Wind Turbine Recovery Forecasting using Survival Analysis

Anton Palets

Master's thesis
2023:E67

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Wind Turbine Recovery Forecasting using Survival Analysis

Master's Thesis

Anton Palets

Supervisor: Dragi Anevski
Examiner: Andreas Jakobsson
Department of Mathematics
Division: Mathematical Statistics
June 2023

Abstract

The goal of this thesis is to present a methodology for predicting time until recovery of failed wind turbines. The necessity is motivated by the potential for more accurate wind energy export forecasts. The current approach rests entirely on having an expert examine the turbine and produce a time estimate. Due to its nature, such a prediction cannot be made immediately upon failure. Five common survival analysis models are evaluated in regard to their ability to correctly classify recovery as happening within or after 24 hours of failure, and point prediction error in the case that the failure event is predicted to resolve within 24 hours. A method for nonparametric clustering of survival curves is developed, that is used to reduce the number of variables in the examined models. The Weibull Accelerated Failure Time model with the clustered error codes and logarithm of energy produced in the month prior to failure is found to perform significantly better than alternatives. Classification is optimized by finding optimal thresholds using ROC curves. An attempt is made to present theory necessary to motivate the models used.

Acknowledgments

I would like to extend the deepest gratitude to my Power Factors supervisor Pramod Bangalore, who guided me along the way and helped me stay focused on the goal and not get lost in the confusion surrounding doing research. I would also like to thank Power Factors in general for providing access to data which would be unobtainable without them. I would like to thank Edmund Hood Highcock and Margarida Nabais in Power Factors for sacrificing their time to help me with the tedious setup of data access.

I would also like to thank Dragi Anevski, my Lund University supervisor. Without his guidance, the work would likely have major theoretical oversights. His stamp of approval allowed me to focus my effort on furthering the work instead of on double-checking the overall reasonableness of methodology. I would also like to thank Andreas Jakobsson for dedicating his time to examine my thesis and providing suggestions on how to further improve the work.

Finally, I would like to thank my parents and my girlfriend for supporting me throughout the process of writing this thesis. Their support was critical in staying focused on the project.

Popular Scientific Summary

This Master's thesis in Mathematical Statistics has as a goal the development of a model that can reliably predict whether a broken wind turbine will recover within 24 hours or not, and if it will, in how many hours. We opt to develop a methodology that could be readily applied to data different from that examined in our work. In particular, this means that the data at hand could be approached with simpler models, but we opt to give it a more general treatment, as our data is simply an example of how a turbine failure and recovery dataset can look like. Such preferences lead us to the choice of survival analysis as a preferred framework. Survival analysis is a statistical field that addresses any time-to-event data, originally getting its name from the event being death in clinical trials.

In our case, the event will be recovery of a turbine; nonetheless, we opt to keep all terminology standard, so what intuitively would be called a 'probability of recovery' is still referred to as 'survival probability'. As the goal is prediction on unseen data, we pursue a Machine Learning approach where the data available is randomly split into a training and test dataset. The training data is used to fit the models and any associated optimizations, whereas the test data is used for evaluation. This approach has the benefit of us knowing what the true recovery times are, while the fitted models predict it.

In this work we will examine five different models. These models vary substantially in their nature, which considerably reduces the shared common ground their performance could be evaluated on. In more formal terms, the goal is to first *classify* a recovery event into two categories: recovery within 24 hours of failure, and after 24 hours of failure. If the classification is that the recovery will occur within 24 hours, we would like to know a more precise value – this we refer to as the *point prediction* (of recovery time). The two problems are coupled as the point prediction is only of interest based on the classification, so they will be evaluated together. To evaluate the classification we will use two common metrics – sensitivity and accuracy, where the former is a good indicator of how a model deals with false negatives (recovery events that were predicted to resolve within 24 hours, but in reality recovered after 24 hours). For the point prediction, we will use mean absolute error, which is preferred over mean squared error in this case as it does not give additional weight to predictions that were far off from the true value.

As is usual for any survival analysis dataset, the two categories we will classify into are very unbalanced. What this means is that the number of turbines that recover within 24 hours is substantially higher than those that do not. This presents additional challenges with classification, which we address by relatively standard means – through the study of ROC curves and optimal thresholds. This approach allows us to balance the classification in our situation where one of the categories is inherently more likely.

We use several variables to model the recovery time distribution. The one always used is the

error codes, which enter into the models not in their raw form but in clusters. Error codes are unique representations of potential final causes of failure, and each failure event in our dataset is accompanied by one of these. The use of clusters of codes instead of codes directly allows us to reduce the number of variables in the subsequent models, and is justified by the reasonable assumption that recovery from many problems is similar enough as to not warrant distinguishing. We develop a clustering method that allows us to group the codes in a way that does not impose unwanted assumptions. Another variable we use is the logarithm of the energy produced by the turbine in the month before it failed. The assumption behind this variable is that this value could act as a proxy for wear and tear which could ultimately impact the recovery time. Alternatively, well performing turbines could see priority treatment, and the recovery could be accelerated. Finally, we investigate the impact of past failure history on present recovery. For each turbine and for each cluster of failures, a counter is constructed recording the number of times a failure in the cluster happened to the given turbine.

With all of the above, the work attempts to develop the theory to justify the models and methods used. This results in a fairly lengthy theoretical section, most of which can be skipped by those that find no inherent interest in the subject. The model we find to be best is a fully parametric model with fairly few variables, which has the benefit of quick fitting and prediction.

Abbreviations

AFT	Accelerated Failure Time
AIC	Akaike Information Criterion
AUC	Area Under the ROC Curve
BIC	Bayesian Information Criterion
CLT	Central Limit Theorem
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FWER	Family-Wise Error Rate
MAE	Mean Absolute Error
OLS	Ordinary Least Squares
PH	Proportional Hazards
QQ	Quantile-Quantile
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
TN	True Negative
TP	True Positive
TPR	True Positive Rate

Notation

T	Random variable – survival or recovery time
$S(t)$	Survival function
$h(t)$	Hazard rate
$H(t)$	Cumulative hazard rate or predictable process
M	Martingale, counting process martingale
$\{\mathcal{F}\}$	Filtration of σ-algebra $\mathcal{F}$ in discrete or continuous time
$[X]$	Optional variation process of X
$[X_1, X_2]$	Optional covariation process between X_1 and X_2
$\langle X \rangle$	Predictable variation process of X
$\langle X_1, X_2 \rangle$	Predictable covariation process between X_1 and X_2
$H \bullet X$	Transformation of X by H
$N(t)$	Counting process
$\lambda(t)$	Intensity process of counting process
$N_\cdot, \lambda_\cdot, M_\cdot$	Counting process, intensity process, and martingale summed over a relevant index
$W(t)$	Wiener process
$U(t)$	Gaussian martingale
$J(t)$	Indicator taking on value 1 if number at risk at time t is not zero

Contents

Introduction

Plenty of research has been devoted to wind power forecasting [1][2][3], which is necessary in part due to the way electricity markets are structured. The market operates on a day-ahead commitment principle, where a turbine has to be declared available to generate electricity 24 hours in advance. The problem of forecasting wind energy production can be split in two – forecasting of weather, most importantly wind speed, and forecasting of turbine availability. Turbine availability in this context is simply the ratio of turbines available to produce electricity to all turbines, and as such can be impacted by a number of things. For example, Teng et. al. [4] focus on the problem of availability as impacted by curtailment caused by grid export restrictions. Not much focus has been given to forecasting of availability as impacted by technical malfunctions currently present in turbines, with the best approach being an expert opinion of a technician, who provides an estimate based on experience. Such an approach, albeit more reliable, is not ideal, as the 'prediction' cannot be made as soon as a failure occurs. The goal of this thesis work is to investigate various data driven approaches to predict the time to recovery of a turbine after failure.

Time until event data requires a specific theoretical framework, since during model estimation a turbine might be experiencing failure, so only a lower bound of some recovery times is known. Omitting such data points from consideration would be theoretically unjust, as that would overestimate the probability of recovering. Furthermore, a prediction needs to be sound not just at the time the turbine fails, but also at any time point after, and as such needs to be able to incorporate the information on being broken up until that time point.

This view of the problem casts it into the field of survival analysis, with a wealth of theory available. Survival analysis is a field of statistics that has been developed primarily for the medical field, where survival of individuals was under consideration. Perhaps the first development was the life table – a tabulation of probability for an individual at a given age to die before their next birthday. The first life table was created by demographer John Graunt and astronomer Edmond Halley in the 17th century [5].

In 1958 Edward Kaplan and Paul Meier submitted similar papers to the Journal of the American Statistical Association, and were convinced by John Tukey, the journal's editor, to combine their work [6][7]. Their contribution is what is now known as the Kaplan-Meier estimator – a nonparametric estimator of the survival function. Meier was inspired by Major Greenwood's studies into cancer duration [8], after whom the variance estimator of the Kaplan-Meier estimator is named [9].

In 1972 Wayne Nelson suggested a nonparametric estimator for the cumulative hazard rate [10], further studied by Odd Aalen [11], now called the Nelson-Aalen estimator. The Kaplan-Meier and Nelson-Aalen estimators are closely related, where both can be derived from counting process theory and shown to be asymptotically equivalent [12][13]. This thesis takes a point process approach to derive the estimators, and it is not unnatural to treat the Nelson-Aalen estimator as the foundation, despite the popularity of the Kaplan-Meier estimator.

David Cox greatly contributed to the field in 1972 by introducing the Proportional Hazards model, which assumed a splitting of the hazard function into a covariate dependent scaling factor, and a time dependent baseline hazard equal for all individuals [14]. The beauty of the proportional hazards model is that a substantial amount of inference, like comparisons between groups, can be done without ever delving into the baseline hazard. If one would like to use the Cox proportional hazards model to generate predictions for new individuals, however, like in our case with turbines, one does require an estimator of the baseline hazard – most commonly given by the Breslow estimator [15].

Accelerated Failure Time models, named first by Nelson and Hahn in 1972 [16], provide a fully parametric alternative to the aforementioned approach. In AFT models, covariates act linearly to accelerate or decelerate the lifetime instead of Proportional Hazards' multiplicative scaling of the unspecified baseline hazard.

The application of the model from this work would be to forecast recovery times for turbines based on its own historical behavior. Generalization capability is considered an important characteristic, so a Machine Learning approach is taken where the data for model fitting and model evaluation are kept separate. Five models have been evaluated: Cox regression, Aalen's additive hazards, and Weibull, Log-normal, and Log-logistic AFT. Since this is a 'mixed bag' of non-parametric, semi-parametric, and fully parametric models, evaluation and comparison of these model is not straightforward. Each of these models have various metrics available, and residuals could be studied for each. Unfortunately, such an approach would complicate evaluation of all models, and is of debatable use, as the model fit on training data is not the primary criteria of interest.

Thankfully, the intended goal of these models offers a solution. The need for a point estimate is only necessary if the turbine is likely to recover within 24 hours. We thus first intend on classifying failure events into two categories: those that will recover within 24 hours, and those that will not. For those that will, we would also like a point prediction of time until recovery. As will be further discussed, classification is complicated since the two classes are very unbalanced. We find optimal classification thresholds for every model to address this problem. For point prediction, we will predict the median recovery time, as some models have issues with integrating over all possible recovery times. Integration would be necessary if we were to use the expected value as a predictor. Thus, to systematically compare the models we will employ classification metrics, and then compute the mean absolute error on those failure events where the recovery is predicted to occur within 24 hours.

As stated before, a good model would perform well when classifying and predicting not only immediately after a failure occurs. The ability of a model to classify and predict long after a turbine has broken initially is equivalent to the ability of the model to capture the tail of recovery time distribution well.

The thesis is organized as follows: in the section *Theory and Methods* we motivate the five models mentioned above and opt to develop the theory necessary to introduce them in reasonable depth. For parametric models this is quite simple as fitting is done through standard likelihood maximization approaches. Nonparametric and semi-parametric models require more care. To properly discuss them we introduce some theory on martingales and counting processes, in par-

ticular variation and covariation processes that are useful for obtaining variance estimators. We also discuss nonparametric tests for survival functions, a special case of which is the logrank test. Based on the logrank test we develop a nonparametric iterative clustering method. The need for clustering is motivated by the desire to reduce the number of coefficients in the models, under the assumption that recovery from many issues is similar enough to not require distinguishing. To our knowledge, the clustering approach is novel.

In the section *Experimental Results* we discuss the data at hand and motivate data preparation steps taken. General statistical properties of the recovery time and the continuous variable used in modeling are investigated. We then discuss the clustering of error codes using our nonparametric clustering method. Following this, we outline the general modelling approach taken, motivating our choices. The simplest models with the least parameters are fit and evaluated first, after which we explore if it is possible to improve prediction and classification accuracy further. To do so, we construct integer-valued 'history' variables for every possible issue that keeps track of how many times a given failure happened before the one currently underway. We then attempt to improve the models by refitting them on the variables deemed significant by a Holm-Bonferroni procedure in the preceding step. A final assessment is given by comparing the models that performed best.

Finally, in the *Conclusion*, we discuss limitations of the thesis and the general implications of our results. Many potential approaches to build upon our results are discussed, with some deserving to be their own research projects. Others yet are more so cursory remarks that would be relatively easy to implement, but were not due to scope and time considerations.

Theory and Methods

In this section, we briefly outline and develop the subject that is survival analysis. We mimic the approach in [17], taking a counting process point of view. Basic notions like survival function and hazard are presented and motivated, together with some key models that can prove fruitful in the application of interest. In particular, we will outline some key theory on martingales and variation processes. The connection of variation processes to variance will often be used to derive confidence bounds and test statistics. This discussion does not pretend to be complete, and it is worth noting that a substantial amount of complexity is put aside, especially when considering continuous time martingales. Lastly, a clustering method based on iterative nonparametric tests is introduced.

Many results on martingales and counting processes are useful to motivate nonparametric tests and non- and semi-parametric models. Many results are presented with proofs, where in source material they are given as exercises or presented as short sketches.

Basic Notions

We begin by introducing the most important notions in survival analysis, which happen to be three interconnected quantities, each in turn defining the other two.

Definition 1 (Survival function)**.** Given that a random variable T is the survival time, the survival function is $S(t) = P(T > t)$.

Definition 2 (Hazard rate)**.** Given that survival time T is absolutely continuous, i.e. has a probability density, the hazard rate $h(t)$ considers the probability of an event within a small interval $[t, t + dt)$, given as $h(t)dt$. More explicitly, we have

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t).$$

Definition 3 (Cumulative hazard rate)**.** With hazard rate as defined above, the cumulative hazard rate is given by

$$H(t) = \int_0^t h(s)ds.$$

These quantities are to some extent interchangeable, but this by no means implies that they are equally easy to estimate in all situations; some models prefer to model one over the other. Thus, it is instructive to outline the connection. Note that

$$H'(t) = h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t)$$

$$
\begin{aligned}
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{-(S(t + \Delta t) - S(t))}{S(t)} \\
&= -\frac{S'(t)}{S(t)} \\
&= \frac{f(t)}{S(t)},
\end{aligned}
$$

where $f(t) = F'(t)$, given that F is the cumulative distribution function of T, that is $F(t) = P(T \leq t) = 1 - S(t)$. If we take the integral of the second to last line above, we obtain

$$\int_0^t -\frac{S'(s)}{S(s)} ds = -\log S(t) = \int_0^t h(s) ds = H(t),$$

from which we can write an important relationship

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(s) ds\right\}.$$

Martingales and Variation Processes

Martingales will be the key to nonparametric estimators and models, as via the Doob-Meyer decomposition they naturally arise from counting processes, on which our approach rests. We omit many of the regularity conditions from discussion, as martingales themselves are not the topic of this work. The Doob-Meyer decomposition (Theorem 11) allows us to split a counting process into a predictable part and something that can be treated as a noise term. The central limit theorem for martingales will be useful to derive asymptotic properties of estimators, which allows us to construct standard Normal distribution confidence intervals once variance is found. Optional and predictable variation processes are key to finding variances of estimators, and optional and predictable covariation processes help with deriving test statistics for nonparametric tests.

All relevant results are first done in discrete time before presenting and motivating their equivalents lifted to the continuous case. We assume throughout that all involved processes are absolutely and square integrable at all times, and all following results are stated under these assumptions.

Definition 4 (Filtration in discrete time)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\mathcal{F}_n$ be a sub-σ-algebra of $\mathcal{F}$ for all $n \in \mathbb{Z}^+$, that is $\mathcal{F}_n$ is a σ-algebra and $\mathcal{F}_n \subset \mathcal{F}$. If for all $k \leq l$

$$\mathcal{F}_k \subseteq \mathcal{F}_l,$$

the collection $\{\mathcal{F}_n\}$ is called a filtration.

Definition 5 (Discrete time martingales)**.** Suppose $M = \{M_i\}$ is a process in discrete time, and assume throughout $M_0 = 0$. The process M is a martingale if

$$\mathrm{E}(M_n|\mathcal{F}_m) = M_m,$$

for all $n > m$.

To derive variance estimators, we introduce variation processes.

Definition 6 (Optional variation process)**.** The optional variation process for a martingale M in discrete time is given as

$$[M]_n = \sum_{k=1}^{n}(M_k - M_{k-1})^2 = \sum_{k=1}^{n}(\Delta M_k)^2,$$

where we let $[M]_0 = 0$.

Definition 7 (Predictable variation process)**.** The predictable variation process for a martingale M in discrete time is given as

$$\langle M\rangle_n = \sum_{k=1}^{n}\mathrm{E}\{(M_k - M_{k-1})^2|\mathcal{F}_{k-1}\} = \sum_{k=1}^{n}\mathrm{Var}(\Delta M_k|\mathcal{F}_{k-1}),$$

where we let $\langle M\rangle_0 = 0$.

Next, we present a result that is the reason behind the usefulness of variation processes. The result is from [17, p. 45], where we expand on the proof, particularly for the predictable covariation processes, which is given in source material as Exercise 2.4.

Theorem 1. *Assume that M is a mean zero martingale. Then, $M^2 - [M]$ and $M^2 - \langle M\rangle$ are also mean zero martingales.*

Proof. First, note that by assumption, $M_0^2 - [M]_0 = 0$ and $M_0^2 - \langle M\rangle_0 = 0$. To prove the first of the statements, we write

$$\begin{aligned}\mathrm{E}(M_n^2 - [M]_n|\mathcal{F}_{n-1}) &= \mathrm{E}\left\{(M_{n-1} + M_n - M_{n-1})^2\right.\\ &\quad\left. - ([M]_{n-1} + (M_n - M_{n-1})^2)|\mathcal{F}_{n-1}\right\}\\ &= \mathrm{E}\left\{M_{n-1}^2 + 2(M_n - M_{n-1}) - [M]_{n-1}|\mathcal{F}_{n-1}\right\})\\ &= M_{n-1}^2 - [M]_{n-1}.\end{aligned}$$

The second statement is proved similarly,

$$\begin{aligned}\mathrm{E}(M_n^2 - \langle M\rangle_n|\mathcal{F}_{n-1}) &= \mathrm{E}\left\{(M_{n-1} + M_n - M_{n-1})^2\right.\\ &\quad\left. - \langle M\rangle_{n-1} - \mathrm{E}\{(M_n - M_{n-1})^2|\mathcal{F}_{n-1}\}|\mathcal{F}_{n-1}\right\}\\ &= M_{n-1}^2 - \langle M\rangle_{n-1}.\end{aligned}$$

□

Note that since by the above $\mathrm{E}\, M^2 = E[M] = \mathrm{E}\langle M \rangle$, and $\mathrm{E}\, M = 0$ whenever M is a mean zero martingale, we obtain

$$\mathrm{Var}\, M = \mathrm{E}\, M^2 = E[M] = \mathrm{E}\langle M \rangle.$$

We will more often than not deal with martingales in the context of stochastic integrals. The latter are continuous time generalizations of transformations, which we introduce next.

