A Machine Learning Approach to Enhance the Information on Suspensions in Life Data Analysis

Tamer Tevetoglu[a], Martin Dazer[a]

[a] Institute of Machine Components, University of Stuttgart, Stuttgart, Germany,
tamer.tevetoglu@ima.uni-stuttgart.de

Abstract: Increasing digitalization and implementation of sensors in systems result in high data availability, which enables and benefits data-driven approaches. Commonly, these approaches revolve around predictive maintenance, anomaly detection, or clustering. In this paper, we analyze the practicality and performance of life data analyses based on neural networks for a finite population. To this end, the Weibull analysis is extended with a machine learning approach and compared with conventional approaches.

1. INTRODUCTION AND MOTIVATION

Reliability engineers usually have budget and time constraints regarding testing strategies. These constraints manifest as an inability to accurately verify a system's reliability with a pre-defined confidence due to small sample sizes, insufficient number of failures from testing, or inadequate choice of life data analysis methods. Conventional approaches in life data analysis counteract these constraints by taking suspensions into account or allowing to correct the bias when computing parameter estimates and confidence bounds. Hence, engineers only have limited number of tools to deal with constraints in reliability testing.

Previous studies have shown that these counteracting measures may not be effective under certain circumstances, i.e. despite taking suspensions into account or using bias-corrections, parameter estimates may differ substantially from the ground truth [1]–[3]. This may lead to a false sense of security regarding the operational life of a product. As data-driven approaches become steadily more important in other reliability engineering areas, e.g. Prognostics and Health Management (PHM), the focus of this paper lies on the analysis whether some shortcomings in life data analysis can be mitigated by using data-driven approaches in addition to or instead of conventional approaches. We train a data-driven model that uses a neural network to recognize patterns in sequences of data, e.g. numerical times series data emanating from sensors. A trained model can output the remaining useful lifetime (RUL) of a system based on sequential sensor data like temperature, vibration, etc. In life data analysis, failures and their respective sensor data can be used to train data-driven model. This trained model is then being used to predict the RUL of the suspensions.

If the predicted failure times are close to the unknown real failure times of the suspensions, one may use the predicted failures in addition to the actual failures or use the predicted suspensions times (instead of the initial suspension times). Ideally, these predicted RULs help to obtain more accurate parameter estimates and confidence bounds and also save testing time since not all samples need to be tested until failure. In order to evaluate this proposition, we conduct a study to determine how using neural networks to increase the number of failures (or more accurate suspension time estimates) by predicting the RULs of suspensions actually performs against conventional approaches like bias-corrections. For this purpose, we use a turbofan engine data set from NASA [4] and compare the performances of three Weibull analysis approaches to each other:

1. Maximum likelihood estimation (MLE) with bias-corrections
2. MLE with machine learning
3. MLE with machine learning and bias-corrections

For each approach, the coverage probability of the confidence bounds is evaluated at the B_{10}-life for a censored subset of the NASA data set (considered as a finite population) with varying sample sizes. The first approach is based on the MLE in combination with likelihood-ratio bounds and the Hirose and Ross (HR) bias-correction. This bias-correction method performed best in a previous study [2]. The second approach requires training a machine learning model with actual failures and subsequent prediction of the suspensions' RULs. Then a conventional Weibull analysis is conducted with the actual and predicted failures. The data-driven model we use in this paper is the Weibull Time To Event Recurrent Neural Network (WTTE-RNN) [5]. The third approach includes a subsequent bias-correction after using the machine learning model.

Based on this simulation study, this paper's main objective is to conclude on whether the use of neural networks can mitigate above mentioned shortcomings, and if so, what the precise prerequisites are. These prerequisites include the sample size, number of failures, number of suspensions, censoring share, and choice of methods. Our special attention lies on the definition of use cases, where such analyses would benefit from additional data-driven approaches.

This paper addresses the following research questions:

- What are the requirements (e.g. minimum sample size) for a data-driven approach in the context of life data analyses?
- Can we generally use WTTE-RNN to predict RULs in life data analyses, and if so, does it help to obtain more accurate confidence bounds?
- Can we treat predicted RULs as actual failure time or does treating them as suspensions result in more accurate estimates?
- Does the combination of WTTE-RNN and bias-correction methods perform better than conventional approaches?
- How does a finite population affect the coverage probability in our simulation study?

2. SIMULATION STUDY

In this chapter, we provide the needed information regarding the simulation study. After a brief overview of the current state of the art regarding lifetime analyses, the NASA data set and WTTE-RNN are presented. Lastly, the simulation setup is explained.

