specified for each scenario. Results for all three methods are shown with and without applied delta rules ($\Delta \leq 2$ for ICs, $\Delta \leq 0.1$ for CV), with the winner within each selection strategy (with/without applied rule) highlighted. Results for t-tests with $\alpha = 0.05$ were omitted, as they frequently indicated similar performance among three to four candidates across cases, leading to significant over-selection of the 2P model.

3.1. *Selection accuracy - synthetic data*

The accuracy of each method in identifying the true underlying model for synthetic data is shown in the left panel of Table 1. On average, smaller sample sizes and higher censoring levels decrease model selection accuracy. Across best-, moderate-, and worst-case scenarios, no distinct performance shift between methods is observed; instead, performance appears to depend on the underlying Weibull model. AICc tends to perform best for models with higher parameter counts (Cr and Mix), whereas BIC shows superior results for the simpler 2P models and CV for the 3P models. When the number of failure mechanisms is uncertain, and sample sizes are large with low censoring, BIC consistently performs well and is generally recommended. AICc is preferred when it is known that failure data stems from two failure mechanisms, but it remains unclear whether these follow a Cr or Mix model. Applying delta rules enhances model selection for AICc and CV on average. Especially in selecting the 2P model, accuracy improvements of up to 29% and 32%, respectively, are observed. This suggests that both methods tend to overfit without such adjustments. For BIC, the applied delta rule yields no improvement.

3.2. *Predictive accuracy - synthetic data*

The results for predictive accuracy (RMSE) of each method with synthetic data are illustrated in the right panel of Table 1. They indicate that all methods perform equally well in the best-case scenario. In the moderate-case, BIC excels for 2P and Cr models, while CV outperforms for 3P models and both CV and AICc for Mix models. In the worst-case, CV yields the lowest RMSE across all models. While AICc and BIC remain competitive in moderate cases, CV demonstrates superior generalization under high-censoring conditions. Applied delta rules yield on average improvements for AICc and CV, but not for BIC.

Comparing the left and right panel of the results table indicates that predictive accuracy and true

Table 1. Results for the selection of the true-underlying model and the predictive accuracy.

				True underlying model - Selection accuracy						Predictive accuracy - Average RMSE					
Case	**Model**	**n**	**Cen**	**AICc**	**BIC**	**CV**	**AICcΔ2**	**BICΔ2**	**CVΔ0.1**	**AICc**	**BIC**	**CV**	**AICcΔ2**	**BICΔ2**	**CVΔ0.1**
Best	2P	10,000	20	62%	100%	40%	76%	100%	62%	0.002	0.002	0.002	0.002	0.002	0.002
	3P			92%	87%	95%	92%	84%	95%	0.002	0.002	0.002	0.002	0.002	0.002
	Cr		10	87%	91%	83%	90%	91%	88%	0.002	0.002	0.002	0.002	0.002	0.002
	Mix			97%	91%	97%	95%	91%	97%	0.003	0.003	0.003	0.003	0.003	0.003
Moderate	2P	1,000	60	64%	100%	40%	93%	100%	72%	0.028	0.015	0.025	0.016	0.015	0.022
	3P			60%	54%	76%	61%	50%	70%	0.019	0.018	0.018	0.018	0.019	0.017
	Cr		30	81%	78%	75%	87%	77%	79%	0.020	0.017	0.020	0.019	0.017	0.018
	Mix			81%	63%	79%	74%	59%	74%	0.030	0.032	0.030	0.031	0.033	0.030
Worst	2P	100	80	33%	73%	58%	61%	79%	80%	0.249	0.176	0.154	0.178	0.163	0.142
	3P			40%	35%	43%	37%	31%	28%	0.219	0.188	0.179	0.191	0.177	0.168
	Cr		30	58%	49%	49%	58%	44%	51%	0.060	0.060	0.055	0.060	0.060	0.055
	Mix			55%	39%	51%	43%	35%	47%	0.065	0.062	0.060	0.064	0.062	0.060

model selection accuracy are not always aligned, as an incorrect model choice may result in only marginal RMSE penalties. This is exemplified by the Cr model's ability to adapt to 2P-Weibull data. Because some models win frequently by small margins while others offer rare but substantial gains, the overall RMSE can be driven by a small number of highly accurate selections.

3.3. *Model selection use case - CIBF*

For the CIBF, minimum-score selection correctly identified the Cr model for AICc and CV, whereas BIC selected the 2P model (see Fig. 2, CIBF). With delta rules applied, both ICs favored the 2P model, while CV identified the correct model. Since this case reflects a *worst-case* regarding sample size and censoring, the lack of consistent model selection aligns with synthetic findings. Despite these challenges, CV consistently selects the Cr model.