Definition 8 (Transformation of a process)**.** Assume that $X = \{X_0, X_1, \dots\}$ is some process which is adapted to the filtration $\{\mathcal{F}_n\}$, that is X_n is measurable with respect to $\mathcal{F}_n$ for every n. Further, let $H = \{H_0, H_1, \dots\}$ be a predictable process, that is H_n is measurable with respect to $\mathcal{F}_{n-1}$ for each n. The transformation of X by H is defined as

$$(H \bullet X)_n = H_0 X_0 + \sum_{k=1}^{n} H_k (X_k - X_{k-1}).$$

If the process X is not just any process, but a martingale, we obtain a very important intermediary result. The result is taken from [17, p. 46].

Theorem 2. *If H is a predictable process, and M is a mean zero martingale, then $Z = H \bullet M$ is also a mean zero martingale.*

Proof. To show that Z is a martingale write

$$\begin{aligned} \mathrm{E}(Z_n - Z_{n-1} | \mathcal{F}_{n-1}) &= \mathrm{E}(H_n (M_n - M_{n-1}) | \mathcal{F}_{n-1}) \\ &= H_n \, \mathrm{E}(M_n - M_{n-1} | \mathcal{F}_{n-1}) \\ &= 0. \end{aligned}$$

Since $Z_0 = H_0 M_0 = 0$, we have that Z is a mean zero martingale. □

To derive variance estimators of stochastic integrals, we will use variation processes. This necessitates in particular the following result on optional variation processes of transformations, presented as Exercise 2.7 in [17, p. 47].

Theorem 3 (Optional variation process of a transformation)**.** *With the usual assumptions on H and M, we have that*

$$[H \bullet M] = H^2 \bullet [M],$$

or equivalently,

$$[H \bullet M]_n = \sum_{k=1}^{n} H_k^2 \Delta [M]_k.$$

Proof. From the definition of an optional variation process and transformation, we can write

$$\begin{aligned}[H \bullet M]_n &= \sum_{k=1}^{n} (\Delta(H \bullet M)_k)^2 \\ &= \sum_{k=1}^{n} (H_k \Delta M_k)^2 \\ &= \sum_{k=1}^{n} H_k^2 \Delta[M]_k.\end{aligned}$$

□

We will also need the similar result for the predictable covariation processes, presented with proof in [17, p. 47].

Theorem 4 (Predictable variation process of a transformation). *With the usual assumptions on H and M, we have that*

$$\langle H \bullet M \rangle = H^2 \bullet \langle M \rangle,$$

or equivalently,

$$\langle H \bullet M \rangle_n = \sum_{k=1}^{n} H_k^2 \Delta \langle M \rangle_k.$$

Proof. From the definitions of a predictable variation process and transformation, we can write

$$\begin{aligned}\langle H \bullet M \rangle_n &= \sum_{k=1}^{n} \operatorname{Var}\{\Delta(H \bullet M)_k | \mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} \operatorname{Var}\{H_k \Delta M_k | \mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} H_k^2 \operatorname{Var}\{\Delta M_k | \mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} H_k^2 \Delta \langle M \rangle_k.\end{aligned}$$

□

We can generalize variation processes to covariation processes, which end up having a relationship not unlike that of covariance to variance.

Definition 9 (Covariation processes)**.** Similarly to the variation processes, we define the optional and predictable covariation processes as

$$[M_1, M_2]_n = \sum_{k=1}^{n} \Delta M_{1k} \Delta M_{2k},$$
$$\langle M_1, M_2 \rangle_n = \sum_{k=1}^{n} \mathrm{Cov}\{\Delta M_{1k}, \Delta M_{2k} | \mathcal{F}_{k-1}\}$$

Thus, by the above we have that $[M, M] = [M]$, and $\langle M, M \rangle = \langle M \rangle$. Perhaps unsurprisingly, the equivalent of Theorem 1 holds for covariation processes as well. The result is presented as found in [17, p. 50], where it is presented without proof.

Theorem 5. *If M_1 and M_2 are mean zero martingales with respect to some filtration $\{\mathcal{F}_n\}$, then*

$$M_1 M_2 - [M_1, M_2], \text{ and}$$
$$M_1 M_2 - \langle M_1, M_2 \rangle$$

are mean zero martingales.

Proof. The proof is computationally heavy and involves the same tricks as in Theorem 1, and hence brings nothing of interest to the work. One substitutes $M_{in} = M_{i(n-1)} + M_{in} - M_{i(n-1)}$ and expands the product as before. A similar trick is used for the optional and predictable covariation processes, allowing one to write

$$[M_1, M_2]_n = [M_1, M_2]_{n-1} + \Delta M_{1n} \Delta M_{2n}, \text{ and}$$
$$\langle M_1, M_2 \rangle_n = \langle M_1, M_2 \rangle_{n-1} + \mathrm{Cov}\{\Delta M_{1n}, \Delta M_{2n} | \mathcal{F}_{n-1}\}.$$

The resulting terms are then simplified as before, using standard rules. □

The following result clearly displays why covariation processes are necessary, even if only variation processes are of direct importance to us. The result is stated without proof only for predictable variation processes in [17, p. 50].

Theorem 6. *Let M_1 and M_2 be two martingales, as before. Then,*

$$[M_1 + M_2] = [M_1] + [M_2] + 2[M_1, M_2]$$
$$\langle M_1 + M_2 \rangle = \langle M_1 \rangle + \langle M_2 \rangle + 2\langle M_1, M_2 \rangle.$$

Proof. For the optional variation process we write

$$\begin{aligned}
[M_1 + M_2]_n &= \sum_{k=1}^{n} (\Delta M_{1k} + \Delta M_{2k})^2 \\
&= \sum_{k=1}^{n} (\Delta M_{1k})^2 + \sum_{k=1}^{n} (\Delta M_{2k})^2 + 2 \sum_{k=1}^{n} \Delta M_{1k} \Delta M_{2k} \\
&= [M_1]_n + [M_2]_n + 2[M_1, M_2]_n.
\end{aligned}$$

Similarly, for the predictable variation process

$$\begin{aligned}\langle M_1 + M_2\rangle_n &= \sum_{k=1}^{n} \mathrm{Var}\{\Delta M_{1k} + \Delta M_{2k}|\mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} \mathrm{Var}\{\Delta M_{1k}|\mathcal{F}_{k-1}\} + \sum_{k=1}^{n} \mathrm{Var}\{\Delta M_{2k}|\mathcal{F}_{k-1}\} \\ &\quad + 2\sum_{k=1}^{n} \mathrm{Cov}\{\Delta M_{1k}, \Delta M_{2k}|\mathcal{F}_{k-1}\} \\ &= \langle M_1\rangle_n + \langle M_2\rangle_n + 2\langle M_1, M_2\rangle_n.\end{aligned}$$

□

The following result is formulated from [17, Exercise 2.8] and is the equivalent of Theorems 3 and 4 for covariation processes.

Theorem 7 (Covariation processes of transformations). *Let M_1 and M_2 be martingales and H_1, H_2 predictable processes. Then*

$$[H_1 \bullet M_1, H_2 \bullet M_2]_n = \sum_{k=1}^{n} H_{1k}H_{2k}\Delta[M_1, M_2]_k$$
$$\langle H_1 \bullet M_1, H_2 \bullet M_2\rangle_n = \sum_{k=1}^{n} H_{1k}H_{2k}\Delta\langle M_1, M_2\rangle_k$$

Proof. For the optional covariation process we write

$$\begin{aligned}[H_1 \bullet M_1, H_2 \bullet M_2]_n &= \sum_{k=1}^{n} \Delta(H_1 \bullet M_1)_k \Delta(H_2 \bullet M_2)_k \\ &= \sum_{k=1}^{n} H_{1k}H_{2k}\Delta M_{1k}\Delta M_{2k} \\ &= \sum_{k=1}^{n} H_{1k}H_{2k}\Delta[M_1, M_2].\end{aligned}$$

Similarly, for the predictable covariation process

$$\begin{aligned}\langle H_1 \bullet M_1, H_2 \bullet M_2\rangle_n &= \sum_{k=1}^{n} \mathrm{Cov}\{\Delta(H_1 \bullet M_1)_k, \Delta(H_2 \bullet M_2)_k|\mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} \mathrm{Cov}\{H_{1k}\Delta M_{1k}, H_{2k}\Delta M_{2k}|\mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} H_{1k}H_{2k}\,\mathrm{Cov}\{\Delta M_{1k}, \Delta M_{2k}|\mathcal{F}_{k-1}\} \\ &= \sum_{k=1}^{n} H_{1k}H_{2k}\Delta\langle M_1, M_2\rangle_k,\end{aligned}$$

where second to last line follows because H_{ik} are predictable and hence measurable with respect to $\mathcal{F}_{k-1}$, and so can be treated as deterministic. □

We present the following result to motivate to some extent why the continuous time version is reasonable. The result for discrete time has an elementary proof, which we include for completeness, see [18, pp. 296-298] for a complete discussion.

Theorem 8 (Doob decomposition). *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_n\}, n \in \mathbb{N}$ be a filtration of $\mathcal{F}$. Further, let X_n be an adapted process, that is X_n is F_n measurable for each n. Then there exists a martingale M_n and a predictable process H_n such that for each n*

$$X_n = H_n + M_n.$$

Proof. Define H as

$$H_n = \sum_{k=1}^{n} \mathrm{E}(X_k|\mathcal{F}_{k-1}) - X_{k-1},$$

that is the sum of expected increments of X. Note that with the above definition, H_n is measurable with respect to $\mathcal{F}_{n-1}$. Define the process M as

$$M_n = X_0 + \sum_{k=1}^{n} X_k - \mathrm{E}(X_k|\mathcal{F}_{k-1}).$$

We then have that $X_n = H_n + M_n$. Furthermore, the process M is a martingale since

$$\begin{aligned}\mathrm{E}(M_n - M_{n-1}|\mathcal{F}_{n-1}) &= \mathrm{E}(X_n - \mathrm{E}(X_n|\mathcal{F}_{n-1})|\mathcal{F}_{n-1}) \\ &= \mathrm{E}(X_n|\mathcal{F}_{n-1}) - \mathrm{E}(X_n|\mathcal{F}_{n-1}) \\ &= 0.\end{aligned}$$

□

We now move onto continuous time equivalents of the results above. There is a substantial amount of technicality that is omitted in this section. For clarity, we state the definitions of filtrations, adapted processes, and martingales in continuous time.

Definition 10 (Filtration in continuous time). Let $(\Omega, \mathcal{F}, P)$ be a probability space. For every $t \in [0, \tau]$, let $\mathcal{F}_t$ be a sub-σ-algebra of $\mathcal{F}$. If for all $u \leq s$

$$\mathcal{F}_u \subseteq \mathcal{F}_s,$$

the collection $\{\mathcal{F}_t\}$ is called a filtration.

Definition 11 (Adapted process). A stochastic process $X = \{X(t) : t \in [0, \tau]\}$ is said to be adapted to some filtration $\{\mathcal{F}_t\}$ if $X(t)$ is measurable with respect to $\mathcal{F}_t$ for each t.

Definition 12 (Martingales in continuous time). A stochastic process $M = \{M(t) : t \in [0, \tau]\}$ is a martingale relative to $\{\mathcal{F}_t\}$ if it is adapted to $\{\mathcal{F}_t\}$ and satisfies the martingale property

$$\mathrm{E}(M(t)|\mathcal{F}_s) = M(s),$$

for all times $t > s$.

Definition 13 (Variation and covariation processes in continuous time)**.** Partition the interval $[0, t]$ into n subintervals of equal length and introduce

$$\Delta M_k = M\left(\frac{kt}{n}\right) - M\left(\frac{(k-1)t}{n}\right)$$
$$\mathcal{F}_k = \mathcal{F}_{kt/n}.$$

The continuous versions of the predictable and optional variation processes are then defined as limits in probability of the corresponding discrete process over increasingly finer partitions $[0, t] = \bigcup_k \left((k-1)t/n, kt/n\right]$

$$[M](t) = \lim_{n\to\infty} \sum_{k=1}^{n} (\Delta M_k)^2$$
$$\langle M\rangle(t) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathrm{Var}(\Delta M_k|\mathcal{F}_{k-1})$$
$$[M_1, M_2](t) = \lim_{n\to\infty} \sum_{k=1}^{n} \Delta M_{1k}\Delta M_{2k}$$
$$\langle M_1, M_2\rangle(t) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathrm{Cov}\{\Delta M_{1k}, \Delta M_{2k}|\mathcal{F}_{k-1}\}.$$

Predictability in continuous time is not as trivial of a notion as its discrete counterpart. We thus limit ourselves to sufficient conditions for a process to be predictable, see [17, p. 50].

Theorem 9 (Predictability for continuous processes)**.** *The following are sufficient conditions for a process $H = \{H(t) : t \in [0, \tau]\}$ to be predictable:*

- *H is adapted to the filtration $\{\mathcal{F}_t\}$*
- *Sample paths of H are left continuous.*

With all of the above, we are finally ready to introduce the most important operator in this section – the stochastic integral.

Definition 14 (Stochastic integral)**.** Suppose H is a predictable process and M is a martingale, both with respect to some filtration $\{\mathcal{F}_t\}$. As before, partition the interval $[0, t]$ into n equal subintervals. We introduce

$$H_k = H\left(\frac{(k-1)t}{n}\right),$$

and define ΔM_k as before. The stochastic integral is then defined as

$$\int_0^t H(s)dM(s) = \lim_{n\to\infty} \sum_{k=1}^{n} H_k \Delta M_k.$$

Note that this limiting definition does not cover the general case, for example, integration with respect to Wiener processes. For the general case, one introduces the Itô integral, but the above definition will suffice for our purposes.

With the above definition in mind, we state the following results as found in [17, p. 51] for variation and covariation processes of stochastic integrals.

Theorem 10 (Variation and covariation processes of stochastic integrals). *With the usual assumptions on H, H_1, H_2 and M, M_1, M_2, we have that*

$$\begin{aligned}
\left[\int_0^t H(s)dM(s)\right] &= \int_0^t H(s)^2 d[M](s) \\
\left\langle\int_0^t H(s)dM(s)\right\rangle &= \int_0^t H(s)^2 d\langle M\rangle(s) \\
\left[\int_0^t H_1(s)dM_1(s), \int_0^t H_2(s)dM_2(s)\right] &= \int_0^t H_1(s)H_2(s)d[M_1, M_2](s) \\
\left\langle\int_0^t H_1(s)dM_1(s), \int_0^t H_2(s)dM_2(s)\right\rangle &= \int_0^t H_1(s)H_2(s)d\langle M_1, M_2\rangle(s).
\end{aligned}$$

To present the continuous time version of the Doob decomposition, we first need to introduce the notion of submartingales.

Definition 15 (Submartingale). An adapted process $X = \{X(t) : t \in [0, \tau]\}$ is called a submartingale if for all $t > s$

$$\mathrm{E}(X(t)|\mathcal{F}_s) \geq X(s).$$

Theorem 11 (Doob-Meyer decomposition). *Any submartingale X can be decomposed uniquely as*

$$X = X^* + M,$$

where X^* is a nondecreasing predictable process and M is a mean zero martingale.

Proof. The proof of the theorem is outside the scope of this work, see [19] for a self-contained proof. □

The Doob-Meyer decomposition is perhaps the most important result in this chapter and will allow us to do a variety of things. The usefulness becomes clear when we observe that counting processes are submartingales.

Counting Processes

Definition 16 (Counting process). A counting process $N = \{N(t) : t \in [0, \tau]\}$ is a right-continuous process with jump sizes of 1 at event times, and constant in between.

The result below is adapted from [17, p. 54].

Theorem 12 (Existence of intensity process). *Let N be a counting process adapted to some filtration $\{\mathcal{F}_t\}$. Then there exists a unique cumulative intensity process $\Lambda(t)$ such that $N(t) - \Lambda(t)$ is a mean*

zero martingale. Furthermore, if Λ is absolutely continuous, there exists a predictable intensity process $\lambda(t)$ such that

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

Proof. Note that N is a submartingale since a counting process is nondecreasing by definition. We can then apply the Doob-Meyer decomposition to N, and write it as $N(t) = \Lambda(t) + M(t)$. By rearranging the terms, we obtain the desired result. For a proof of existence of λ in the case that Λ is absolutely continuous, refer to [20]. □

Variation processes are useful to us due to their connection to variance, but even more so because for counting processes they allow us to express variance in terms of known quantities. The result below is taken from [17, pp. 54-55].

Theorem 13 (Variation processes for a counting process martingale)**.** *Suppose*

$$M(t) = N(t) - \int_0^t \lambda(s)ds$$

with mean zero martingale M, counting process N, and intensity process λ as in Theorem 12. Then, the optional and predictable variation processes $[M]$ and $\langle M \rangle$ respectively, are

$$[M](t) = N(t)$$
$$\langle M \rangle(t) = \int_0^t \lambda(s)ds.$$

Proof. Recall the limiting definition of an optional variation process of a continuous time martingale. Consider some small interval of time. The cumulative intensity contribution to the increment of optional variation converges to 0 as $n \to \infty$, due to its continuity. Furthermore, since the counting process is a step function, possible contributions can only be made from time intervals containing the jumps. Thus, we have

$$[M](t) = \sum_{T_j \le t} \big(N(T_j) - N(T_j-)\big)^2 = \sum_{T_j \le t} 1 = N(t).$$

For the predictable variation process we can write heuristically

$$\begin{aligned} d\langle M \rangle(t) &= \mathrm{Var}\{dM(t)|\mathcal{F}_{t-}\} \\ &= \mathrm{Var}\{dN(t) - \lambda(t)dt|\mathcal{F}_{t-}\} \\ &= \mathrm{Var}\{dN(t)|\mathcal{F}_{t-}\} \\ &= \mathrm{E}\{dN(t)|\mathcal{F}_{t-}\}\left(1 - \mathrm{E}\{dN(t)|\mathcal{F}_{t-}\}\right), \end{aligned}$$

where $\mathcal{F}_{t-}$ is interpreted as the information available just before time t. The expected value above can be expressed as

$$\lambda(t)dt = P(dN(t) = 1|\mathcal{F}_{t-}) = \mathrm{E}\{dN(t)|\mathcal{F}_{t-}\},$$

so we obtain

$$\begin{aligned} d\langle M\rangle(t) &= \lambda(t)dt(1-\lambda(t)dt) \\ &\approx \lambda(t)dt. \end{aligned}$$

Thus,

$$\langle M\rangle(t) = \int_0^t \lambda(s)ds.$$

□

Definition 17 (Integrals with respect to counting process martingales). Let M be a counting process (mean zero) martingale. We interpret an integral with respect to M as

$$\int_0^t H(s)dM(s) = \int_0^t H(s)dN(s) - \int_0^t H(s)\lambda(s)ds.$$

Furthermore, the integral with respect to the counting process N is interpreted as

$$\int_0^t H(s)dN(s) = \sum_{T_j \leq t} H(T_j).$$

The following result will be key to finding variance and covariance estimators, taken from [17, p. 56].