2.1. Previous Work and State of the Art

Likelihood ratio bounds (LRB) are a frequently used type of confidence bounds in reliability engineering and generally recommended for small sample sizes. It is based on the results of the maximum likelihood estimation, which has failures and suspensions from testing or field operation as input data. As a result, the accuracy and precision of the LRB are directly dependent on the Weibull parameter estimates from the MLE. Small sample sizes or number of failures result in biased Weibull parameter estimates, since the MLE is only asymptotically unbiased as the sample size increases. Hence, the confidence bounds are also biased as well. While assessing testing results or field data to see whether the reliability specifications are met, biased parameter estimates and confidence bounds could lead to a false sense of security regarding the operational life of a product.

There are many papers regarding the accuracy of ML estimates for the Weibull distribution. References [6] and [7] show that bias-corrected ML estimates are more accurate than noncorrected estimates. But [3] points out that bias-correction methods may increase the mean squared error (MSE) of the estimate, thus making the estimated parameter bounds become less accurate. These less accurate parameter bounds directly affect the accuracy of confidence bounds. We have shown that more accurate (less biased) Weibull parameter estimations do not necessarily results in more accurate confidence bounds [1], [2]. The effectiveness of bias-corrections is heavily dependent on the data type (uncensored or censored (censoring share), sample size, population characteristics, etc.).

The current state of the art counteracts the drawbacks of small sample sizes and biased estimates in two ways ([6], [8], [9]):

1. The maximum likelihood estimation allows to take failures and suspensions into account. For type II right-censored data, all suspension times are equal to the last actual failure time.
2. Bias-correction methods based on data from Monte-Carlo studies.

As mentioned above, these methods do not work well under certain circumstances, e.g. high censoring shares in the data or a relatively small finite population in an enumerative study. Enumerative studies make conclusions about the population from which the sample was drawn [10]. Most research focuses on analytical studies, where the population is theoretically infinite, and samples allow conclusions on a future population. Therefore, we address the problems of small sample sizes in finite populations in this paper with a data-driven RUL prediction of suspension data and unbiasing methods.

2.2. NASA Data Set and WTTE-RNN

The data set for the simulation study is the Turbofan Engine Degradation Simulation Data Set. More specifically, we used the train_FD0001 subset [11]. The data set contains the results of simulations of aircraft turbines. The simulation program C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) was used, which is why the above dataset is referred to as the C-MAPSS dataset. The dataset contains the sensor data and operational settings of the simulated aircraft turbines. For each of the aircraft turbines, data was collected from 21 sensors (e.g. temperature, pressure, etc.) and three operational settings. For each of the turbines included in the data set, no damage is present at the beginning of the data collection. Only as the data record progresses does each of the simulated turbines develop a fault until the data collection ends with the failure of that turbine. The number of the last recorded cycle represents the lifetime of a turbine. Table 1 shows an excerpt of the data set. Each turbine has a specific ID and a cycle count. The very last cycle count represents the end-of-life of each turbine, e.g. turbine 1 has a lifetime of 192 cycles.