3.4. *Predictive accuracy use case - SPS dipole magnets*

Using SPS magnet failure data up to 2017, BIC selected the model with the lowest RMSE (2P, RMSE $= 0.002$) both before and after the delta rule was applied. In contrast, AICc selected the 3P model (RMSE $= 0.009$) initially but switched to the 2P model after the delta rule, while CV consistently selected the Cr model (RMSE $= 0.014$). For failure data up to 2018, all methods selected the 2P model, both before and after applying delta rules. The 2P model shows the second-lowest RMSE (0.003), with the 3P model performing best (0.002). For 2017, only CV did not manage to select the model with best predictive accuracy, despite the better synthetic worst-case scenario performance. However, it can be seen from Fig. 2 that the Cr model fits the training data better and that the 2P model's superior accuracy seems to stem from changing hazard rates, which is untypical for a single failure mechanism. Such hidden effects are not modeled by the synthetic data experiments.

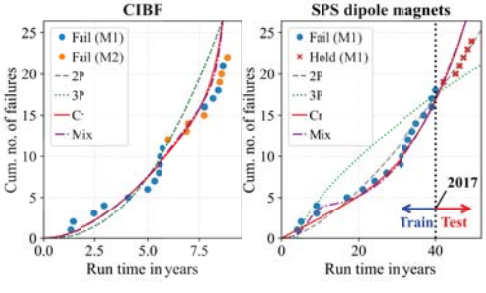

Fig. 2. Fitted Weibull models for real-world use cases.

4. Conclusion

This study compared stratified 5-fold cross-validation with AICc/BIC for model selection in the context of Weibull analysis. The model selection performance was evaluated for single 2P and 3P, as well as competing and mixture Weibull models, across different censoring levels, sample sizes and model parameters, using synthetic datasets. Results show that:

- CV exhibits the lowest predictive errors in small-sample, high-censoring scenarios across all Weibull models; AICc/BIC yield comparable predictive performance for larger samples with less censoring.
- Performance for identifying the true underlying model differs with the type of Weibull model, with BIC performing best for the most parsimonious model (2P), CV winning for 3P models, and AICc for competing/mixture models.
- The applied $\Delta \leq 2$ rule leads to better model selection performance for AICc, but not for BIC; for CV the corrected resampled t-test proved overly sensitive and led to worse results, whereas applying a delta of 0.1 provided better performance.

The methods were also applied to highly censored failure data from two CERN accelerator systems: the power supplies of the Large Hadron Collider's Beam Interlock System (CIBF) for true model selection, and SPS dipole magnets for predictive accuracy. For the CIBF, the observed lack of consistent model selection aligned with the synthetic findings. However, the synthetic setup did not capture all real-world effects, such as the atypical hazard changes observed in the SPS data. Consequently, the findings should be verified and refined across additional real-world scenarios. In

addition, future research should identify improved selection rules for CV, as additional performance gains seem possible.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Bermejo, I. and S. Grimm (2024). MSR17 Can machine learning support survival model selection to inform economic evaluations? Exploring k-fold cross validation based model selection in seven datasets. *Value in Health 27*(12), S441.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8*, 3–62.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research 33*(2), 261–304.

CERN (1976). The 400-GeV proton synchrotron. Technical Report CERN/SIS-PU 76-02, European Organization for Nuclear Research, Geneva, Switzerland.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society: Series B (Methodological) 21*(2), 411–421.

Harter, H. L. and A. H. Moore (1965). Maximum-likelihood estimation of the parameters of gamma and Weibull populations from complete and from censored samples. *Technometrics 7*(4), 639–643.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd ed.). Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, Volume 2, Montreal, Quebec, Canada, pp. 1137–1143. Morgan Kaufmann Publishers Inc.

Lai, C.-D. (2014). *Generalized Weibull Distributions*. SpringerBriefs in Statistics. Springer.

Liu, S. and D. J. Zhou (2024). Using cross-validation methods to select time series models: Promises and pitfalls. *British Journal of Mathematical and Statistical Psychology 77*(2), 337–355.

Mendenhall, W. and R. J. Hader (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika 45*(3/4), 504–520.

Myung, J. I., Y. Tang, and M. A. Pitt (2009). Chapter 11 evaluation and comparison of computational models. In *Computer Methods, Part A*, Volume 454 of *Methods in Enzymology*, pp. 287–304. Academic Press.

Nadeau, C. and Y. Bengio (2003). Inference for the generalization error. *Machine Learning 52*(3), 239–281.

Puccio, B., A. Castañeda Serra, M. Kwiatkowski, I. Romera Ramirez, and B. Todd (2010). The CERN Beam Interlock System: Principle and operational experience. In *Proceedings of the 1st International Particle Accelerator Conference (IPAC'10)*, Kyoto, Japan, pp. 2866–2868. JACoW. Paper WEPEB073.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology 25*, 111–163.

Reid, M. (2025). Reliability – a Python library for reliability engineering. [Computer software], Version 0.9.0. Zenodo. DOI: 10.5281/zenodo.3938000.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Takezawa, K. (2018). A simulation study for the AIC and likelihood cross-validation: The case of exponential versus Weibull distributions. *Journal of Advances in Mathematics and Computer Science 28*(3), 1–13.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics 18*(3), 293–297.