Theorem 14. *Let M be a counting process martingale. Then*

$$\begin{aligned} \left[\int_0^t H(s)dM(s)\right] &= \int_0^t H(s)^2 dN(s) \\ \left\langle \int_0^t H(s)dM(s) \right\rangle &= \int_0^t H(s)^2\lambda(s)ds. \end{aligned}$$

Proof. The result follows by Theorems 10 and 13. □

The following result will be important in deriving test statistics for nonparametric tests and is presented in [17, p. 55] as Exercise 2.10.

Theorem 15. *Suppose N_1 and N_2 are two counting processes with no simultaneous jumps. Let M_1 and M_2 be the corresponding martingales. Then,*

$$\begin{aligned} [M_1, M_2](t) &= 0, \\ \langle M_1, M_2\rangle(t) &= 0, \end{aligned}$$

for all t. Martingales satisfying this property are said to be orthogonal.

Proof. Since N_1 and N_2 have no simultaneous jumps, the process $N_\cdot = N_1 + N_2$, is also a counting process. By the Doob-Meyer decomposition we can write $N_\cdot$ as

$$N_\cdot(t) = \int_0^t \lambda_\cdot(s)ds + M_\cdot(t),$$

from which we can identify that

$$\int_0^t \lambda_\cdot(s)ds = \int_0^t \lambda_1(s)ds + \int_0^t \lambda_2(s)ds$$
$$M_\cdot(t) = M_1(t) + M_2(t).$$

By Theorem 6, the results follow if we can show that $[M_1 + M_2] = [M_1] + [M_2]$ and $\langle M_1 + M_2\rangle = \langle M_1\rangle + \langle M_2\rangle$. For the optional variation process we can write

$$[M_1 + M_2] = [M_\cdot] = N_\cdot = N_1 + N_2 = [M_1] + [M_2].$$

Similarly, for the predictable variation process

$$\langle M_1 + M_2\rangle = \langle M_\cdot\rangle = \lambda_\cdot = \lambda_1 + \lambda_2 = \langle M_1\rangle + \langle M_2\rangle.$$

□

The following result will allow us to derive a nonparametric test statistic which will prove useful for clustering. The result is taken from [17, p. 56].

Theorem 16. *Suppose $N_1, N_2, \ldots, N_k$ are counting processes with no simultaneous jumps and are defined with respect to the same filtration. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ and $M_1, M_2, \ldots, M_k$ be the corresponding intensities and martingales respectively. Then*

$$\left[\sum_{j=1}^k \int_0^t H_j(s)dM_j(s)\right] = \sum_{j=1}^k \int_0^t H_j(s)^2 dN_j(s)$$
$$\left\langle \sum_{j=1}^k \int_0^t H_j(s)dM_j(s)\right\rangle = \sum_{j=1}^k \int_0^t H_j(s)^2\lambda_j(s)ds$$

Proof. As the martingales are orthogonal, so are the stochastic integrals with respect to them. The result follows by Theorems 6 and 15, since $d\langle M_i, M_j\rangle = 0$ and $d[M_i, M_j] = 0$ for all $i \neq j$. □

The remainder of this chapter is devoted to continuous time stochastic processes. The theory here is a minimal justification for obtaining asymptotic confidence intervals for our nonparametric estimators. We first introduce the notion of a Wiener process (also known as Brownian motion), which arises as a limit of random walks.

Definition 18 (Wiener process)**.** The Wiener process $W(t)$ is a real-valued continuous time stochastic process characterized by the following

- $W(0) = 0$,
- W has continuous paths, i.e. $W(t)$ is continuous in t,
- W has independent increments over non-overlapping time intervals, i.e. for any times $s < t < u$, increment $W(t) - W(s)$ is independent of $W(u) - W(t)$,
- W has normally distributed increments, $W(t) - W(s) \sim N(0, t - s)$.

Of interest to us is the following result, given in [17, pp. 61-62] as Exercise 2.12.

Theorem 17 (Gaussian martingale)**.** *Let $V(t)$ be some strictly increasing continuous function with $V(0) = 0$. Then the Gaussian martingale $U(t) = W(V(t))$ retains the first three properties of the Wiener process $W(t)$, and*

$$U(t) - U(s) \sim N(0, V(t) - V(s)).$$

Furthermore, $U(t)$ is a martingale and $\langle U \rangle(t) = V(t)$.

Proof. The first property of W is satisfied because $V(0) = 0$. Since $V(t)$ is continuous and strictly increasing, one can think of $V(t)$ as a rescaling of the time axis. Composition of continuous functions is continuous, and so the second property follows. Since V is a strictly increasing function, it maps non-overlapping intervals to non-overlapping intervals, so the third property holds. The distribution of increment follows for similar reasons.

To see that $U(t)$ is a martingale, write

$$\begin{aligned} \mathrm{E}(U(t)|\mathcal{F}_s) &= \mathrm{E}(U(t) - U(s) + U(s)|\mathcal{F}_s) \\ &= \mathrm{E}(U(t) - U(s)|\mathcal{F}_s) + \mathrm{E}(U(s)|\mathcal{F}_s) \\ &= U(s). \end{aligned}$$

To show that $\langle U \rangle(t) = V(t)$, we show that $U(t)^2 - V(t)$ is a martingale, and the result follows by Theorem 11, since the decomposition must be unique.

$$\begin{aligned} \mathrm{E}\{U(t)^2 - V(t)|\mathcal{F}_s\} &= \mathrm{E}\{(U(t) - U(s) + U(s))^2 - V(t)|\mathcal{F}_s\} \\ &= \mathrm{E}\{(U(t) - U(s))^2 + 2U(s)(U(t) - U(s) + \\ &\quad + U(s)^2 - V(t)|\mathcal{F}_s\} \\ &= \mathrm{E}\{(U(t) - U(s))^2|\mathcal{F}_s\} + 2U(s)\,\mathrm{E}\{(U(t) - U(s)|\mathcal{F}_s\} + \\ &\quad + U(s)^2 - V(t) \\ &= V(t) - V(s) + U(s)^2 - V(t) \\ &= U(s)^2 - V(s). \end{aligned}$$

□

Asymptotic results of interest to us will be a consequence of the following theorem, taken from [17, pp. 63-65].

Theorem 18 (CLT for counting process martingales). *Suppose $H^{(n)}(t)$ is a sequence of predictable processes for all $n \geq 1$ and $M^{(n)}(t)$ is a corresponding sequence of counting process martingales*

$$M^{(n)}(t) = N^{(n)}(t) - \int_0^t \lambda^{(n)}(s)ds.$$

Assume that $V(t)$ is a strictly increasing function, and that we can write

$$V(t) = \int_0^t v(s)ds.$$

If

$$H^{(n)}(s)^2\lambda^{(n)}(s) \xrightarrow{P} v(s) > 0 \quad \textit{and}$$
$$H^{(n)}(s) \xrightarrow{P} 0$$

for all $s \in [0, \tau]$ as $n \to \infty$, then

$$\int_0^t H^{(n)}(s)dM^{(n)}(s) \xrightarrow{d} W(V(t)), \quad \textit{as } n \to \infty.$$

Proof. The proof is outside the scope of this work. The original result is due to Rebolledo presented in [21]. □

Nelson-Aalen Estimator

In a survival analysis setting, for some individuals the event of interest is never observed during the examined time period. In such a case, for individual i we say that the actual survival time T_i was censored. We thus observe potentially censored survival times $\tilde{T}_i$ and introduce censoring indicators $d_i = \mathbb{1}\{\tilde{T}_i = T_i\}$, taking on value 1 if the observed time is an event time and 0 otherwise. There are many assumptions of various strictness about the censoring mechanism at play. We will assume what is referred to as independent censoring, that is for an individual who has not experienced the event of interest at time t, the risk of experiencing the event during $[t, t + dt)$ is the same as in the situation without censoring. Formally, the condition is given as

$$P(t \leq \tilde{T}_i \leq t + dt, d_i = 1 | \tilde{T}_i \geq t, \mathcal{F}_{t-}) = P(t \leq T_i \leq t + dt | T_i \geq t).$$

For each of the individuals we introduce a counting process taking on values 0 or 1, given by $N_i(t) = \mathbb{1}\{\tilde{T}_i \leq t, d_i = 1\}$. From these individual counting processes, we define the aggregate process that records all events

$$N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{i=1}^{n} \mathbb{1}\{\tilde{T}_i \leq t, d_i = 1\}.$$

If we assume independent censoring and that $h_i(t) = h(t)$ for all i, that is that the hazard rate is the same for all individuals, it can be shown that the counting process $N(t)$ has an intensity process of the form

$$\lambda(t) = Y(t)h(t), \tag{1}$$

where h is the hazard rate. In the above $Y(t)$ is a predictable process that denotes the number of individuals at risk just before time t, and can be expressed as

$$Y(t) = \sum_{i=1}^{n} \mathbb{1}\{\tilde{T}_i \geq t\}.$$

A counting process with an intensity that can be written as in (1) is said to fulfill the multiplicative intensity model. The Nelson-Aalen estimator then estimates the cumulative hazard

$$H(t) = \int_0^t h(s)ds.$$

Estimating the cumulative hazard is significantly easier than estimating the hazard directly, for similar reasons as to why it is easier to estimate the cumulative distribution function rather than the density. All results on the Nelson-Aalen estimator are taken from [17, pp. 70-90].

We know that $N(t)$ is a submartingale as it is a nondecreasing process, and hence by the Doob-Meyer decomposition there exists a unique predictable cumulative intensity process $\Lambda(t)$ such that $N(t) = \Lambda(t) + M(t)$, and $M(t)$ is a mean zero martingale. We will assume throughout that Λ is absolutely continuous, hence there exists a predictable process $\lambda(t)$ such that

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

With the form of the intensity process in (1), we can thus write

$$dN(t) = Y(t)h(t)dt + dM(t).$$

We wish to isolate $h(t)$ in the expression above, but $Y(t)$ can take on zero value. To address this, we introduce the indicator $J(t) = \mathbb{1}\{Y(t) > 0\}$. Multiplying by $J(t)$ and dividing by $Y(t)$, we obtain in integral form

$$\int_0^t \frac{J(s)}{Y(s)} dN(s) = \int_0^t J(s)h(s)ds + \int_0^t \frac{J(s)}{Y(s)} dM(s), \tag{2}$$

where we say $J(t)/Y(t) = 0$ whenever $Y(t) = 0$. The left-hand side of the above is the Nelson-Aalen estimator, which due to integration with respect to a counting process being equivalent to summation over jump times, can equivalently be written as

$$\hat{H}(t) = \int_0^t \frac{J(s)}{Y(s)} dN(s) = \sum_{T_j \leq t} \frac{1}{Y(T_j)}.$$

If we introduce

$$H^*(t) = \int_0^t J(s)h(s)ds,$$

we can rewrite (2) as

$$\hat{H}(t) - H^*(t) = \int_0^t \frac{J(s)}{Y(s)} dM(s).$$

Note that since M is a mean zero martingale, then so is the integral on the right-hand side. This implies that $\mathrm{E}(\hat{H}(t) - H^*(t)) = 0$, so $\hat{H}$ is an unbiased estimator of H^*. Of course, we wish to have an estimator of H, but this is not possible in a nonparametric setting since we cannot estimate $h(t)$ when $Y(t) = 0$, i.e. when no individuals are at risk just before time t.

To derive the variance of the estimator $\hat{H}$, recall that $\mathrm{Var}\, M(t) = \mathrm{E}[M](t)$ and write

$$\begin{aligned}[\hat{H} - H^*](t) &= \left[\int_0^t \frac{J(s)}{Y(s)} dM(s)\right] \\ &= \int_0^t \left(\frac{J(s)}{Y(s)}\right)^2 d[M](s) \\ &= \int_0^t \frac{J(s)}{Y(s)^2} dN(s).\end{aligned}$$

We thus have that

$$\hat{\sigma}^2(t) = \mathrm{E}[\hat{H} - H^*](t) = \sum_{T_j \le t} \frac{1}{Y(T_j)^2}.$$

The Nelson-Aalen estimator is asymptotically normal, which can be motivated by applying the CLT for martingales, see Theorem 18. We apply the CLT to the martingale

$$\sqrt{n}(\hat{H}(t) - H^*(t)) = \int_0^t \sqrt{n}\frac{J(s)}{Y(s)} dM(s).$$

We assume that $Y(t)/n \xrightarrow{P} y(t)$, where y is some positive function. We check the conditions of Theorem 18, with $H^{(n)}(t) = \sqrt{n}J(t)/Y(t)$ and $\lambda^{(n)}(t) = Y(t)h(t)$,

$$\begin{aligned}\left(\sqrt{n}\frac{J(t)}{Y(t)}\right)^2 Y(t)h(t) = \frac{J(t)h(t)}{Y(t)/n} &\xrightarrow{P} \frac{h(t)}{y(t)}, \\ \sqrt{n}\frac{J(t)}{Y(t)} = \frac{1}{\sqrt{n}}\frac{J(t)}{Y(t)/n} &\xrightarrow{P} 0,\end{aligned}$$

as $n \to \infty$. Thus, $\sqrt{n}(\hat{H}(t) - H^*(t))$ converges to a mean zero Gaussian martingale with variance $\int_0^t h(s)/y(s)ds$. Asymptotically, H^* and H are equivalent, so the result holds for $\sqrt{n}(\hat{H}(t) - H(t))$ as well. In particular, for a fixed time t, the Nelson-Aalen estimator is asymptotically normally distributed. In line with this, a significance α confidence interval for $H(t)$ can be given as

$$\hat{H}(t) \pm z_{1-\alpha/2}\hat{\sigma}(t),$$

where z indicates a quantile of the standard normal distribution.

Usually we want an estimator of the hazard rate, not of the cumulative hazard. One approach is to smooth the Nelson-Aalen estimate with a kernel smoother, which is a weighted moving average of the increments. We denote $\Delta\hat{H}(T_j)$ to be the increments of the Nelson-Aalen estimate over the jump time T_j and write

$$\hat{h}(t) = \frac{1}{b}\sum_{T_j} K\left(\frac{t - T_j}{b}\right)\Delta\hat{H}(T_j).$$

The kernel K is a bounded function that vanishes outside $[-1, 1]$ and integrates to 1. The parameter b is the bandwidth, so $\hat{h}(t)$ is then a weighted average of Nelson-Aalen increments over $[t - b, t + b]$. There are several choices for the kernel function, a popular one for this application is the Epanechnikov kernel $K(x) = 3(1 - x^2)/4$. The choice of bandwidth b determines the smoothness of the estimated hazard and constitutes a trade-off between bias and variability. There are optimal ways to choose the bandwidth which we will not delve into; see for example [22]. With the estimator above, we cannot produce an estimate for $t \in [0, b]$. The solution has been to adopt boundary kernels, see for example [23]. There also exist extensions that allow the bandwidth to be time-dependent, as to adjust smoothing based on the amount of data available locally.

Increments of the Nelson-Aalen estimator are uncorrelated, so the variance of $\hat{h}(t)$ is simply

$$\widehat{\mathrm{Var}}(\hat{h}(t)) = \frac{1}{b^2} \sum_{T_j} K\left(\frac{t - T_j}{b}\right)^2 \Delta\hat{\sigma}^2(T_j),$$

where $\Delta\hat{\sigma}^2(T_j)$ are increments of the variance estimator derived above.

Kaplan-Meier Estimator

The Kaplan-Meier estimator is a nonparametric estimator of the survival function $S(t)$ of the random variable T. The estimator can be derived heuristically as a product of conditional survival probabilities over an increasingly finer partition. The partition can be made fine enough where at most one event occurs in each subinterval, and then the conditional probabilities are 1 in intervals with no events, and $1 - 1/Y(T_j)$ otherwise, yielding

$$\hat{S}(t) = \prod_{T_j \leq t} \left\{1 - \frac{1}{Y(T_j)}\right\}.$$

We will approach the estimator from a somewhat more formalized direction – in its connection to the Nelson-Aalen estimator, as done in [12, pp. 255-258]. In view of this, we introduce the product integral.

Definition 19 (Product integral)**.** Let $X(t)$, $t \in [0, \tau]$ be a cadlag function (right-continuous with left limits) and define the product integral as

$$\prod_{u \in [0,t]} (1 + dX(u)) = \lim_{\max |t_i - t_{i-1}| \to 0} \prod (1 + X(t_i) - X(t_{i-1})),$$

where $0 = t_0 < t_1 < \cdots < t_n = t$ is a partition of $[0, t]$. If X is a step function with ordered jump times T_j, the above becomes a product over the jump times

$$\prod_{u \in [0,t]} (1 + dX(u)) = \prod_{T_j \leq t} (1 + \Delta X(T_j)).$$

Of interest to us will be the following result, see [12, Therorem II.6.1].

Theorem 19. *The product integral* $Y(t) = \prod_{u \leq t}(1 + dX(u))$ *exists and is the unique solution to the integral equation*

$$Y(t) = 1 + \int_0^t Y(u-)dX(u).$$

Note that the right-hand side of the definition of the product integral is the result of applying a first order Euler discretization scheme, so the result is somewhat intuitive.

With this theory, we are now ready to state the definition of the cumulative hazard rate for an arbitrary distribution, which will allows us to draw a connection between the Kaplan-Meier and Nelson-Aalen estimators.

Definition 20 (Cumulative hazard rate)**.** Denote the survival function with $S(t)$. For an arbitrary distribution, the cumulative hazard rate can be defined as

$$H(t) = -\int_0^t \frac{dS(u)}{S(u-)}.$$

For the absolutely continuous case, the above is

$$H(t) = -\int_0^t \frac{S'(u)}{S(u)}du,$$

and for a purely discrete distribution

$$H(t) = \sum_{u \leq t} P(T = u | T \geq u) = \sum_{u \leq t} -\frac{S(u) - S(u-)}{S(u-)}.$$

Of key importance will be the following result that connects the survival function and cumulative hazard in the general case, see [12, Therorem II.6.6].

Theorem 20. *If* $H(t)$ *is the cumulative hazard rate, the survival function can be written as*

$$S(t) = \prod_{u \leq t}(1 - dH(u)).$$

Proof. Writing the definition of the hazard rate as

$$dH(u) = -\frac{dS(u)}{S(u-)},$$

we can express the survival function as

$$S(t) = 1 - \int_0^t S(u-)dH(u).$$

Theorem 19 can be immediately applied to the above with $X = -H$, yielding the result. □

We can thus estimate the survival function using the plug-in estimator, by introducing $S(t) = \eta(H)$ and writing

$$\begin{aligned}\hat{S}(t) &= \eta(\hat{H}) \\ &= \prod_{u \leq t} (1 - d\hat{H}(u)) \\ &= \prod_{T_j \leq t} (1 - \Delta\hat{H}(T_j)),\end{aligned}$$

where $\hat{H}$ is the Nelson-Aalen estimator of the cumulative hazard. Hence, the Kaplan-Meier estimator is related to the Nelson-Aalen estimator in exactly the same way as the survival function is to the cumulative hazard.