Table 1: Turbofan Engine Degradation Simulation Data Set

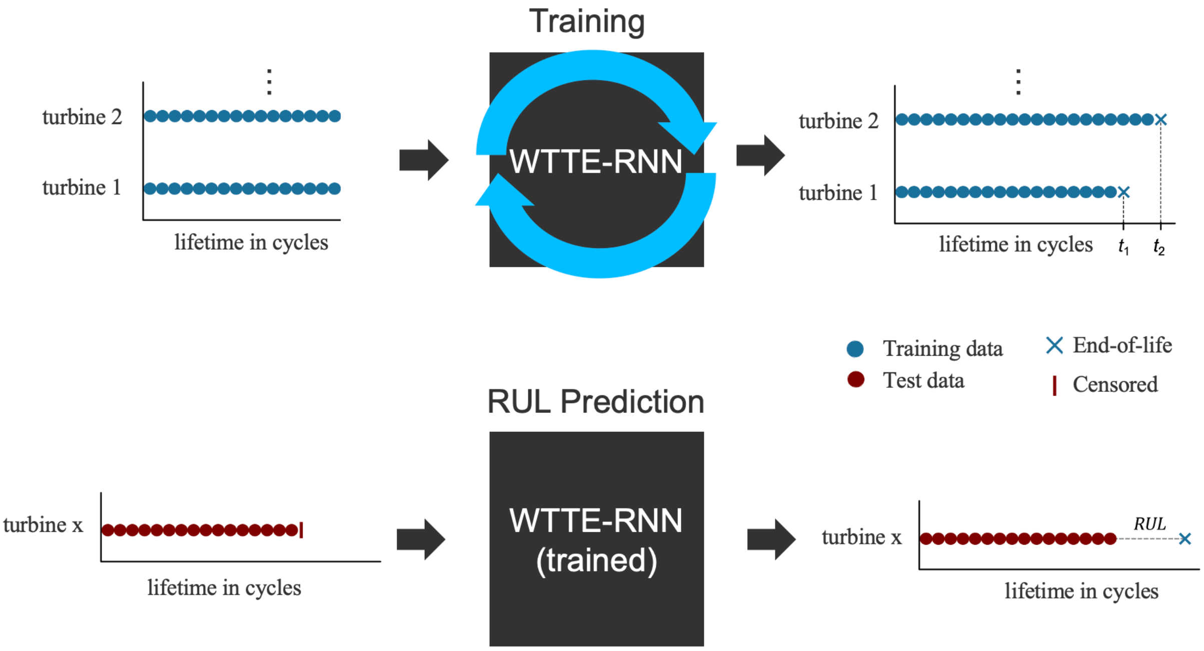
TURBINE ID	**CYCLE**	**SETTING 1**	**…**	**SETTING 3**	**SENSOR 1**	**…**	**SENSOR 21**
1	1	-0,0007	…	100	518,67	…	234.190
1	2	0,0019	…	100	518,67	…	234.236
1	3	-0,0043	…	100	518,67	…	233.442
			⋮				
1	190	-0,0027	…	100	518,67	…	230.675
1	191	0,0000	…	100	518,67	…	231.295
1	192	0,0009	…	100	518,67	…	229.649
2	1	-0,0018	…	100	518,67	…	234.585
2	2	0,0043	…	100	518,67	…	234.085
2	3	0,0018	…	100	518,67	…	234.250
			⋮				
2	285	-0,0007	…	100	518,67	…	231.155
2	286	-0,0010	…	100	518,67	…	230.169
2	287	-0,0005	…	100	518,67	…	230.848
3	1	0,0008	…	100	518,67	…	233.205
3	2	-0,001	…	100	518,67	…	234.369
3	3	0,0013	…	100	518,67	…	233.162
			⋮				

This specific subset contains data for 100 turbines until failure. All turbines fail due to wear of the high-pressure compressor (HPC).

WTTE-RNN

Weibull Time To Event Recurrent Neural Network is a method for predicting future events, such as system failures. It is a machine learning model that uses recurrent neural networks (RNN) with long short-term memory for backpropagation. These networks can handle sequential data, like the NASA data set used in this paper, very well. The developers used the very same NASA turbofan data to validate the WTTE-RNN model and its concept. Hence, we can assume that the trained WTTE-RNN results in accurate RUL predictions, and therefore we did not modify layers, number of neurons, etc. of the model. We solely optimized some hyperparameters regarding the loss decrease during training, i.e. the batch size. The model is free and available for Python [12].
As shown in figure 1, the model needs to be trained before using it for predictions. The model learns to predict the RUL by using a subset of the available data set (training set, blue circles). Each blue circle represents one cycle of the turbine containing the sensor data. WTTE-RNN iterates multiple times over this training set and decreases its loss function in order to predict RULs accurately.

Figure 1: WTTE-RNN Training and RUL Prediction



After training the model and evaluating its accuracy, we can now use WTTE-RNN for predictions on the test subset of our data (red circles). For this purpose, we censor the test data set (type II right-censored) and provide the signal data until the censoring time to the model. WTTE-RNN then predicts the RUL based on the given signal data. Thus, we can compute each end-of-lifetime for suspensions in our data set given that we have actual failures to train the model.
Please check [5] for a thorough explanation on how WTTE-RNN works.

2.3. Assessment Criteria

The B_{10}-life's coverage probability (*CP*) is used as an assessment criterion for evaluating the different approaches mentioned in chapter 1. The coverage probability is the frequentist probability that the confidence interval contains the ground truth statistics. As mentioned above, we solely compute two-sided likelihood ratio bounds with a confidence level of $1 - \alpha = 0.9$, where α is the significance level. Therefore, the coverage probability

$$CP = \frac{n_\mathrm{b}}{n_\mathrm{t}} = 1 - \hat{\alpha} \tag{1}$$

is the Monte Carlo approximation of $1-\alpha$, where n_b is the frequency of how often the confidence intervals contain the true B$_{10}$-life and n_t is the number of simulation trials, i.e. the number of randomly drawn samples from the NASA data set.

In this paper, the true value is the B$_{10}$-life on the ground truth Weibull curve. In order to calculate the ground truth references using the MLE, all 100 failure times of the finite NASA data set are considered to compute the reference statistics assuming a two-parameter Weibull distribution (see table 2).