To study the statistical properties of the Kaplan-Meier estimator, we proceed similarly to Nelson-Aalen. As before, put $J(t) = \mathbb{1}\{Y(t) > 0\}$, and $H^*(t) = \int_0^t J(u)h(u)du$. We can then introduce

$$S^*(t) = \prod_{u \leq t} (1 - dH^*(u)).$$

Before we continue, we need to state an important result that will be of use, see [12, Theorem II.6.2].

Theorem 21 (Duhamel's equation). *Let* $Y_1 = \prod(1 + dX_1)$*, and* $Y_2 = \prod(1 + dX_2)$*. Then*

$$\frac{Y_1(t)}{Y_2(t)} - 1 = \int_0^t \frac{Y_1(u-)}{Y_2(u)} d(X_1 - X_2)(u).$$

We can now apply the Duhamel equation with $\hat{S}$ and S^*

$$\frac{\hat{S}(t)}{S^*(t)} - 1 = \int_0^t \frac{\hat{S(u-)}}{S^*(u)} d(\hat{H} - H^*)(u).$$

As was shown previously, $\hat{H} - H^*$ is a mean zero martingale, so the integral is as well, which leads to

$$\mathrm{E}\left(\frac{\hat{S}(t)}{S^*(t)}\right) = 1.$$

One may further prove that the Kaplan-Meier estimator is uniformly consistent. This allows us to make asymptotic approximations $\hat{S}(u-)/S^*(u) \approx 1$, $S^*(t) = S(t)$, and $H^*(u) = H(u)$, with which we can write

$$\begin{aligned}\frac{\hat{S}(t)}{S(t)} - 1 &\approx -\int_0^t d(\hat{H} - H)(u) \\ \hat{S}(t) - S(t) &\approx -S(t)\left(\hat{H}(t) - H(t)\right).\end{aligned}$$

We then have that

$$\operatorname{Var}\hat{S}(t) \approx S(t)^2 \operatorname{Var}\hat{H}(t),$$

which we estimate as

$$\hat{\tau}^2 = \hat{S}(t)^2\hat{\sigma}^2(t),$$

where $\hat{\sigma}^2$ is the variance of the Nelson-Aalen estimator.

Nonparametric Tests

The results in this section are presented as adapted from [17, pp. 104-113]. Suppose that we wish to test if the survival functions, or alternately the hazard rates, of two groups are the same over some time interval $[0, t_0]$. To this end, we assume that both groups follow the multiplicative intensity model detailed above, that is, with intensities for counting processes N_i for $i = 1, 2$ of the form

$$\lambda_i = h_i(t)Y_i(t), \quad i = 1, 2.$$

We wish to test

$$\begin{aligned} H_0 : &\quad h_1(t) = h_2(t), \text{ for all } t \in [0, t_0] \\ H_1 : &\quad h_1(t) \leq h_2(t) \text{ or } h_1(t) \geq h_2(t), \text{ for all } t \in [0, t_0], \end{aligned}$$

that is the alternate hypothesis is that one of the hazard rates dominates the other for all times of interest. We denote the common value of the hazard rates under the null hypothesis as $h(t)$. The test statistic will be based on a comparison of increments of the Nelson-Aalen estimators for each of the counting processes. As before, the Nelson-Aalen estimator for each of the processes is

$$\hat{H}_i(t) = \int_0^t \frac{J_i(s)}{Y_i(s)} dN_i(s), \quad i = 1, 2.$$

We then introduce a nonnegative predictable weight process L such that $L(t) = 0$ whenever $J_1(t)J_2(t) = 0$. We study the process

$$Z(t_0) = \int_0^{t_0} L(t) \left(d\hat{H}_1(t) - d\hat{H}_2(t) \right),$$

that sums up the weighted differences in the Nelson-Aalen estimators for the respective groups. We rewrite Z using the definition of the Nelson-Aalen estimator as

$$\begin{aligned} Z(t_0) &= \int_0^{t_0} \frac{L(t)J_1(t)}{Y_1(t)} dN_1(t) - \int_0^{t_0} \frac{L(t)J_2(t)}{Y_2(t)} dN_2(t) \\ &= \int_0^{t_0} \frac{L(t)}{Y_1(t)} dN_1(t) - \int_0^{t_0} \frac{L(t)}{Y_2(t)} dN_2(t), \end{aligned}$$

where the equality follows because $L(t) = L(t)J_1(t) = L(t)J_2(t)$. Using the Doob-Meyer decomposition, under the null hypothesis we can write the processes as

$$dN_i(t) = h(t)Y_i(t)dt + dM_i(t),$$

where as usual, M_i are martingales. We now insert this into the expression for Z presented above

$$\begin{aligned} Z(t_0) &= \int_0^{t_0} \frac{L(t)}{Y_1(t)} \Big(h(t)Y_1(t)dt + dM_1(t) \Big) \\ &\quad - \int_0^{t_0} \frac{L(t)}{Y_2(t)} \Big(h(t)Y_2(t)dt + dM_2(t) \Big) \\ &= \int_0^{t_0} \frac{L(t)}{Y_1(t)} dM_1(t) - \int_0^{t_0} \frac{L(t)}{Y_2(t)} dM_2(t). \end{aligned}$$

Hence, $Z(t_0)$ is a mean zero martingale as it is a difference of two mean zero martingales. The estimator of the variance of Z can be found by looking at the predictable variation process of Z

$$\begin{aligned}\langle Z\rangle(t_0) &= \left\langle \int_0^{t_0} \frac{L(t)}{Y_1(t)} dM_1(t) - \int_0^{t_0} \frac{L(t)}{Y_2(t)} dM_2(t) \right\rangle \\ &= \int_0^{t_0} \left(\frac{L(t)}{Y_1(t)}\right)^2 d\langle M_1\rangle(t) + \int_0^{t_0} \left(\frac{L(t)}{Y_2(t)}\right)^2 d\langle M_2\rangle(t) \\ &= \int_0^{t_0} \frac{L(t)^2(Y_1(t)+Y_2(t))}{Y_1(t)Y_2(t)} h(t)dt \\ &= \int_0^{t_0} \frac{L(t)^2 Y_\cdot(t)}{Y_1(t)Y_2(t)} h(t)dt,\end{aligned}$$

where $Y_\cdot(t) = Y_1(t) + Y_2(t)$ is the number at risk just before time t for both groups considered together. By definition of cumulative hazard, $dH(t) = h(t)dt$, so we can find an estimate of the above by replacing $dH(t)$ with the estimated Nelson-Aalen increments $d\hat{H}(t)$

$$d\hat{H}(t) = \frac{1}{Y_\cdot(t)} dN_\cdot(t),$$

where $N_\cdot = N_1 + N_2$. This is possible since under the null hypothesis there is no distinction between the groups, so we can use the Nelson-Aalen estimator computed on the two groups together. We thus obtain the estimator

$$V(t_0) = \int_0^{t_0} \frac{L(t)^2}{Y_1(t)Y_2(t)} dN_\cdot(t).$$

To see that the variance estimator is unbiased, first note that the intensity process for $N_\cdot$ is $\lambda_\cdot = h(t)Y_\cdot(t)$ under the null hypothesis and write

$$\begin{aligned}\mathrm{E}\{V(t_0)\} &= \mathrm{E}\left\{ \int_0^{t_0} \frac{L(t)^2}{Y_1(t)Y_2(t)} \Big(h(t)Y_\cdot(t)dt + dM_\cdot(t) \Big) \right\} \\ &= \int_0^{t_0} \frac{L(t)^2 Y_\cdot(t)}{Y_1(t)Y_2(t)} h(t)dt \\ &= \mathrm{E}\{\langle Z\rangle(t_0)\}.\end{aligned}$$

One can show that $Z(t_0)$ is approximately normally distributed under the null hypothesis, see [17, pp. 112-113] for details. This allows us to construct tests for the null hypothesis based on one of

$$\begin{aligned}\frac{Z(t_0)}{\sqrt{V(t_0)}} &\sim N(0,1) \\ \frac{Z(t_0)^2}{V(t_0)} &\sim \chi^2(1).\end{aligned}$$

The remaining problem is that of the choice for the weight process $L(t)$. The most common choice, and the one we will use for our clustering algorithm, is the logrank test. The weight process for it is given by

$$L(t) = \frac{Y_1(t)Y_2(t)}{Y_\cdot(t)}.$$

There are numerous other choice for $L(t)$ which produce alternate tests; see [17, Table 3.2] for examples and key references.

There exist extensions of the above to more than two samples and stratified versions. For a discussion of these and handling of ties, see [17, pp. 109-112].

Relative Risk Regression

This section is adapted from [17, pp. 133-142]. For this class of models, the hazard rate is modeled, and for a vector of covariates $\mathbf{x}_i$ for individual i, the relation

$$h(t|\mathbf{x}_i) = h_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t))$$

is assumed to hold. Here, $h_0(t)$ is the *baseline hazard*, which is left unspecified, and constitutes the nonparametric part of the model. The relative risk function r is specified, and regression coefficients $\boldsymbol{\beta}$ determine the effect of the covariates on the baseline hazard. We normalize such that $r(\boldsymbol{\beta}, \mathbf{0}) = 1$, and thus have that the baseline hazard is the hazard for an individual with all covariates equal to 0.

Perhaps the most important case arises when r is the exponential relative risk function $r = \exp\big(\boldsymbol{\beta}^T \mathbf{x}_i(t)\big)$, giving us Cox regression. Furthermore, if the covariates are not time-dependent, we have the Cox proportional hazards model.

The estimation of coefficients $\boldsymbol{\beta}$ is complicated by the semi-parametric nature of the model. Instead of using the regular likelihood methods, we use partial likelihood.

The data set is assumed to be of the form $\{(\tilde{T}_i, d_i, \mathbf{x}_i(t))\}_{i=1}^n$, where as before $\tilde{T}_i$ is the observed survival time, d_i is the censoring indicator, and $\mathbf{x}_i(t)$ is a covariate vector for individual i. For each of the individuals i, we introduce counting processes $N_i(t) = \mathbb{1}\{\tilde{T}_i \leq t, d_i = 1\}$ which then have intensity processes $\lambda_i(t)$ of the form

$$\lambda_i(t) = Y_i(t)h(t|\mathbf{x}_i),$$

where Y_i is an at risk indicator for the individual, taking value 1 if the individual is at risk in the time interval $[0, t)$, and 0 otherwise. Combining this with the model specified hazard rate above, we obtain

$$\lambda_i(t) = Y_i(t)h_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t)).$$

We can then introduce the aggregated counting process $N_\cdot(t) = \sum_{l=1}^n N_l(t)$, which counts the number of events across all individuals. Observe that the intensity of this process is given by

$$\lambda_\cdot(t) = \sum_{l=1}^n \lambda_l(t) = \sum_{l=1}^n Y_l(t)h_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t)).$$

We then introduce

$$\pi(i|t) = \frac{\lambda_i(t)}{\lambda_\cdot(t)} = \frac{Y_i(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{l=1}^n Y_l(t)r(\boldsymbol{\beta}, \mathbf{x}_l(t))},$$

which allows us to factorize $\lambda_i(t) = \lambda_\cdot(t)\pi(i|t)$. The quantity above is the conditional probability for individual i to observe an event at time t given the past and, crucially, that some event is observed at that time.

Taking the product of the conditional probabilities over the observed event times yields the partial likelihood for $\boldsymbol{\beta}$. Assuming there are no ties, let the event times be ordered such that $T_1 < T_2 < \dots$ and denote i_j as the individual with the j'th ordered survival time. We can then write the partial likelihood as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^n \pi(i_j|T_j) = \prod_{j=1}^n \frac{Y_{i_j}(T_j)r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{l=1}^n Y_l(T_j)r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))}.$$

If we introduce the notion of a risk set at time t to be $\mathcal{R}(t) = \{l|Y_l(t) = 1\}$, we can write the partial likelihood as

$$L(\beta) = \prod_{j=1}^n \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{l\in\mathcal{R}(T_j)} r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))}.$$

Note that the partial likelihood has no dependence on the baseline hazard h_0, so we can estimate the regression coefficients $\boldsymbol{\beta}$ without knowing h_0. We can treat $L(\boldsymbol{\beta})$ as a usual likelihood and use standard optimization methods, e.g. Newton–Raphson, to find the maximum. We will refer to the value of $\boldsymbol{\beta}$ that maximizes the partial likelihood L above, as $\hat{\boldsymbol{\beta}}$.

Due to the ease of interpretation and commonality, we will now restrict ourselves to the case when $r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \exp\{\boldsymbol{\beta}^T\mathbf{x}_i(t)\}$, i.e. Cox regression. For this case, we express below the log-likelihood which is usually easier to optimize

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{j=1}^n \boldsymbol{\beta}^T\mathbf{x}_{i_j}(T_j) - \sum_{j=1}^n \log\left(\sum_{l\in\mathcal{R}(T_j)} \exp\{\boldsymbol{\beta}^T\mathbf{x}_l(T_j)\}\right).$$

Note that we assumed all event times are unique, which is usually not realistic due to the way real data is actually entered. One approach to this is to add a random noise to the event times as to break the tie. There are other approaches, see for example [24, section 8.4].

To estimate the cumulative hazard, we first need an estimator of the cumulative baseline hazard. Examining the intensity $\lambda_\cdot(t)$ of the aggregate counting process $N_\cdot(t)$ one can see that it $\boldsymbol{\beta}$ were known, we would have a multiplicative intensity model on hand. In such a case, the Nelson-Aalen estimator could be used. Since we have an estimate $\hat{\boldsymbol{\beta}}$ derived above, we can use it as the best alternative in the Nelson-Aalen estimator. We thus obtain the Breslow estimator of cumulative baseline hazard

$$\hat{H}_0(t) = \int_0^t \frac{dN_\cdot(u)}{\sum_{l=1}^n Y_l(u)r(\hat{\boldsymbol{\beta}}, \mathbf{x}_l(u))} = \sum_{T_j\le t} \frac{1}{\sum_{l\in\mathcal{R}(T_j)} r(\hat{\boldsymbol{\beta}}, \mathbf{x}_l(T_j))}.$$

Aalen's Additive Regression

In Aalen's additive hazards model we assume the following form for the hazard rates

$$h(t|\mathbf{x}_i) = \beta_0(t) + \beta_1(t)x_{i1}(t) + \cdots + \beta_p(t)x_{ip}(t),$$

where as before $\mathbf{x}_i$ is the vector of covariates for the i'th individual. All results in this section are adapted from [17, pp. 154-160]. With the assumption of a multiplicative intensity model, we have that

$$\lambda_i(t) = Y_i(t)\big(\beta_0(t) + \beta_1(t)x_{i1}(t) + \cdots + \beta_p(t)x_{ip}(t)\big).$$

By introducing the cumulative regression functions $dB_q(t) = \beta_q(t)dt$ and with the usual decomposition $dN_i(t) = \lambda_i(t)dt + dM_i(t)$, we can write

$$dN_i(t) = Y_i(t)dB_0(t) + \sum_{j=1}^{p} Y_i(t)x_{ij}(t)dB_j(t) + dM_i(t).$$

Thus, for a fixed t the above takes the form of usual multiple linear regression, with $dN_i(t)$ acting as the observations, $Y_i(t)x_{ij}(t)$ acting as covariates, and $dM_i(t)$ acting as random errors. We write the above in matrix notation as

$$d\mathbf{N}(t) = \mathbf{X}(t)d\mathbf{B}(t) + d\mathbf{M}(t),$$

where $\mathbf{N}(t) = (N_1(t), \ldots, N_n(t))^T$ a vector of counting processes, and similarly $d\mathbf{M}(t) = (dM_1(t), \ldots, dM_n(t))^T$ a vector of corresponding counting process martingales, and

$$\mathbf{X}(t) = \begin{pmatrix} Y_1(t) & Y_1(t)x_{11}(t) & \cdots & Y_1(t)x_{1p}(t) \\ \vdots & \vdots & \vdots & \vdots \\ Y_i(t) & Y_i(t)x_{i1}(t) & \cdots & Y_i(t)x_{ip}(t) \\ \vdots & \vdots & \vdots & \vdots \\ Y_n(t) & Y_n(t)x_{n1}(t) & \cdots & Y_n(t)x_{np}(t) \end{pmatrix}.$$

Applying the standard result for OLS regression, we can solve for $d\mathbf{B}(t)$ as

$$d\hat{\mathbf{B}}(t) = \big(\mathbf{X}(t)^T\mathbf{X}(t)\big)^{-1}\mathbf{X}(t)^T d\mathbf{N}(t),$$

when $\mathbf{X}(t)$ is full rank. Let $J(t)$ be an indicator of $\mathbf{X}(t)$ being full rank and introduce for brevity the left inverse

$$\mathbf{X}^-(t) = \big(\mathbf{X}(t)^T\mathbf{X}(t)\big)^{-1}\mathbf{X}(t)^T.$$

With the above, we obtain

$$\begin{aligned} \hat{\mathbf{B}}(t) &= \int_0^t J(u)\mathbf{X}^-(u)d\mathbf{N}(u) \\ &= \sum_{T_j \le t} J(T_j)\mathbf{X}^-(T_j)\Delta\mathbf{N}(T_j), \end{aligned}$$

where $\Delta\mathbf{N}(T_j)$ is zero but for the component corresponding to the individual that experienced an event at time T_j, where it is one.

To study the covariance of the estimator $\hat{\mathbf{B}}(t)$ we proceed as before. Put

$$\mathbf{B}^*(t) = \int_0^t J(u)d\mathbf{B}(u),$$

and write

$$\begin{aligned}
\hat{\mathbf{B}}(t) &= \int_0^t J(u)\mathbf{X}^-(u)d\mathbf{N}(u) \\
&= \int_0^t J(u)\mathbf{X}^-(u)\{\mathbf{X}(u)d\mathbf{B}(u) + d\mathbf{M}(u)\} \\
&= \int_0^t J(u)\mathbf{X}^-(u)\mathbf{X}(u)d\mathbf{B}(u) + \int_0^t J(u)\mathbf{X}^-(u)d\mathbf{M}(u) \\
&= \mathbf{B}^*(t) + \int_0^t J(u)\mathbf{X}^-(u)d\mathbf{M}(u).
\end{aligned}$$

Since $\mathbf{M}$ is a mean zero martingale, we have as a consequence that $\hat{\mathbf{B}} - \mathbf{B}^*$ is a mean zero martingale and hence that $\mathrm{E}\{\hat{\mathbf{B}}(t) - \mathbf{B}^*(t)\} = 0$, so $\hat{\mathbf{B}}(t)$ is an unbiased estimator of $\mathbf{B}^*(t)$, and an almost unbiased estimator of $\mathbf{B}(t)$. To obtain the covariance matrix of $\hat{\mathbf{B}}(t)$, we find the predictable variation process of $\hat{\mathbf{B}} - \mathbf{B}^*$

$$\begin{aligned}
\langle\hat{\mathbf{B}} - \mathbf{B}^*\rangle(t) &= \left\langle \int_0^t J(u)\mathbf{X}^-(u)d\mathbf{M}(u) \right\rangle \\
&= \int_0^t J(u)\mathbf{X}^-(u)\,\mathrm{diag}\{\boldsymbol{\lambda}(u)du\}\mathbf{X}^-(u)^T \\
&= \sum_{T_j \le t} J(T_j)\mathbf{X}^-(T_j)\,\mathrm{diag}\{\Delta\mathbf{N}(T_j)\}\mathbf{X}^-(T_j)^T \\
&= \hat{\boldsymbol{\Sigma}}(t),
\end{aligned}$$

where we estimate $\boldsymbol{\lambda}(u)du$ as $d\mathbf{N}(u)$.