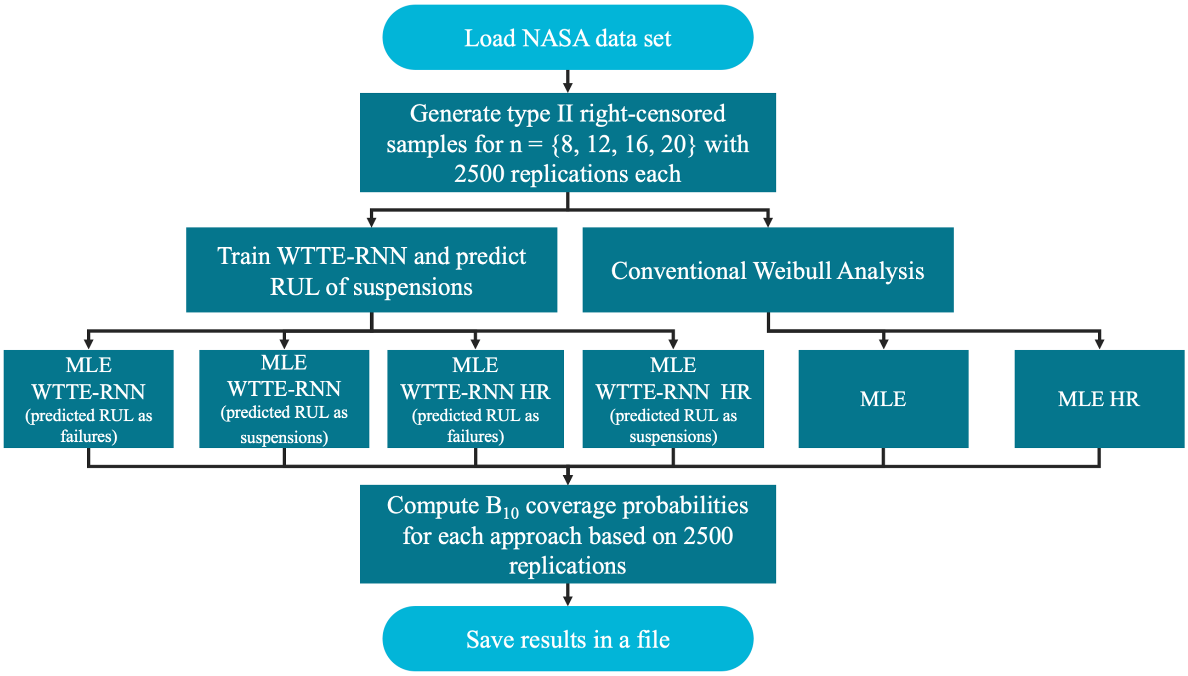
Table 2: Ground Truth Statistics

B$_{10}$	**Weibull Shape Parameter β**	**Weibull Scale Parameter η**
135.06 cycles	4.41	225.03 cycles

2.4. Simulation Setup

The first steps involve loading the complete NASA data set (train_FD0001 subset) to the workspace and generate type II right-censored samples with 2500 replications each. The censoring share is set to 25% in order have enough training data for the WTTE-RNN in each sample. Previous studies showed that the MLE performs well for sample sizes greater than 20 (with or without bias-corrections, and independently from the chosen confidence bounds methods) [1], [2]. Therefore, the samples sizes in this simulation study range from 8 to 20. Also, these samples sizes represent a more realistic scenario as long testing time and large numbers of test specimens are expensive. Figure 2 illustrates the sequential steps involved in the simulation study.