Similarly as for the Nelson-Aalen estimator, one can show that $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}^*)$ converges in distribution to a mean zero multivariate Gaussian martingale with covariance that is the limit in probability of $n\hat{\boldsymbol{\Sigma}}(t)$, see [12, section VII.4.2] for details. Asymptotically $\mathbf{B}^*(t)$ and $\mathbf{B}(t)$ are identical, so $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B})$ also converges weakly to this Gaussian martingale. This allows us to construct an approximate confidence interval for the cumulative regression functions as

$$\hat{B}_q(t) \pm z_{1-\alpha/2}\sqrt{\left(\hat{\boldsymbol{\Sigma}}(t)\right)_{qq}}.$$

If the hazard rates are of interest, smoothing is carried out similarly to the Nelson-Aalen estimator, see [17, pp. 159-160] for details.

Accelerated Failure Time Models

Accelerated failure time models are fully parametric, and their fitting is done through usual likelihood methods. The material in this section is adapted from [24, pp. 393-406]. The models get their name from the impact of the covariates – unlike in Cox regression, the effect is multiplicative on the time, not the hazards. Assuming that S_0 denotes the baseline survival function, i.e. for an individual with all covariates identically zero, we have a defining relationship

$$S(t|\mathbf{x}_i) = S_0(\exp\{\boldsymbol{\theta}^T\mathbf{x}_i\}t),$$

for all individuals i with $\mathbf{x}_i$ as their covariates. A different perspective is given by considering a log-linear model for survival time. Write

$$V = \log T = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i + \sigma W,$$

and suppose that

$$S_0(t) = P(\exp\{\beta_0 + \sigma W\} > t).$$

We then have that the two representations are equivalent with $\boldsymbol{\theta} = -\boldsymbol{\beta}$

$$\begin{aligned} S(t|\mathbf{x}_i) &= P(\exp\{\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i + \sigma W\} > t) \\ &= P(\exp(\beta_0 + \sigma W) > \exp\{-\boldsymbol{\beta}^T\mathbf{x}_i\}t) \\ &= S_0(\exp\{-\boldsymbol{\beta}^T\mathbf{x}_i\}t). \end{aligned}$$

The various AFT models are characterized by assuming specific distributions for W or S_0, with one defining the other.

We begin by discussing the most common choice, where the survival function is assumed to be Weibull distributed. The case arises if we assume W to follow the extreme value distribution with survival function $S_W(w) = \exp\{-e^w\}$. As before, put $V = \log T = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i + \sigma W$, and write

$$\begin{aligned} S_V(v_i) &= P(V > v_i) \\ &= P\left(W > \frac{v_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{x}_i}{\sigma}\right) \\ &= \exp\left\{-\exp\left\{\frac{v_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{x}_i}{\sigma}\right\}\right\}. \end{aligned}$$

We reparametrize $\alpha = 1/\sigma$, and $\lambda(\mathbf{x}_i) = \exp\{\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i\}$ and obtain

$$S_V(v_i) = \exp\{-e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}\}.$$

To obtain the density we differentiate the above to get

$$f_V(v_i) = \alpha e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}\exp\{-e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}\}.$$

The above can be expressed on the original timescale. We can introduce $t = e^v$, by noting that $S_V(v) = S_T(e^v)$. Inserting this into the expression for S_V found above, we get

$$S_T(t_i) = \exp\left\{ -\left(\frac{t_i}{\lambda(\mathbf{x}_i)}\right)^{\alpha} \right\},$$

which is one of many parametrizations of the Weibull distribution. This in particular has been chosen because of its use in the `lifelines` package in Python.

To write the likelihood, we introduce censoring indicators d_i taking on value 1 if the event has been observed for the i'th individual and 0 otherwise. We can then write the likelihood as

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma | v, \mathbf{x}) &= \prod_{i=1}^{n} \left(f_V(v_i)\right)^{d_i} \left(S_V(v_i)\right)^{1-d_i} \\ &= \prod_{i=1}^{n} \left(\alpha e^{v_i \alpha} \lambda(\mathbf{x}_i)^{-\alpha} \exp\{-e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}\}\right)^{d_i} \\ &\quad \times \left(\exp\{-e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}\}\right)^{1-d_i}. \end{aligned} \tag{3}$$

Another common choice is to assume that W follows the Log-logistic distribution

$$S_W(w) = \frac{1}{1+e^w},$$

and similarly to the Weibull case, find that

$$S_V(v_i) = S_W\left(\frac{v_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{x}_i}{\sigma}\right) = \frac{1}{1+\exp\{\frac{v_i-\beta_0-\boldsymbol{\beta}^T\mathbf{x}_i}{\sigma}\}}.$$

If we, as before, introduce $\alpha = 1/\sigma$ and $\lambda(\mathbf{x}_i) = \exp\{\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i\}$, we get

$$S_V(v_i) = \frac{1}{1+e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}}.$$

Differentiating as before, we obtain the density

$$f_V(v_i) = \alpha \frac{e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha}}{(1+e^{v_i\alpha}\lambda(\mathbf{x}_i)^{-\alpha})^2}.$$

Performing the timescale change, we get

$$S_T(t_i) = \frac{1}{1+t_i^{\alpha}\lambda(\mathbf{x}_i)^{-\alpha}} = \frac{1}{1+\left(\frac{t_i}{\lambda(\mathbf{x}_i)}\right)^{\alpha}}.$$

Estimation is once again done using Maximum Likelihood, with the same form as for the Weibull case in (3), substituting $f_V(v)$ and $S_V(v)$ with those found above.

The last AFT model we will consider is the Log-normal distribution. For this model, W is assumed to follow a standard normal distribution, and hence $V = \log T$ is also normal. With Φ denoting the standard normal cumulative distribution function, the survival function is simply

$$S_V(v_i) = 1 - \Phi\left(\frac{v_i - \beta_0 - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma}\right) = 1 - \Phi\left(\frac{v_i - \beta(\mathbf{x}_i)}{\sigma}\right),$$

and the density is

$$f_V(v_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{v_i - \beta(\mathbf{x}_i)}{\sigma}\right)^2\right\},$$

where for brevity $\beta(\mathbf{x}_i) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i$. Parameter estimation is done the same way as for the other AFT models.

Significance of Variables and Multiple Testing

Suppose we wish to test whether a variable is significant or not in the models presented above. For Cox regression and the AFT models this turns out to be a relatively easy problem. Beginning with Cox regression, as before, let $\boldsymbol{\beta}$ be the true value of coefficients and $\hat{\boldsymbol{\beta}}$ be their maximum partial likelihood estimates. By [12, Theorem VII.2.2], the estimates $\hat{\boldsymbol{\beta}}$ are asymptotically normal, so a standard Wald test can be used. Recall that the test is given by

$$H_0 : \beta_q = 0$$
$$H_1 : \beta_q \neq 0,$$

and the test statistic is

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2}{\operatorname{Var} \hat{\boldsymbol{\beta}}},$$

which is χ^2 distributed with one degree of freedom. See [12, Chapter VII.2.1] for a derivation of the variance of $\hat{\boldsymbol{\beta}}$. The test is implemented in nearly all survival analysis statistical packages, in particular Python's `lifelines`, which is used in this work.

In AFT models estimates of parameters are given by the maximum likelihood estimator, which is asymptotically normal, see for example [25]. The Wald test can thus be used for AFT models as well.

Suppose we fit a model on some collection of variables and would like to know which variables are statistically significant in the model. A naive approach would be to simply carry out the Wald test for each of the variables, thus yielding as many null and alternate hypotheses as variables, and simply labeling variables as significant if the corresponding Wald test p-value is under the desired significance level α. This simple procedure, however, yields a high family-wise error rate (FWER) – the probability that at least one of the null hypotheses under consideration was erroneously rejected. Recall that the significance level α is defined as the probability of incorrectly rejecting the null hypothesis. If we denote the individual null hypotheses as H_{0q},

with $q = 1, \ldots, m$, we can write the FWER as

$$\begin{aligned}
\text{FWER} &= P(\text{at least one } H_{0q} \text{ rejected} | \text{all } H_{0q} \text{ true}) \\
&= 1 - P(\text{no } H_{0q} \text{ rejected} | \text{all } H_{0q} \text{ true}) \\
&= 1 - \prod_{q=1}^{m} P(\text{do not reject } H_{0q} | H_{0q} \text{ true}) \\
&= 1 - \prod_{q=1}^{m} \big(1 - P(\text{reject } H_{0q} | H_{0q} \text{ true})\big) \\
&= 1 - (1-\alpha)^m.
\end{aligned}$$

If we let α be some standard choice, for example 0.05, it takes only 7 hypotheses to have a FWER of over 0.3, and 15 hypotheses for a FWER of over 0.5. A simple way to control for the FWER would be to considerably lower the significance levels for all of the individual tests, but this is likely to negatively affect the power of the procedure as the impact on Type II errors is not evaluated. The suggested solution is the Holm-Bonferroni procedure as introduced by Holm in [26] which provides a uniformly most powerful procedure, while controlling for the FWER.

Suppose we wish to control the FWER at level α. To carry out the Holm-Bonferroni procedure, we first order the p-values $p_{(1)}, \ldots, p_{(m)}$ and the corresponding null hypotheses $H_{0(1)}, \ldots, H_{0(m)}$. The method is iterative, and starts with the first ordered p-value and null hypothesis. If

$$p_{(1)} < \frac{\alpha}{m},$$

we reject the first null hypothesis and consider the next one; if the inequality does not hold, we do not reject any hypotheses. For all following hypotheses $H_{0(k)}$, we reject each iteratively if

$$p_{(k)} < \frac{\alpha}{m+1-k}.$$

If on any given step the p-value is above the corrected significance level, we do not reject that and any following hypotheses.

Clustering Methodology

Suppose each individual has an associated categorical variable that we would like to use to model survival time T. If there are many categories, it is reasonable to expect that the distribution of the recovery time for some of the categories is similar enough. Placing the 'similar' categories into clusters and then using the clusters instead of the original categories in the models would then result in fewer coefficients to estimate in the models, since each additional category necessitates a separate coefficient.

A naive approach to clustering would be to take some quantile, e.g. median, of the Kaplan-Meier estimate of the survival function for each of the categories and use a standard approach

like K-means or K-medians. Unfortunately, this would mean comparing the survival distributions at only one point, which we deem to be overly reductive. In turn, we wish to compare the survival distributions for each of the categories.

Our approach rests on logrank tests to cluster the categories. Comparisons will be done pairwise – two categories with their resulting populations will be compared for equality of hazard rates. Recall that the logrank test has the following structure

$$\begin{aligned} H_0 : &\quad h_1(t) = h_2(t), \\ H_1 : &\quad h_1(t) \leq h_2(t) \text{ or } h_1(t) \geq h_2(t). \end{aligned}$$

From the test above we will obtain a p-value, which together with a chosen significance level, can be used to decide if the two categories can be placed in a cluster. Ideally, we would have a test where the null and alternate hypotheses were swapped, since in reality our 'null hypothesis' is that the categories are not equivalent. Unfortunately, such a test is not theoretically possible since the null hypothesis would be composite and the distribution under it would be unknown. Furthermore, as the goal is variable reduction without a severe information loss, we prefer to keep the categories separate unless they are similar enough. We thus prefer to minimize not Type I, but Type II error, equivalent to maximizing the power of the test. A simple way to control this is through the choice of the significance level, with a higher value corresponding to higher power.

The above detailed the simple case where we pretend there are only two categories. The basis of our clustering algorithm is the 'pairwise' equivalent of the above procedure. We carry out the logrank test between each pair of categories and record the resulting p-values. Tests where the p-value is below the chosen significance level α can be immediately discarded, as they correspond to categories for which the survival distributions are different. The remaining pairwise tests are all that could not be rejected, that is those for which the hazard rates are equal. By construction, a given category might be involved in several of these candidate pairs. This presents a challenge, as it is not obvious what the clusters should be. To address this, we will cluster the most 'similar' pairs. In practice, we iteratively cluster the pairs as ordered by the p-values from highest to lowest if neither of the categories in the pair has been clustered before.

The above forms a single iteration of the clustering algorithm proposed. Note that with just the above, the clusters can contain at most two categories. To cluster further, we repeat the above step, with the newly formed clusters acting as categories did before. Note that now, as before, the new clusters consist of at most two old clusters. The key difference is that the old clusters can correspond to one or two categories, so the new clusters can contain between one and four categories. The above can be repeated many times, with the naive stopping criterion being that all resulting clusters have different hazard rates, that is that all p-values for pairwise comparisons of clusters are below the chosen significance level α.

Unfortunately, the above presents the problem of post hoc analysis. On all iterations after the first, the hypotheses are presented based on previous test decisions on the same data set. To address this, each iteration is done on a new subset of data. This involves introducing the number of iterations as a variable, which then allows us to split the original data set into separate data sets for each iteration.

Iterative multiple logrank clustering

$p_{max} \leftarrow 1$
while $p_{max} > p$ **or** `iteration` $<$ `maxit` **do**
    `test_pairs` = `pairwise_log_rank`({data set for `iteration` })
    `candidate_groups` $\leftarrow$`sort`({pairs in `test_pairs` : logrank p-value $> p$})
    `taken_groups` $\leftarrow$ [ ]
    for `group` **in** `candidate_groups` **do**
        if `group` **not in** `taken_groups` **then**
            for `code` **in** `group` **do**
                `code` $\leftarrow$ `group`
            end for
            `taken_groups` $\leftarrow$ `taken_groups` + `group`
        end if
    end for
    $p_{max} = \max$\{logrank test p-values\}
    `iteration` $\leftarrow$ `iteration` $+1$
end while

Experimental Results

In the *Theory and Methods* section above, we have introduced five models: Cox regression, Aalen's additive hazards, and Weibull, Log-logistic and Log-normal AFT. We will now evaluate these on the case of wind turbine recovery data. The models are within the scope of survival analysis, and many terms are motivated by the case where time until death is modeled. This distinction is entirely linguistic, but there is potential for confusion, so some clarification of terminology is in order. The random variable T in this section will denote the recovery time of a failed turbine. Thus, the survival function at time t gives the probability that the recovery time will be equal to or larger than t. Similarly, the hazard rate gives the instantaneous probability of recovery given that it has not yet occurred. As a consequence, the interpretation of cumulative hazard rate is the risk of recovery accumulated up until a given time.

We begin by introducing the data set and discussing the data preparation steps carried out, together with the motivation behind them. We then discuss some general statistical tendencies in the data and the continuous variable used in the models. As to keep the discussion of each of the models structured and nonrepetitive, we discuss and motivate the general modelling approach beforehand.

Data set

Our data set consists of turbine failure (and recovery) events for a single customer with a large portfolio. Most crucially, for each event there is a timestamp `timestamp_start` corresponding to the failure time and `timestamp_end` for when the turbine recovered. Taking the difference between the recovery timestamp and the failure timestamp yields the duration modeled by the random variable T, which is what we will we work with. For each of the failure events we also create a censoring indicator `has_timestamp_end`; the observation is censored if the recovery timestamp is missing. Note that with this setup, 'time zero' is whenever the turbine failed, that is, using a medical equivalent, we have that the individuals enter the study when they have failed, and do so on a common relative timescale.

For each of the failure events we further know a unique identifier of the turbine that experienced the event `TurbineId`. The turbine identifier will later be used to construct 'history' variables which track the past number of failure events for each of the turbines.

The most import covariate used will be the error code associated with the failure. There is only one associated code for each event, which corresponds to the issue that ultimately caused the turbine to malfunction. The error codes vary quite a bit in their meaning, but this can be any mechanical issue like *'Breaks Overheating'* or software like *'Yaw Signals Invalid'*. The error codes are numerical values which uniquely map to the causes of failure; most importantly, this mapping is only unique for a given model of turbine. To reduce the number of variables in

the models, we will cluster the error codes. To keep different turbine models separate, we will only do this within a given turbine model. The unique identifier of the model is given by the variable `TurbineTypeId`.

We would like to use some variable that could correlate with the wear endured by the turbine, as it is not unreasonable to suspect that this could have an effect on the recovery times. The variable we opt to construct is the logarithm of energy `log_energy` produced by the turbine in the month before the failure occurred. For each of the turbines, we have a data set indexed by time of the energy produced up until that time. The variable is then simply the logarithm of the difference between the energy produced at the time of failure and the month before the failure.

Data Preparation

Wind turbines have automatic recovery procedures, so it is of little interest to study failures from which recovery can generally be done automatically and promptly. To address this, we find the median recovery time from each of the failures as identified by the error codes. If this value is under a chosen threshold, we chose to discard all events that arose due to this failure code. This choice results in a reduction of the data set, and as a consequence also has the added benefit of being able to feasibly fit all of the models on consumer-grade hardware. We heuristically choose this threshold to be 30 minutes because this provides a good balance between retaining data and studying events of interest. Note that we do not remove individual data points, instead opting to not consider some causes of failure.

Another issue arises due to how data has been entered in the past. Due to customer acquisition and resulting retroactive filling in of historical data, there can be mistakenly unresolved recovery events, which have recovery times on the magnitude of years. To address this, we remove data points with recovery time past one year, as we deem these to be erroneous values.

As briefly mentioned above, definitions of turbine error codes are unique to the particular make and model of turbine, so say code '1' for one turbine may not be equivalent in interpretation to the same code for another turbine. We would like to retain as much data as possible while minimizing the number of unique code sets in the data set, so we discard from consideration turbine types with less than $15\,000$ observations. Furthermore, we only consider codes that have at least 200 observations.

With the aforementioned data preparation complete, we retain $115\,376$ observations, with 21 unique collections of codes, with 139 unique codes in total. Of the $115\,376$ observations, only one is censored, so there is essentially no censoring left in our data set. The unfiltered data set contains $1\,372\,587$ observations.

Discarding codes that do not satisfy our preferences from consideration in particular implies that any following analysis does not necessarily generalize to the entire population, i.e. to all turbines for any event. This however is not an issue, as we only wish to predict recovery times for codes that are left in the data set, be it due to longer recovery times or because the observations that are left carry enough information for each code, so there can be more confidence in the

resulting predictions of models.

Data Exploration

We begin by briefly analyzing the filtered data set. See Figure 1 for an overview of the distribution of recovery times. As can be seen, the recovery times are quite skewed, with many plants recovering very quickly. In fact, the median recovery time is exactly one hour. This will present issues with classification, as we will be working with a very unbalanced data set. If we view the recovery times as belonging to two categories *<24h* and *>24h*, the former constitutes nearly 97% of the data set.

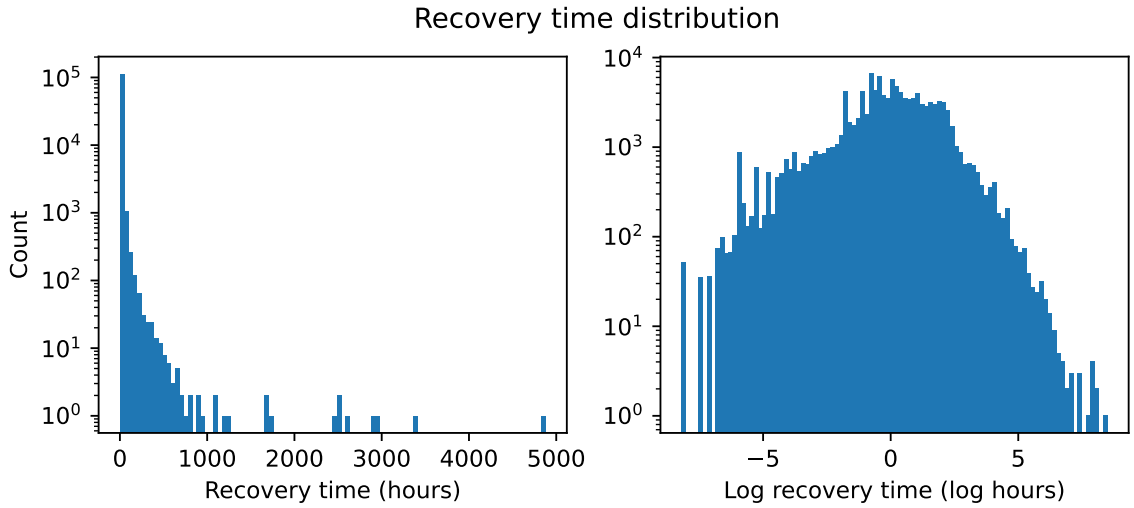


Figure 1: Histogram of recovery times with 100 bins used. Presented in linear time *(left)*, and logarithmic time *(right)*.

As we will be evaluating some parametric models, we present in Figure 2 Quantile-Quantile plots for the three distributions assumed for recovery time. Note that Weibull distribution seems to describe the data best. Both from the log-log histogram in Figure 1 and the QQ-plots we can see that the left tail is poorly described by Log-normal and Log-logistic distributions. For all of the distributions, most of the observations seem relatively well described by the distributions, the differences manifest in the tails of recovery time distribution.

With this in mind, it is important to note that the distributions must hold conditional on the covariates, not unconditionally. Thus, it would be unjust to discard the Log-logistic and Log-normal AFT models just now.

We will use the logarithm of energy produced in the month prior to failure as a variable in the models. We inspect its distribution by means of a histogram and normal QQ-plot in Figure 3. Note that albeit visually the values seem normal, the QQ-plot paints a different picture. Part of the data seems relatively well described by the normal distribution, but the tails are not symmetric, which can also be seen from the histogram.

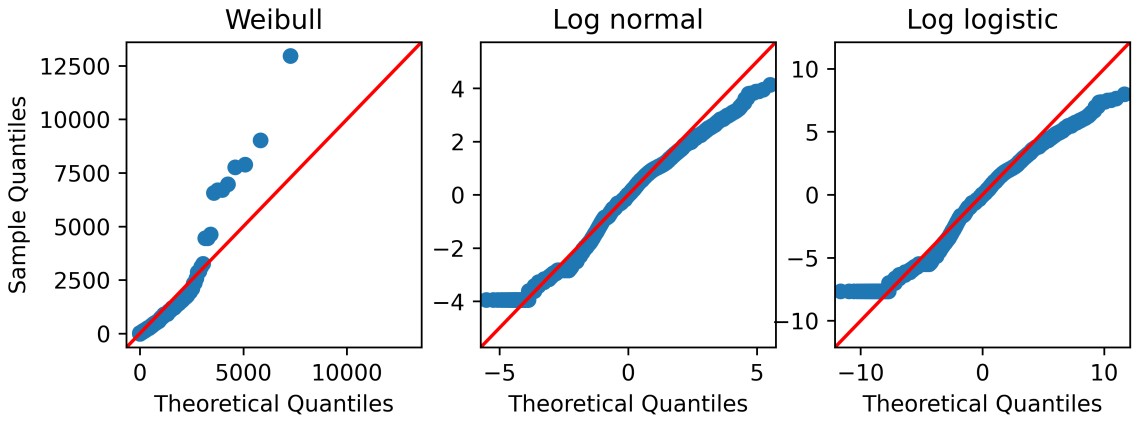


Figure 2: QQ-plots of recovery time for distributions assumed in the three AFT models under consideration.

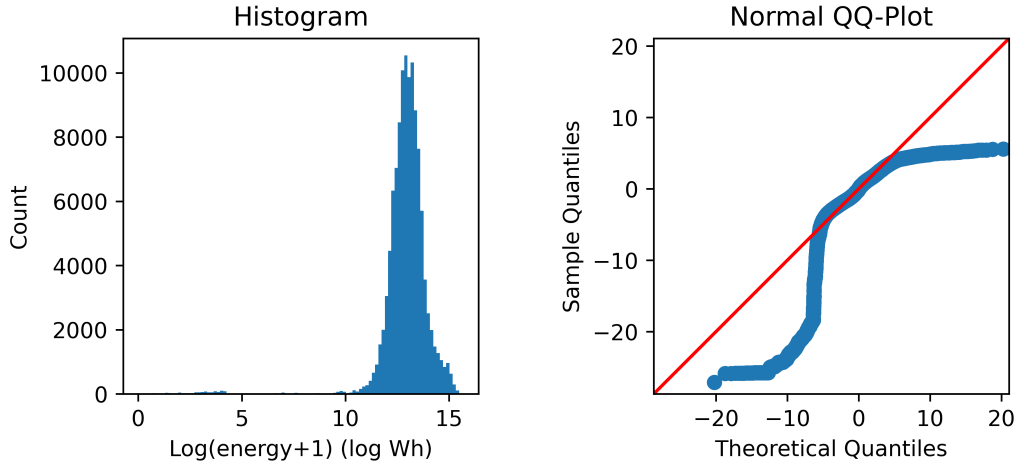


Figure 3: Distribution of logarithm of energy produced in the month before failure. Histogram of log energy *(left)*, Normal distribution QQ-plot of log energy *(right)*.

As discussed before, for each failure event there is an associated error code generated by the turbine. Each event has only one associated cause of failure. This code corresponds to a specific issue that ultimately caused the turbine to malfunction.

As a general check of reasonableness of the use of Cox regression, we check that the proportional hazards assumption holds. Recall that this implies that any failure event has the same baseline hazard, with the only difference between hazards for different events being a constant scaling factor. To do so, we use the built-in `lifelines` method `check_assumptions`. The method tests whether scaled Schoenfeld residuals are 0, and relies on the findings of Grambsch and Therneau in [27]. For all models discussed in this section, the Proportional Hazards assumption fails. This is unsurprising as due to the size of the data set, even very small deviations from the assumption result in the null hypothesis of proportional hazards getting rejected. Furthermore, as our goal is prediction, one could argue that we should avoid placing too much importance on

whether the assumption holds. The assumption might not hold strictly, but the model overall could produce better predictions on average than alternatives.

Error Code Clustering

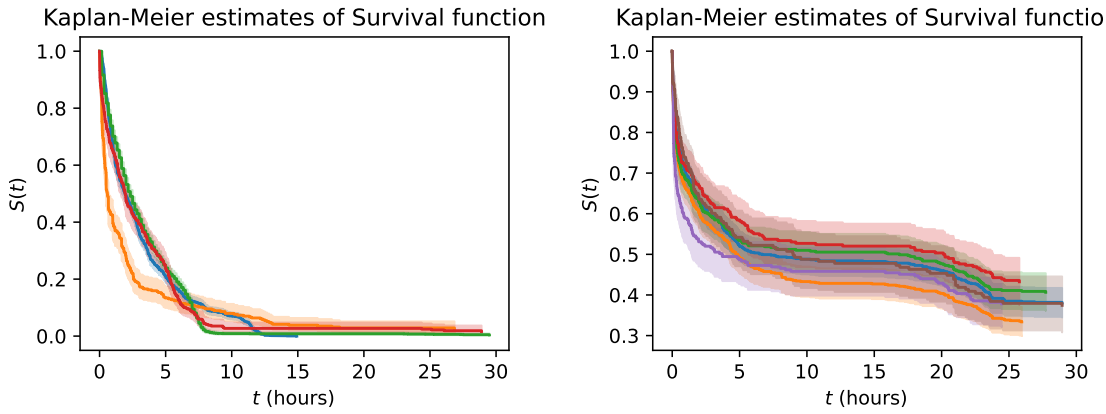
It is reasonable to assume that many of the issues represented by the error codes exhibit similar enough recovery. The iterative clustering method will be used, with the various categories being error codes. Instead of directly using the error codes as a variable in the models, the resulting clusters will be used. This has a benefit of reducing the number of variables in the models, as each additional category is associated with a dummy variable.

Since there is only one censored observation, the logrank test becomes equivalent to a two sample Kolmogorov-Smirnov test for most tests performed, and to our knowledge this iterative testing approach has not been considered before. To retain a larger degree of separation between manufacturers and specific models of turbines, we only cluster codes within the same turbine type. In other words, error codes between two different manufacturers or different models of turbines would not be compared or clustered.

For the iterative multiple logrank clustering, we somewhat arbitrarily set the number of iterations to 4. This number of iterations allows for a sufficient amount of variable reductions, while not performing the tests on overly small data sets. The p-value was set to be 0.9, so as to ensure that the power of each individual logrank test was high. Furthermore, we are primarily interested in variable reduction and the clustering of similar codes is more of an immediate goal. Thus, we choose not to cluster codes for turbine types where there are 4 or fewer codes in the data set.

With these parameters we end up somewhat reducing the number of future dummy variables, with 112 clusters remaining, 99 of them corresponding to one code, 3 of them to 2 codes, 8 to 3 codes, and 1 for 4 and 6 codes. We visually inspect the reasonableness of the clustering by comparing Kaplan-Meier estimated survival functions for the most clustered codes, see Figure 4.

Note that in Figure 4a recovery from one of the codes differs significantly from other codes in the cluster up until the 5th hour. The clustering of this error code is likely explained by little to no difference later on in recovery. Recall also that the alternate hypothesis in the logrank test assumes that survival functions do not cross. For the cluster with 6 codes presented in Figure 4b no such issue is present, but this is likely due to the extremely wide confidence intervals. Overall, we deem the resulting clustering to be reasonable enough, given the limitations of the logrank test and a lack of better alternatives. The resulting clustering is quite conservative, as most error codes are left untouched. Where most clustering occurred, the sanity check provided by Kaplan-Meier estimates of the most clustered codes, should put to rest any worries about clustering overly dissimilar codes.



(a) 4 Codes (b) 6 Codes

Figure 4: Kaplan-Meier estimates of survival functions for the most clustered codes.

Modelling Approach

A critical reader might question the necessity of survival analysis in a setting where censoring is essentially non-existent. We, however, believe that several justifications exist as to the use of this seemingly overcomplicated framework for our case. Primarily, there is nothing inherently incorrect about the use of the framework, since all results hold true when no censoring is present. As will be discussed below, survival analysis approaches allow us to predict conditioned on a failure not resolving for some period of time, which is of key importance for the application. Furthermore, it is very likely that the absence of censoring in our case is a result of the data preparation steps and choice of customer. Thus, even if a simpler approach could be taken with the data at hand, the optimal methodology would accommodate data with a higher censoring rate. Furthermore, it could be argued that our results do not translate to a data set with censoring. The truth of this point heavily depends on the nature of the censoring; the general assumption is of *independent censoring*, which is entirely non-informative of the recovery time distribution. If the independent censoring assumption were true in this fictional data set, the results would be equivalent; if not, then the methodology presented below could be applied in full to that data set, with perhaps a different resulting 'best' model.

The goal of the models is not necessarily the best fit on data, but the ability to predict recovery times on previously unseen data. Recall that the goal is two-fold: classification of recovery into within and after 24 hours, and evaluation of accuracy on the within 24 hours case. There is also an issue of evaluating models, as they are not special cases of some one model. If this were the case, selecting the best model would be a matter of testing if some parameter is significant, and if not a reduction to a less general model would occur. For parametric AFT models, we have the likelihood and hence AIC and BIC values. For Cox regression, we have access to 'partial' variants of these. The Aalen's additive hazards model has none of these. Concordance index is the only metric that all of the models share, but one should be wary of using it as a guide to model selection. Monotone transformations of data do not affect the concordance index, so predicted recovery times could differ significantly from the actual ones without it being reflected

in the metric.

Combined with the goal of the models, and the aforementioned pitfalls, we proceed as follows. We perform an 80-20 split of the data into a training and testing data sets, using the cluster as a stratifying variable. All fitting of models is performed on the training data set, and all evaluation on the testing. Due to the censoring nature of survival data, common metrics like mean square and mean absolute error are not usually used. However, in our case, there is essentially no censoring present. Furthermore, we will only compute the prediction accuracy on those data points, where the model has predicted the recovery time to be under 24 hours. Since the individual data points considered in the computation of metrics for each of the models may be different due to classification, we will also report a confusion matrix. Ideally, all cases would be correctly classified, but, of course, misclassification is unavoidable.

One approach to classification could be to use the median from the predicted survival function. If the predicted median is over 24 hours, we classify the turbine as recovering after 24 hours, and the converse with the under 24 hours case. Using the median for classification is equivalent to finding the value of the predicted survival function at 24 hours and seeing if it is over 0.5. Unfortunately, the data at hand, viewed as belonging to these two classes, is severely skewed towards the under 24-hour category. The approach outlined above then results in almost always classifying recoveries in the under 24-hour category.

A way to alleviate this issue is to find some optimal classification threshold. To discuss this further, we introduce some standard terminology. We will refer to *over 24 hours* as the Positive class, and *under 24 hours* as the Negative class. Correct classification is distinguished by the prefix 'True', and on the contrary, incorrect as 'False'. We will be studying the *True Positive Rate* TPR and *False Positive Rate* FPR, which are defined as follows

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

Another name for TPR is *Sensitivity*, and 1-FPR is *Specificity*. Unless the classifier is perfect, that is can always predict correctly, there will always be a trade-off between sensitivity and specificity. To study how good our models are at classifying, we will plot Receiver Operating Characteristic curves, or ROC curves. These plot the TPR against the FPR for all possible thresholds. In practice, as data sets are finite, many thresholds will result in equivalent TPR and FPR. Thus, it is only necessary to compute the curve at the unique number of survival probabilities at 24 hours present in the data set. The area under the ROC curve is simply called Area Under the Curve or AUC, and by construction takes values between 0 and 1. AUC values have many interpretations, which we will not dwell on. For our purposes, it suffices to say that the higher the AUC value the better, and a value less than or equal to 0.5 corresponds to a random or worse than random classifier.

Selecting an optimal threshold is thus a problem of formulating the desired trade-off between sensitivity and specificity, or alternatively the TPR and the FPR. As a simple solution, we opt for the threshold that maximizes the geometric mean of sensitivity and specificity, that is, maximizes $\sqrt{\mathrm{TPR}(1 - \mathrm{FPR})}$. Thus, for each model we consider, we compute an ROC curve, and as threshold, take the value that maximizes the geometric mean of sensitivity and specificity. Note that we say the threshold is 'optimal' if it satisfies this criterion; optimality in general is a direct

result of each end user's risk preferences – the criterion we choose is optimal in the sense of being better than the naive choice. Threshold selection is done on the training data set, same as model fitting. See Figure 5 for a visualization of classification.

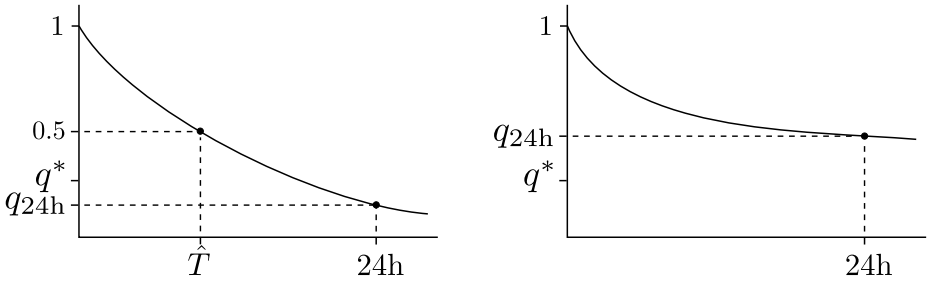


Figure 5: Classification and point prediction of recovery time. Two predicted survival curves, one that is classified to recover within 24 hours *(left)*, and another that is classified to recover after 24 hours *(right)*. The optimal threshold is denoted by q^*.

If we classified a failure event to recover within 24 hours, we would like to produce a point estimate of the recovery time. Two ways of generating predictions were considered – using the median or expectation of the predicted survival function. Unfortunately, finding the expected value is equivalent to finding the area under the survival function, that is integrating it. Aalen's additive hazards model presents a challenge with the above, as convergence in the tail becomes an issue for codes where there are not enough observations available. This results in the area under the survival function being infinite, or poorly defined, as it is unknown when the function reaches value 0. Due to these considerations, we choose to use the median as the point estimator for recovery time, as we wish to examine all models in a unified framework, see Figure 5. Using the median over the mean is fairly common in survival analysis, albeit usually this is due to censoring, not convergence issues.

Unfortunately, Aalen's additive hazards model exhibits yet another problem if it were to be used as discussed so far. Due to the convergence issues discussed above, the lowest available value for the survival function for many data points is above 0.5, so the predicted median value is infinite. If these data points are predicted to recover within 24 hours, the MAE will be infinite, even if some recovery time predictions were finite. If, however, we do not attempt to optimize the classification threshold, and instead use the median directly as detailed above, these cases would be predicted to recover after 24 hours, so they would not be involved in the MAE calculation. We thus use 0.5 as the classification threshold for Aalen's additive hazards models. Although this is suboptimal, as now classification is not equivalent in all models, we deem this to be a better alternative over discarding Aalen's additive hazards from consideration entirely. Naive classification will impact the classification metrics of interest, so it is still possible to evaluate all models jointly.

Determining the optimal model is complicated by all of the above. Classification of the recovery into the two categories impacts the point prediction error, as we consider the latter only on the data points where the recovery is predicted to occur within 24 hours. With this in mind, it is important to consider the mean absolute error in conjunction with classification metrics.

A benefit of modelling the entire distribution is that we can perform classification and prediction conditional on a turbine being broken a certain amount of time. This is enabled by the simple observation

$$P(T > t|T > s) = \frac{P(\{T > t\} \cap \{T > s\})}{P(T > s)} = \frac{P(T > t)}{P(T > s)} = \frac{S(t)}{S(s)}.$$

Thus, to classify and predict with an account of how long a turbine has been broken, we simply divide the predicted survival function for that particular event by the value of the survival function at the conditioning time. The resulting function is a monotonically decreasing function from 1 to 0, so it is a survival curve as well, see Figure 6 for a visualization.