Figure 2: Flowchart: Simulation Study

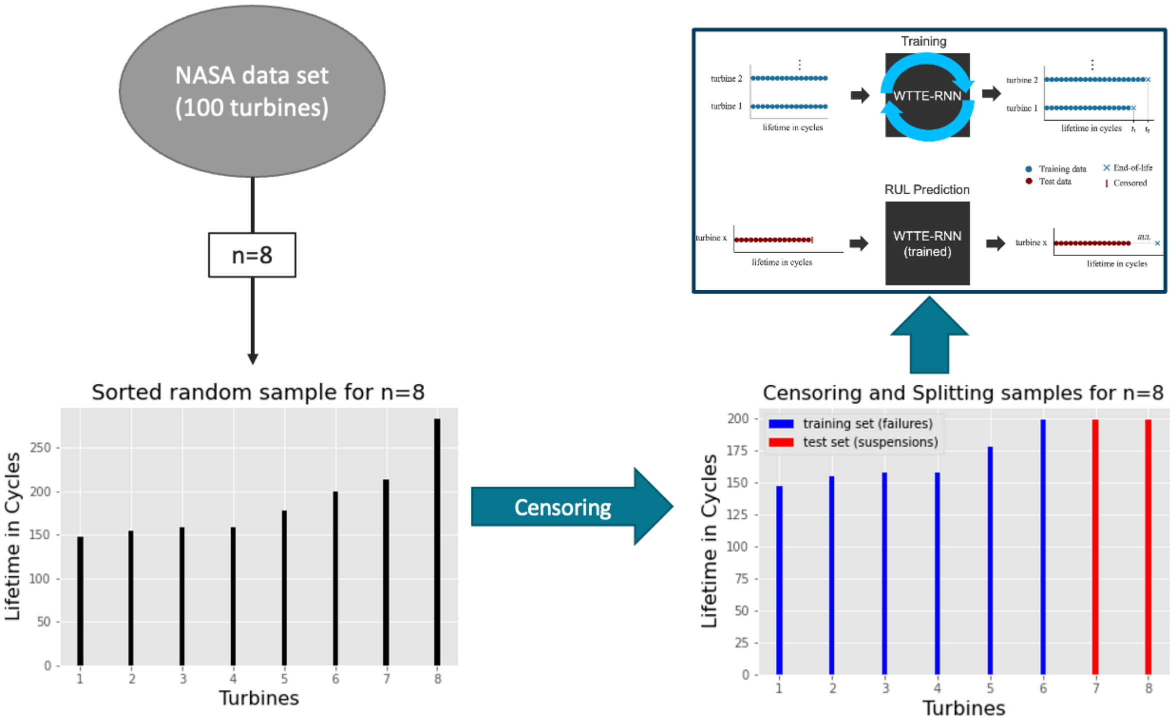

Load NASA data set
Generate type II right-censored samples for n = {8, 12, 16, 20} with 2500 replications each
Train WTTE-RNN and predict RUL of suspensions
Conventional Weibull Analysis
MLE WTTE-RNN (predicted RUL as failures)
MLE WTTE-RNN (predicted RUL as suspensions)
MLE WTTE-RNN HR (predicted RUL as failures)
MLE WTTE-RNN HR (predicted RUL as suspensions)
MLE
MLE HR
Compute B10 coverage probabilities for each approach based on 2500 replications
Save results in a file

For both WTTE-RNN and conventional approaches, the MLE with likelihood ratio bounds is being used. The LRB are two-sided and have a set confidence interval of 90%. The conventional Weibull analysis involves:

- MLE: This approach uses the censored sample as an input and computes the LRB without bias-corrections
- MLE HR: In addition to the MLE, the Hirose and Ross (HR) bias-correction is applied.

Using WTTE-RNN requires to split the generated samples into a training and test set. The training set is based on the actual failures, and the test set (for which we want to predict the RUL) consists of the sample's suspensions. Figure 3 shows how to generate a training and test data set for a random sample.

Figure 3: Training the WTTE-RNN and Predicting the RUL of Suspensions



The WTTE-RNN Weibull analysis involves:

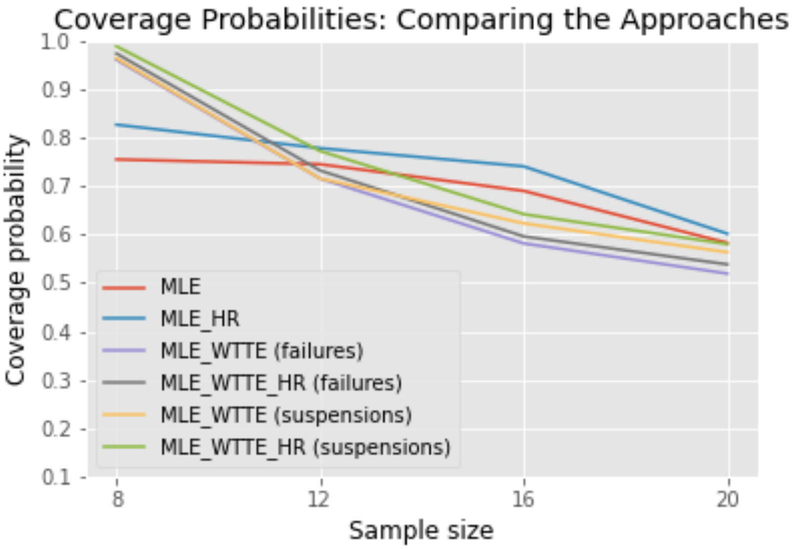
- MLE WTTE-RNN (suspensions): The predicted remaining useful lives of the suspensions are added to the suspensions (censoring) time to compute the end-of-life time of the suspensions (see figure 1). In this approach, the new, predicted times will be used as suspensions in the MLE.
- MLE WTTE-RNN HR (suspensions): Same approach as above, however the MLE will be bias-corrected using the Hirose and Ross method.
- MLE WTTE-RNN (failures): The predicted remaining useful lives of the suspensions are added to the suspensions (censoring) time to compute the end-of-life time of the suspensions. In this approach, the new, predicted times will be used like actual failure times in the MLE.
- MLE WTTE-RNN HR (failures): Same approach as above, however the MLE will be bias-corrected using the Hirose and Ross method.

Each approach outputs 2500 estimated B_{10}-life confidence limits, which are then compared the ground truth B_{10}-life (see table 2). The last step is to save and interpret the results of the simulation study.

3. Results

Figure 4 shows the coverage probabilities of the B_{10}-life for different sample sizes. All WTTE_RNN perform better than the conventional approaches for n = 8. As expected from previous studies [1], [2], bias-corrected conventional Weibull analyses result in a higher coverage probability than non-corrected ones.

Figure 4: Coverage Probabilities



For n = 12 all approaches perform the same. However, the *CPs* of conventional approaches don't decrease as fast as the WTTE-RNN approaches as the sample size increases. This results in higher *CPs* of the conventional approaches for n = 16. For n = 20, conventional approaches still perform better. However, all approaches perform pretty much the same again.

In contrast to infinite populations in analytical studies, a decreasing coverage probability in finite population is not unexpected in enumerative studies (see chapter 4 for a discussion). If you just look at the data for the WTTE approach, it is evident that bias-corrected MLE with predicted suspension times perform best throughout all generated sample sizes. As the sample size increases, the performance difference between WTTE-RNN (failures) and WTTE-RNN (suspensions) increases.

4. DISCUSSION AND CONCLUSION

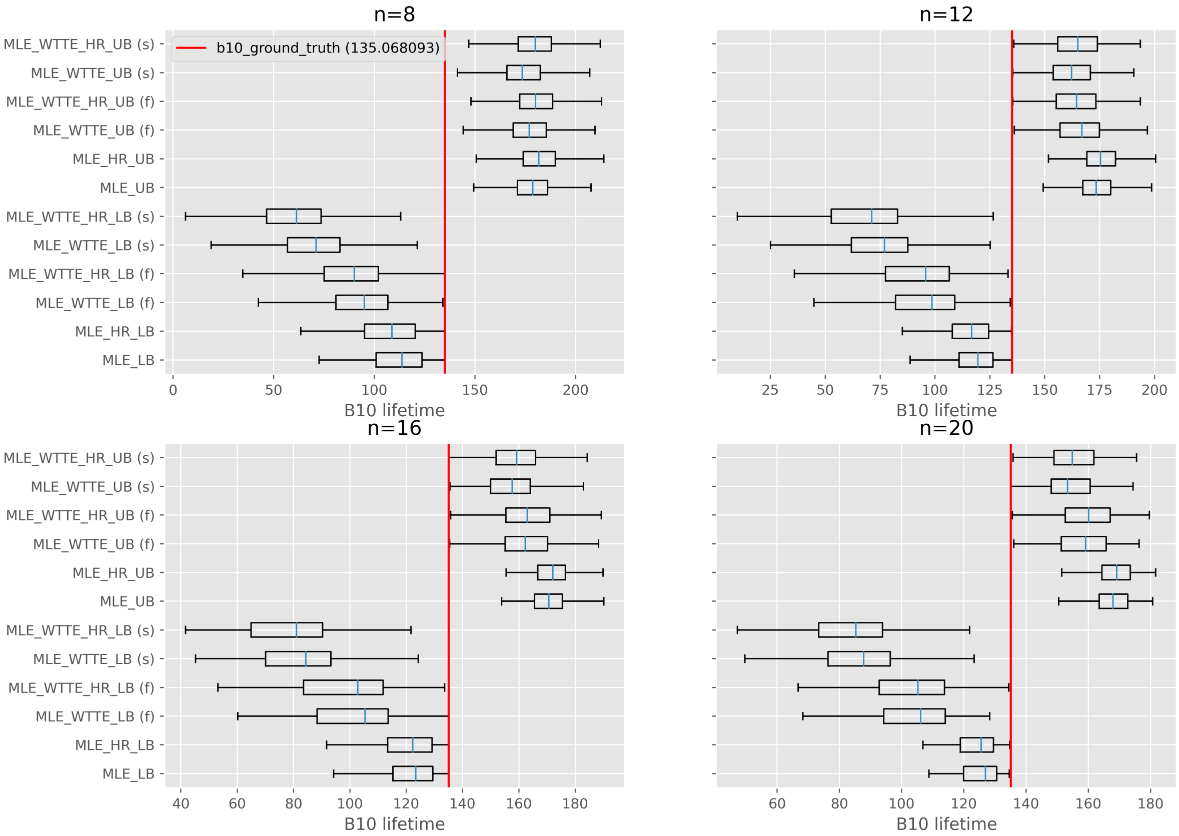
It is important to understand why the approaches vary so much in performance. Figure 5 shows the distribution of the lower and upper bounds around the ground truth (red) for different sample sizes. Solely bounds that cover the ground truth B_{10}-life are presented as boxplots. Lower bounds (LB) are located left from the ground truth and upper bounds (UB) are located right from the ground truth. We will discuss the lower and upper bound deviation separately.