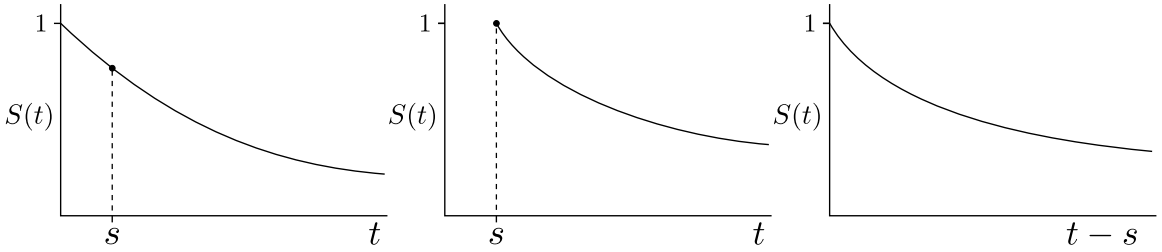


Figure 6: Obtaining a survival function conditioned on not recovering for s units of time from the unconditional one. The timeline can be redefined to be the remaining recovery time, which allows the survival function to start at time 0.

The optimal model would be good not only at classification and point prediction at time 0, but also conditioned on having not recovered for a certain amount of time. This is especially important since the largest differences in the models manifest in their ability to capture the tail of the survival function. It makes little sense to include observations where the true recovery time is below the conditioning time in the computation of the conditional metrics. As a consequence, the number of data points considered will decrease with increasing conditional time. Such a choice reflects the real world needs, where a prediction for a turbine that is known to have recovered would not be necessary.

To assess whether differences in the classification metrics and MAE between models are significant, we will employ bootstrapping to compute confidence intervals for them. Bootstrapping works by assuming that the distribution of the available data follows that of the population, which allows us to construct new data sets by sampling from the empirical cumulative distribution function, originally introduced by Efron in [28]. Unfortunately, it is computationally infeasible to carry out this approach with bootstrap data sets equal in size to that of the original training data set. To compute the confidence intervals, we will obtain the bootstrap data set by subsampling with replacement from our training data set, fit each model on the bootstrapped data, and compute metrics for the resulting models on the test data as before. The above will be repeated $b = 50$ times, to obtain 50 values of each considered metric. For each of the metrics evaluated, we perform a graphical normality check by means of QQ-plots. This allows us to construct asymptotic confidence intervals of the form

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\frac{\sum_{i=1}^{b}(\hat{\theta}_i^* - \hat{\theta})^2}{b(b-1)}},$$

where $\hat{\theta}$ is the estimate of a metric from the full training data set and $\hat{\theta}_i^*$ is the estimate for a given bootstrap model, see [29, p. 110]. The subsample size for each of the bootstrap data sets will be taken to be 10 000, which allows for the subsequent model fitting and prediction to be computationally feasible. Unfortunately, issues are once again presented with the Aalen additive hazards model. On numerous bootstrap data sets, the model exhibits convergence issues accompanied by a 'linear algebra error' message. This results in defective survival functions, which in turn results in being unable to estimate the median. In such situations, all events are predicted to recover after 24 hours (median is treated as infinite in these cases), so no value is available for the MAE. Due to these issues, we do not present confidence intervals for the Aalen additive hazards models.

The above approach allows us to simultaneously perform classification and point prediction. If the problem were of just classifying immediately upon failure, resampling approaches could be employed to balance the data set. Unfortunately, we require the survival function to be that of the actual data, since all conditional survival functions must also be those of the actual data set as to enable conditional classification and point prediction.

We proceed by fitting our models on increasing numbers of variables, starting with just cluster and then cluster and logarithm of energy produced in the month before failure. To see if historical failures have an impact on recovery, we will also construct a 'history' variable corresponding to each cluster. These will count the number of times the corresponding failures occurred before the current failure. We will refer to the cluster, logarithm of energy, and all of the history variables as the full collection.

Base Models

We begin by fitting all of our models on just the cluster, representing the simplest models. Since there are 112 clusters, we end up with 111 dummy variables in each model. For all models but Aalen's additive hazards, we compute ROC curves, see Figure 7. Note that the Weibull model has the highest AUC value at 0.78, closely followed by Cox regression. All of the models have an AUC above 0.5, implying that all of them viewed as classifiers are better than random guesses. As detailed above, we find the optimal classification thresholds, see Table 1. Note that the optimal classification thresholds differ significantly from the naive choice of 0.5, that is, using the median as a threshold.

Table 1: Classification thresholds for models fitted with cluster as variable.

Cox	Weibull	Normal	Logistic
0.0295	0.0210	0.0380	0.0450

We now consider how accurately the model predicts recovery time on the events that were classified to recover within 24 hours. The results are presented in Table 2. Evaluated using the MAE, all of the models are quite similar at time 0, but the differences come to light when considering conditional predictions. At a first glance, Aalen's additive hazards model performs

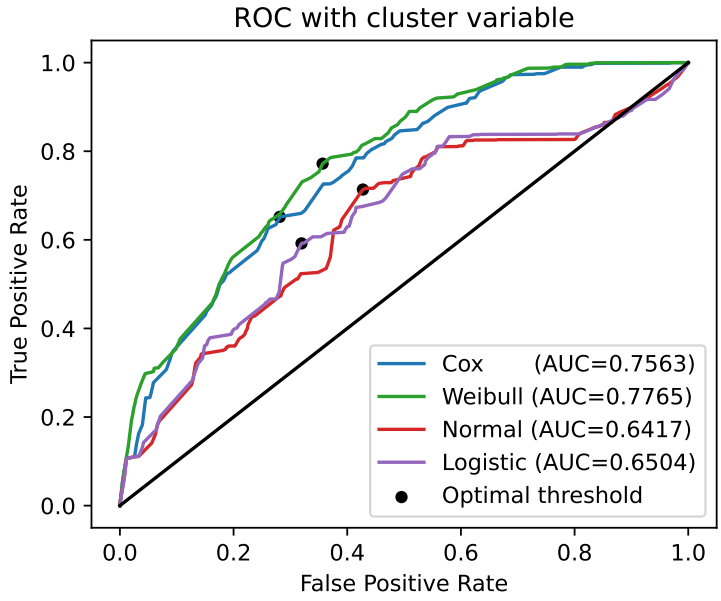


Figure 7: ROC curves for models fitted with cluster as the variable.

best conditioned on a turbine having been broken for 30 minutes. This, however, is misleading, since the MAE as computed by us, is in part a function of classification in the previous step. To make a more educated assessment, we inspect the confusion matrices for classification conditioned on not recovering for 30 minutes in Figure 8. Note that Aalen's additive hazards, Log-normal AFT and Log-logistic AFT models produce more False Positives than True Negatives. To evaluate how classification and MAE change with increasing conditioning time, we inspect the metrics of interest in Figure 9.

Table 2: Average prediction errors in hours for models fitted with cluster as variable.

	Unconditional		Conditional >30 min	
Model	MAE	RMSE	MAE	RMSE
Cox regression	2.725	26.150	3.322	22.335
Aalen's additive hazards	2.715	18.319	1.356	11.959
Weibull AFT	2.517	26.898	2.924	21.691
Log-normal AFT	2.458	16.192	6.179	28.081
Log-logistic AFT	2.751	18.164	3.412	19.029

Evaluating classification and MAE together produces a much clearer view of the differences between models. First, note that there are no MAE values for Aalen's additive hazards and Log-logistic AFT models if conditioned past 1 hour. This can be immediately seen to be a product of poor classification. Note that the sensitivity for these models is 1 for conditioning times past 1 hour, and that all events are classified to recover after 24 hours. For Aalen's additive hazards model this is a product of us setting the threshold to 0.5 due to convergence issues, and hence the classification very strongly preferring False Positives, thus resulting in never predicting that

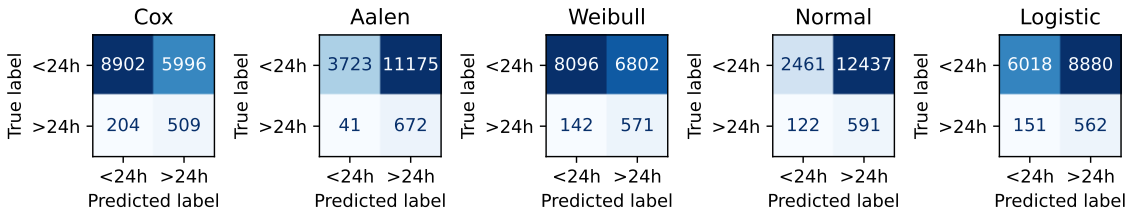


Figure 8: Confusion matrices for classification conditional on not recovering for 30 minutes for models fitted with cluster as variable.

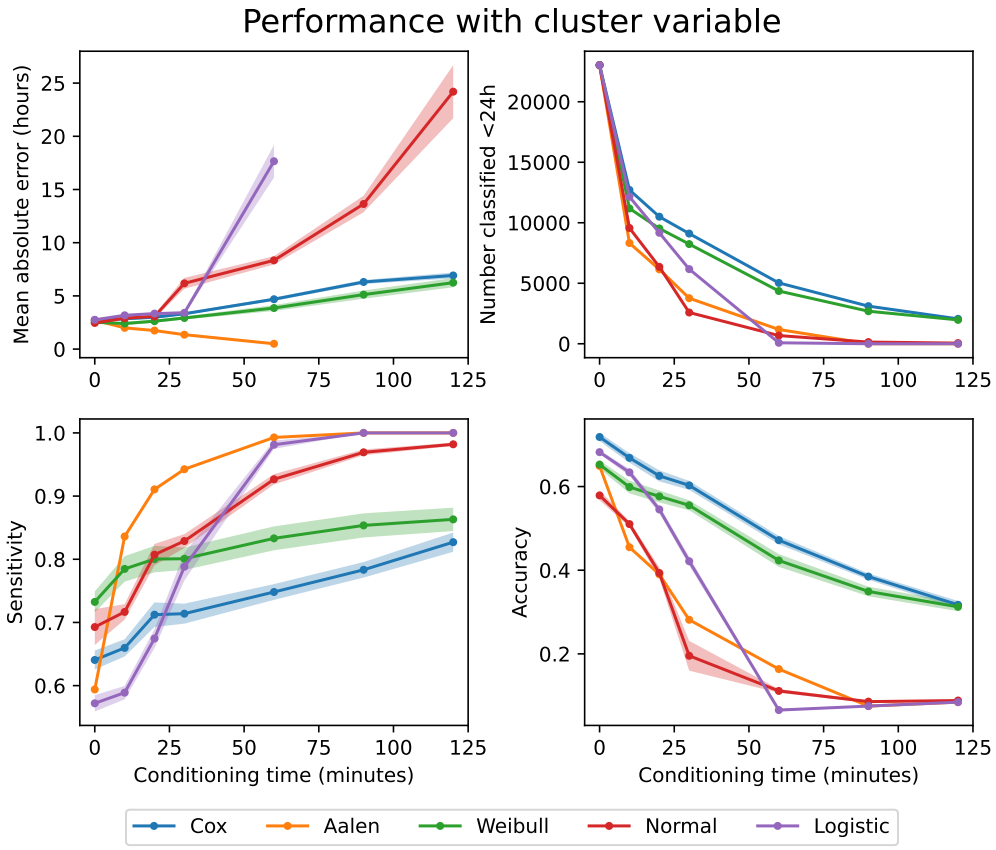


Figure 9: Performance of models fitted with cluster as variable for different conditioning times.

a turbine will recover within 24 hours. For the Log-logistic AFT model, this is likely best explained simply by the inability of it to model the tail of the survival function well. Similar reasoning applies to the Log-normal AFT model, which albeit classifies some turbines to recover within 24 hours, does this quite poorly, as seen by the sensitivity and accuracy values. High sensitivity together with low accuracy implies that the model produces a high number of False Positives. The Log-normal AFT model also evidently models the tail poorly, as seen by the comparatively worse MAE values for high conditioning times. Note that Cox regression and the Weibull AFT model produce quite similar results. In particular, the MAE, albeit increases

with conditioning time, does so at a considerably lower rate, indicating that the tail of the survival function is modeled better than for other models. The classification is also significantly better compared to other models.

We now consider the models with cluster and the logarithm of energy produced in the month before failure as variables. The modelling process is repeated as before. See Figure 10 for the ROC curves and Table 3 for optimal thresholds for each model. Note that the thresholds are quite similar to their counterparts for the models with cluster as the only variable. The ROC curves are now significantly more refined as a product of including a continuous variable as a covariate. The number of unique energy values in the data set is significantly higher than the number of clusters, thus resulting in a larger possible number of survival curves and hence candidate thresholds. Note also that the AUC values have increased for all models, albeit insignificantly, and that optimal thresholds are quite similar to their counterparts obtained from models with just cluster as a variable.

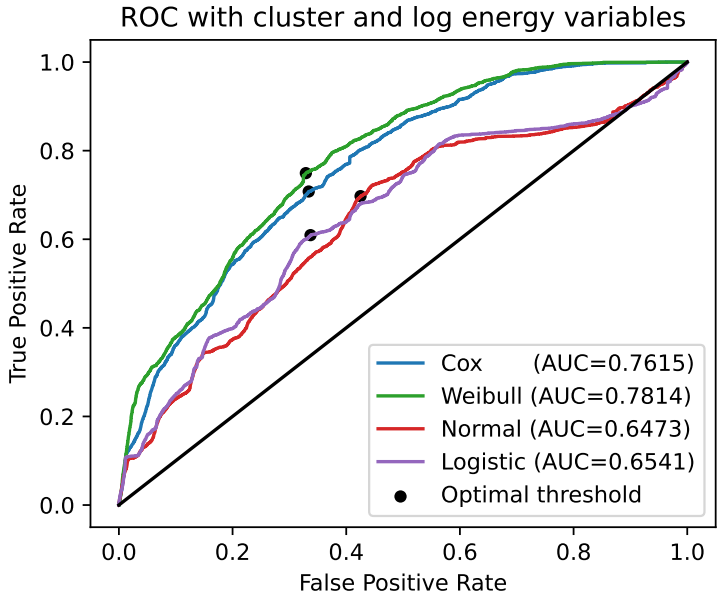


Figure 10: ROC curves for models fitted with cluster and logarithm of energy as variables.

On the test data we perform classification and point prediction as before. Note that there is a very marginal improvement in both the conditional and unconditional MAE values for most models as compared to the models with just cluster as a variable. There is a significant reduction in the MAE for the Log-normal AFT model conditioned on not having recovered for 30 minutes. Inspecting the confusion matrices in Figure 11, we can see that with 30 minutes conditioning the general classification has not changed drastically. Note that with Cox regression there is a lower number of True and False Negatives, whereas with the Weibull AFT model there are fewer False Negatives but simultaneously more True Negatives. As in the cluster only equivalents, Aalen's additive hazards, Log-normal AFT, and Log-logistic AFT produce considerably more False Positives than True Negatives.

As before, the complete picture is given by considering classification performance simultaneously, see Figure 12. Aalen's additive hazards and Log-logistic AFT experience the same issues as before, with extremely poor classification resulting in no MAE for conditioning times past 30 minutes. The improvement in the MAE for the Log-normal AFT model mentioned above is even more clear when considering conditioning on 2 hours, where a reduction of nearly 5 hours is achieved as compared to the same model fitted on just cluster as the variable. The Weibull AFT model and Cox regression are once again significantly better than other models. Of note is that both models classify much more similarly, with a near identical accuracy and number of events classified to recover within 24 hours. There is also remarkable similarity in the resulting conditional MAE.

It is reasonable to say at this stage that Cox regression and the Weibull AFT model perform significantly better than other models, so we can move forward with just what has currently been established as the best. We however opt to give the benefit of the doubt to the other models, and evaluate them on the full collection of variables to see if there is any significant impact.

Table 3: Classification thresholds for models fitted with cluster and logarithm of energy as variables.

Cox	Weibull	Normal	Logistic
0.0250	0.0224	0.0377	0.0425

Table 4: Average prediction errors in hours for models fitted with cluster and logarithm of energy as variables.

	Unconditional		Conditional >30 min	
Model	MAE	RMSE	MAE	RMSE
Cox regression	2.656	27.009	3.005	18.725
Aalen's additive hazards	2.706	16.172	1.352	11.965
Weibull AFT	2.511	26.410	3.042	31.895
Log-normal AFT	2.542	16.389	4.905	25.035
Log-logistic AFT	2.640	15.879	3.267	18.670

Full Models

We now consider models that are on the other end of complexity. For each of the possible clusters, we introduce a 'history' variable which records how many times the turbine has experienced a failure in the cluster. If a turbine has been experiencing certain problems before, this might impact its recovery time.

As in the previous section, we compute ROC curves and find optimal thresholds, see Figure 13 and Table 5 respectively. Note that once again there is a marginal improvement in AUC

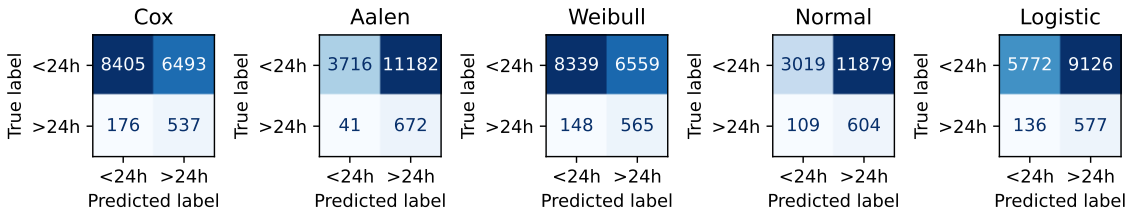


Figure 11: Confusion matrices for classification conditional on not recovering for 30 minutes for models fitted with cluster and logarithm of energy as variables.