As the sample size increases, the WTTE-RNN approaches generally deviate closer around the ground truth than conventional approaches. Yet there is no significant difference between all approaches. WTTE-RNN that use the RUL as predicted suspension times are generally the closest to the ground truth for n > 8.

Usually, the focus in lifetime analysis lies on the lower bounds as it is essential to draw conclusions on the minimum reliability or lifetime that can be guaranteed (with a certain significance). There are some significant differences:

- Lower bounds of conventional approaches deviate closer around the ground truth
- WTTE-RNN (f) using predicted RUL as actual failure data are significantly closer to the ground truth than WTTE-RNN (s)
- Overall, conventional approaches have less deviating lower and upper bounds
- As the sample increase, the deviation for all approaches decreases

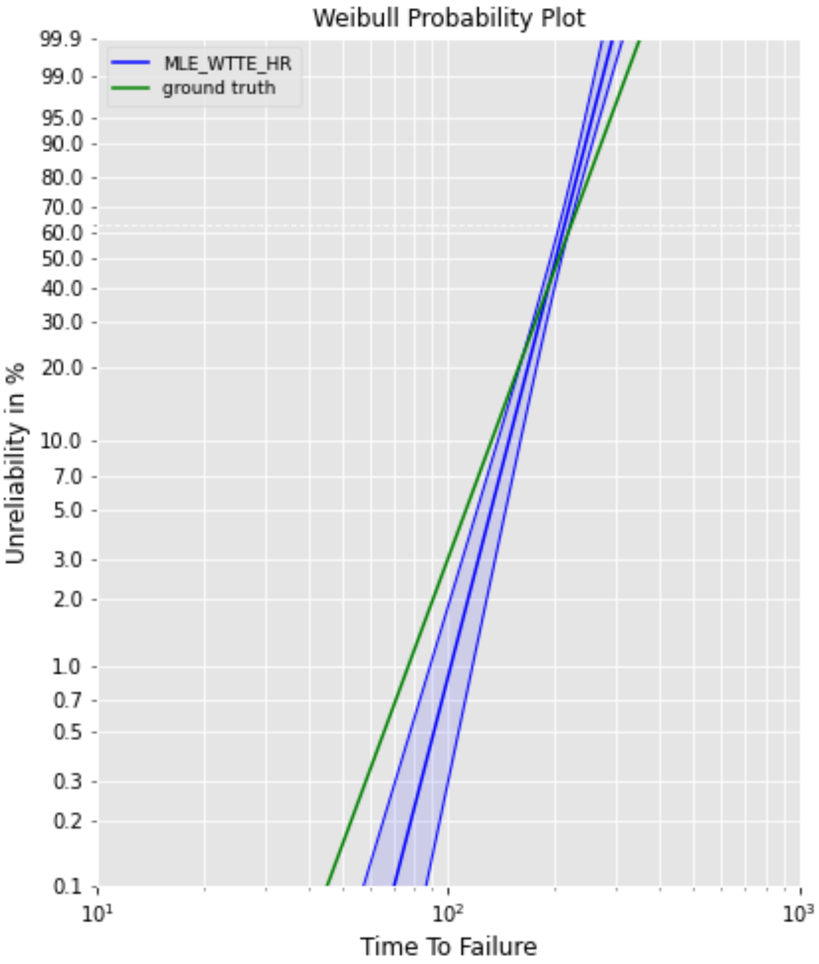
Figure 5: B_{10} Deviation for Different Sample Sizes



When we try to understand the systematic differences which affect the results shown in figure 4, the deviations might help to answer some unclarities:

- When we deal with populations (finite or infinite) in an enumerative study, generally the confidence intervals get smaller as the sample size increases. Since we have more information in form of failures and suspensions, the MLE estimates smaller variances (which are the basis for the confidence bounds). However, if the parameter estimation (Weibull shape and scale parameters) is just slightly biased, the bounds could miss the true B_{10}-life. Figure 6 shows the bias-variance tradeoff for this finite population drawback. The green line represents the ground truth with 100 failures. The blue line is computed using a sample size of 80 (60 failures, 20 suspensions). As can be seen, the confidence bounds seem close to the ground truth but still miss it. Hence, a smaller estimated variance does not necessarily perform better. This could also explain the decreasing *CP* as the sample size increases when dealing with finite populations.
- WTTE-RNN requires training data for training. For small sample sizes (e.g. n= 8) the loss of the model is large, i.e. the model does not have enough data to properly train and this results in large deviations when predicting RUL. Nevertheless, the high variance in estimation based on these predicted RUL results in wider confidence bounds that manage to cover the ground truth B_{10}-life. WTTE-RNN would be a conservative approach as it systematically underestimates the lower bound. It is up to the individual reliability engineers to decide if they want to accept big variances in estimation to be more on the safe side when computing the lower lifetime bound.
- The data shows that using the predicted RUL as suspensions rather than as failures in the MLE performs better. The slightly less deviating lower bound of the WTTE-RNN (f) is insignificant.
- Bias-corrections lead to wider confidence bounds throughout all sample sizes in this study. This reinforces the tendency to estimate larger variances (and therefore confidence bounds) when using the MLE.

Figure 6: Bias-Variance Tradeoff: Finite Populations



The results of the simulation study show that the study type (enumerative or analytical) and population type (focus of this paper: finite) are very important regarding the coverage probability. This paper shows that the *CP* decreases for a finite population in an enumerative as the sample size increase. Previous work shows that the *CP* increases with the sample size for infinite population in an analytical study [1], [2].

WTTE-RNN can help to increase the CP for small sample sizes, However, the above-mentioned drawbacks need to be considered. If one must decide on how to use the predicted information (RUL) in life data analysis, treating them as suspensions should be the choice when using the MLE.

Outlook

This paper shows that there is a high potential in optimizing data-driven approaches that are used in Weibull analysis. Therefore, future research work should focus on the bias-variance trade-off when using hybrid approaches like in this simulation study. This could help to deal with the drawbacks of finite populations and sampling errors. Future data-driven models in life data analysis should find the sweet spot between prediction accuracy and variance.

Based on this work, the authors will

- develop a new data-driven approach to address the above-mentioned drawbacks, and
- conduct simulation studies that solely focus on finite populations in enumerative studies to have some general recommendations regarding estimation accuracy.

References

[1] T. Tevetoglu and B. Bertsche, “Bias Corrected Weibull Parameter Estimation and Impact on Confidence Bounds,” 2020.

[2] T. Tevetoglu and B. Bertsche, “On the Coverage Probability of Bias-Corrected Confidence Bounds,” in *2020 Asia-Pacific International Symposium on Advanced Reliability and Maintenance Modeling (APARM)*, Aug. 2020, pp. 1–6. doi: 10.1109/APARM49247.2020.9209464.

[3] U. Genschel and W. Q. Meeker, “A Comparison of Maximum Likelihood and Median-Rank Regression for Weibull Estimation,” *Qual. Eng.*, vol. 22, no. 4, pp. 236–255, Aug. 2010, doi: 10.1080/08982112.2010.503447.

[4] NASA, *NASA/SP-2011-3422 NASA Risk Management Handbook*. 2011. [Online]. Available: https://www.researchgate.net/publication/314158409_NASA_Risk_Management_Handbook

[5] E. Martinsson, “WTTE-RNN : Weibull Time To Event Recurrent Neural Network,” CHALMERS UNIVERSITY OF TECHNOLOGY, 2017. [Online]. Available: https://publications.lib.chalmers.se/records/fulltext/253611/253611.pdf

[6] R. B. Abernethy, *The new Weibull handbook : Reliability & statistical analysis for predicting life, safety, risk, support costs, failures, and forecasting warranty claims, substantiation and accelerated testing, using Weibull, Log normal, crow-AMSAA, probit, and Kapla*. R.B. Abernethy, 2006.

[7] M. Chen, Z. Zhang, and C. Cui, “On the Bias of the Maximum Likelihood Estimators of Parameters of the Weibull Distribution,” *Math. Comput. Appl.*, 2017, doi: 10.3390/mca22010019.

[8] H. Hirose, “Bias correction for the maximum likelihood estimates in the two-parameter Weibull distribution,” *IEEE Trans. Dielectr. Electr. Insul.*, vol. 6, no. 1, pp. 66–68, 1999, doi: 10.1109/94.752011.

[9] M. F. Barmoav, “Reduced Bias Factor Distribution to reduce the likelihood estimate bias of small sample sizes,” in *2010 Proceedings - Annual Reliability and Maintainability Symposium (RAMS)*, Jan. 2010, pp. 1–7. doi: 10.1109/RAMS.2010.5448021.

[10] D. C. Montgomery, *Engineering Statistics*, Fifth Edit. John Wiley & Sons, Inc., 2010.

[11] “Turbofan Engine Degradation Simulation Data Set.” https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/ (accessed Dec. 03, 2021).

[12] E. Martinsson, *WTTE-RNN*. Accessed: Jan. 02, 2022. [Online]. Available: https://github.com/ragulpr/wtte-rnn