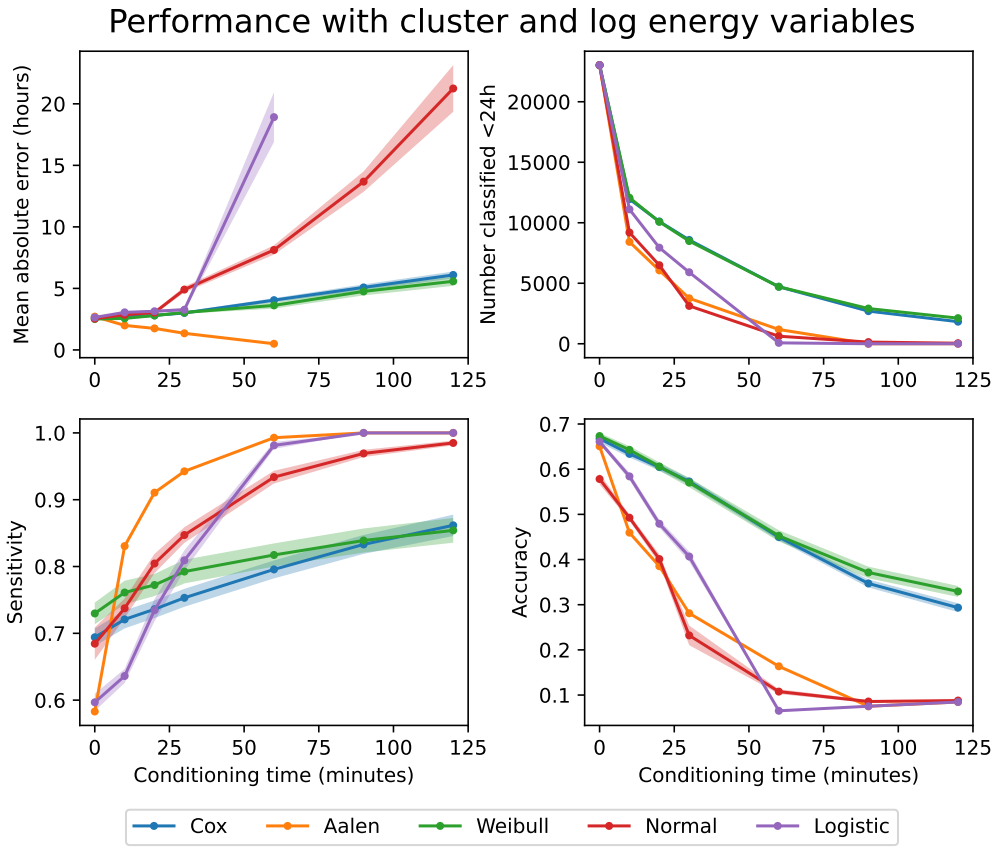


Figure 12: Performance of models fitted with cluster and logarithm of energy as variables for different conditioning times.

values for Cox regression and the Weibull AFT model. The changes in AUC for Log-normal and Log-logistic models is negligible in comparison. The analysis of these models follows near verbatim from previous sections. Of note is a further minor improvement in unconditional MAE for Cox regression and the Weibull AFT model. The improvement in MAE conditional on 30 minutes is notable for the Weibull AFT model. Looking at the confusion matrices in Figure 14 we can see that the number True and False Negatives has increased for both Cox regression and the Weibull AFT model. Similar classification as before is exhibited by Aalen's additive hazards, and the Log-normal and Log-logistic AFT models.

Once again, we evaluate the models using MAE and classification metrics simultaneously, see Figure 15. There is a notable improvement in MAE conditional on 2 hours for the Log-normal AFT model, reducing to under 15 hours. As before, Cox regression and the Weibull AFT model are quite similar, with the notable difference being in sensitivity. Given the similar accuracy for the models, a higher sensitivity is preferred, since in our case False Positives are generally preferred over False Negatives.

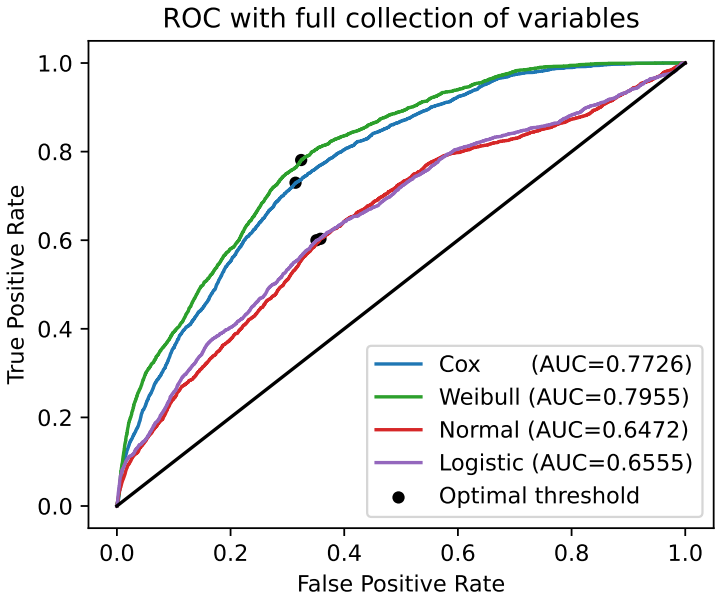


Figure 13: ROC curves for models fitted with the full collection of variables.

Table 5: Classification thresholds for models fitted with the full collection of variables.

Cox	Weibull	Normal	Logistic
0.0261	0.0219	0.0387	0.0399

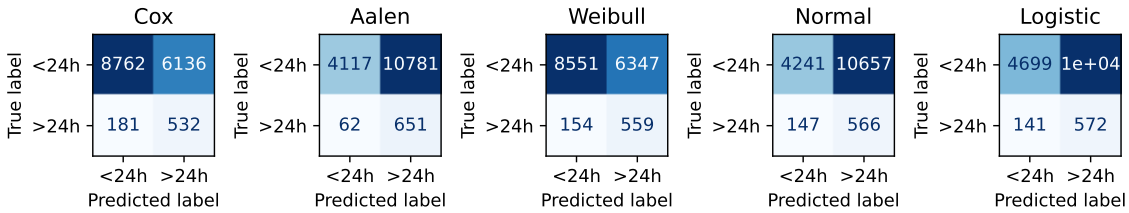


Figure 14: Confusion matrices for classification conditional on not recovering for 30 minutes for models fitted with the full collection of variables.

Table 6: Average prediction errors in hours for models fitted with the full collection of variables.

	Unconditional		Conditional >30 min	
Model	MAE	RMSE	MAE	RMSE
Cox regression	2.336	16.685	2.934	18.196
Aalen's additive hazards	2.761	17.661	1.907	17.878
Weibull AFT	2.430	41.177	2.682	14.697
Log-normal AFT	2.945	19.143	4.243	21.601
Log-logistic AFT	2.755	18.502	3.813	20.485

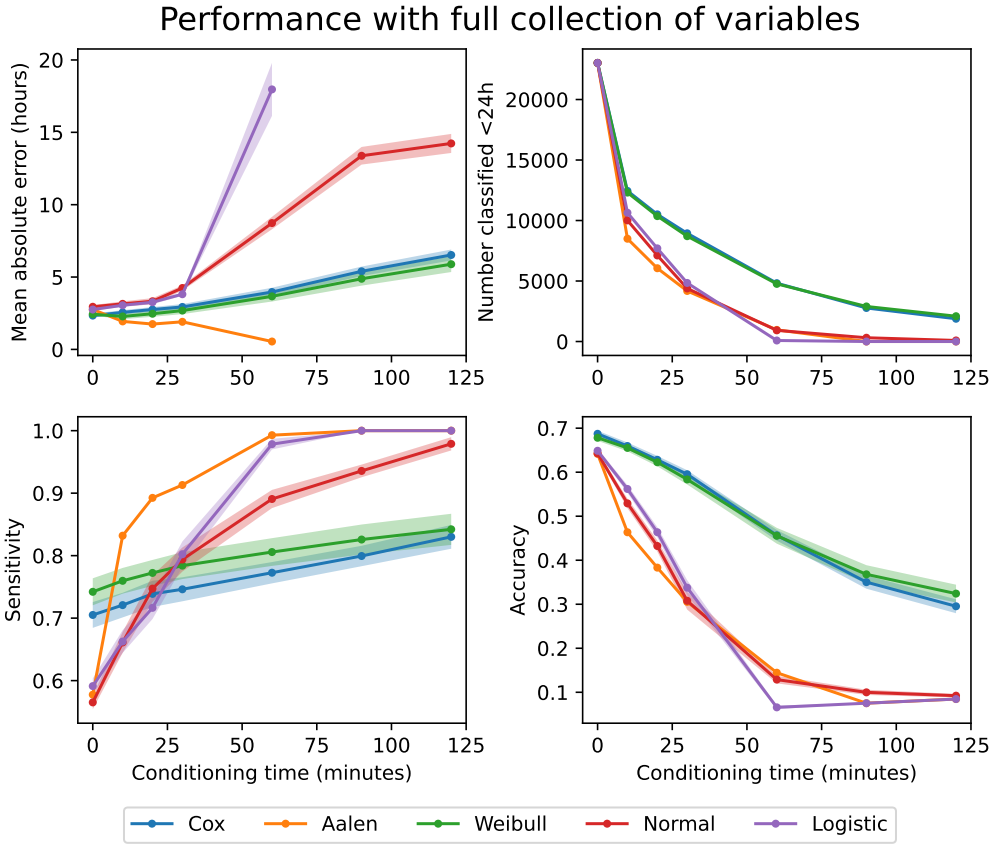


Figure 15: Performance of models fitted with the full collection of variables for different conditioning times.

Variable Selection and Subsequent Models

It is likely that not all history variables are equally significant in the models. Unfortunately, there is no simple way to perform variable selection in Aalen's additive hazards model. The values of coefficients change with time, so the significance of a covariate in the model is hard to quantify. A variable might be significant during some times of the recovery, but become entirely insignificant during others. As Aalen's additive hazards model has exhibited numerous issues and performed poorly overall, we discard it from consideration from this point onwards.

A potential approach to selecting significant variables could be stepwise AIC or BIC selection procedures. Unfortunately, these are computationally infeasible with the size of the data sets and number of variables involved. We thus opt for a significantly simpler approach, where we will select variables that are deemed significant in the models. As we wish to consider the significance of multiple variables simultaneously, the problem we are met with is multiple testing. As discussed in the *Significance of Variables and Multiple Testing* section, simply selecting all variables with p-values below a threshold would result in a high Family-Wise Error Rate (FWER). To address this issue, we will use the Holm-Bonferroni procedure, with the desired FWER being 0.05, as this is a fairly standard significance level for individual tests. Note that we omit from consideration any intercept terms and the dummied variables associated with the individual clusters. Omission of the cluster from consideration is justified by our clustering done before, which we deem to be good enough.

Of note is that in all of the models the logarithm of energy produced in the month before failure was deemed significant. Of the 112 history variables, in Cox regression 33 were significant, 40 in the Weibull AFT model, 32 in the Log-normal AFT model, and 31 in the Log-logistic AFT model.

We then refit the models on the corresponding significant variables, thus ending up with cluster, logarithm of energy, and selected history variables. As for the models above, we compute ROC curves and optimal thresholds, see Figure 16 and Table 7 respectively. There is a minor reduction in AUC values, but this tradeoff could be justified by the considerably smaller models.

Table 7: Optimal classification thresholds for models fitted with variables deemed significant using the Holm-Bonferroni procedure.

Cox	Weibull	Normal	Logistic
0.0258	0.0260	0.0357	0.0410

As can be seen in Table 8, both the unconditional and conditional on 30 minutes of being unrecovered MAE's have increased for all models but for Log-normal AFT. Inspecting the confusion matrices in Figure 17, we see that for Cox regression, the number of True Negatives has decreased while the number of False Negatives has increased. For the Weibull AFT model, we observe slightly better behavior, where the number of both True and False Negatives has increased.

Inspecting how MAE and classification changes with increasing conditioning time in Figure

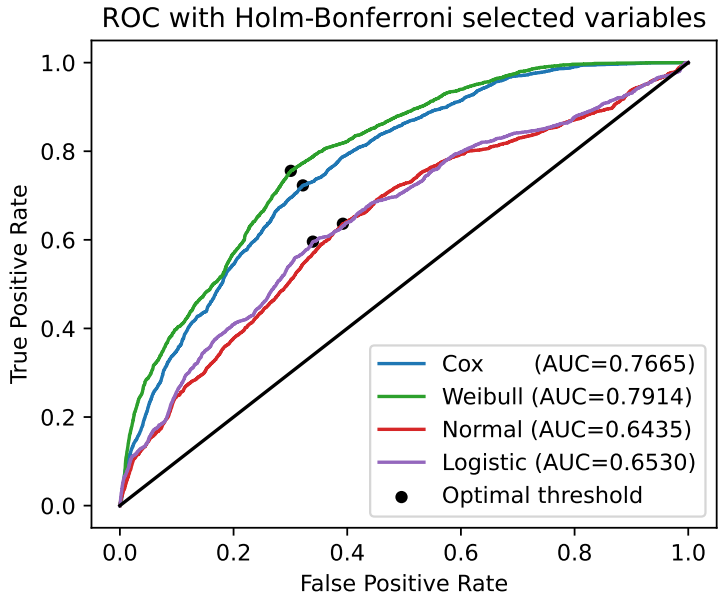


Figure 16: ROC curves for models fitted with variables deemed significant by the Holm-Bonferroni procedure.

Table 8: Average prediction errors in hours for models fitted with variables deemed significant using the Holm-Bonferroni procedure.

	Unconditional		Conditional >30 min	
Model	MAE	RMSE	MAE	RMSE
Cox regression	2.405	17.058	3.133	21.803
Weibull AFT	2.512	26.001	2.804	14.816
Log-normal AFT	2.771	16.728	4.893	24.727
Log-logistic AFT	2.827	18.764	3.802	20.171

18, we can see that the behavior is very similar to that of the full models.

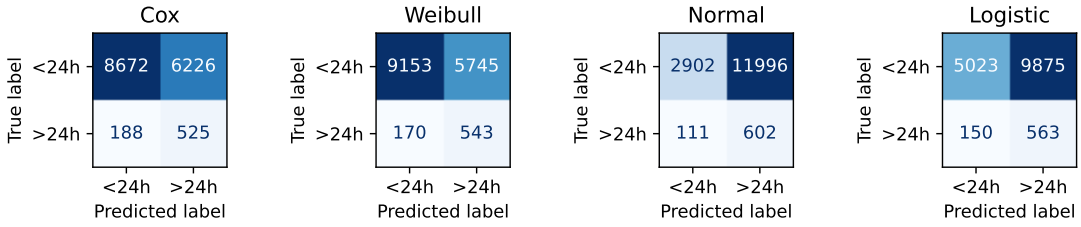


Figure 17: Confusion matrices for classification conditional on not recovering for 30 minutes for models fitted with variables deemed significant using the Holm-Bonferroni procedure.

The differences between various models have been larger than any change due to a different

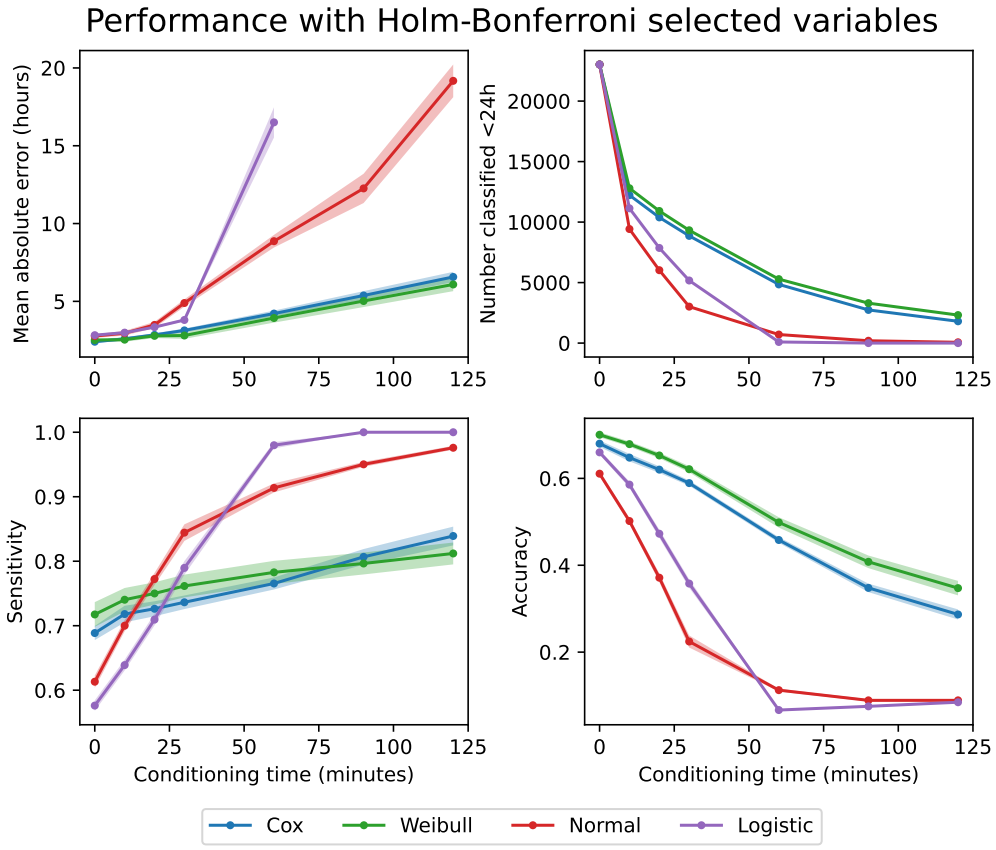


Figure 18: Performance of models fitted with variables deemed significant using the Holm-Bonferroni procedure for different conditioning times.

set of included variables. Cox regression and the Weibull AFT model have consistently been significantly better than Aalen's additive hazards and the other AFT models. The proportional hazards assumption evidently holds in the data set at least to some extent, given the comparatively good classification and prediction.

Given the similarity between the Cox regression and the Weibull AFT models, we opt for the Weibull AFT as the optimal model. Across various variable inclusions, the Weibull AFT model is consistently better at classification than Cox regression, having a higher sensitivity with a near identical, or higher accuracy, as in the case for Holm-Bonferroni selected variables. There is a further benefit of the Weibull AFT model in that it is fully parametric. The Breslow estimator in Cox regression is incapable of producing an estimate of the cumulative baseline hazard past the last available observation, so we are incapable of classifying and predicting far out into the tail of the survival function. The benefit of this is up for debate, as the fit to the tail of the Weibull AFT model, even if better, would likely produce predictions of a nearly useless quality.

Comparison of Weibull Models

As the differences between the Weibull models were hard to distinguish when presented in the context of significantly inferior models, we present them here once again on their own. As can be seen from Figure 19, the ROC curves are largely similar for the Full and the Holm-Bonferroni Weibull models, and similarly for the cluster and cluster with logarithm of energy models.

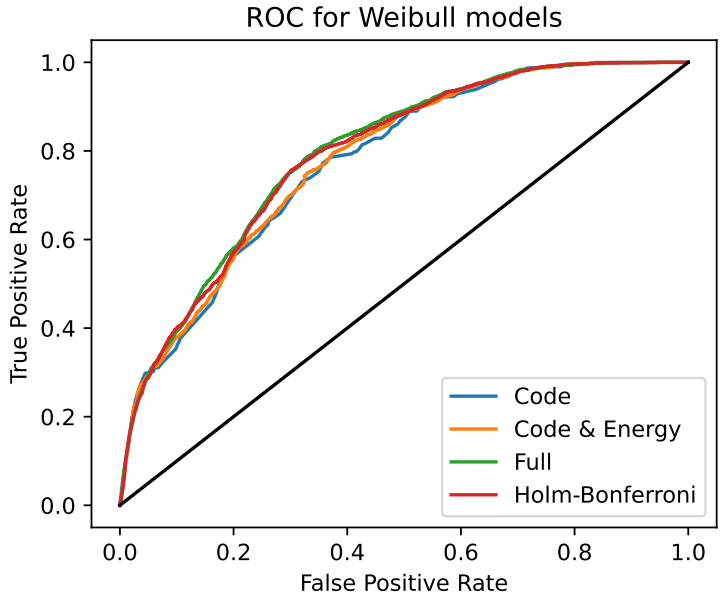


Figure 19: ROC curves for the Weibull AFT models considered.

See Figure 20 for the performance of models with increasing conditioning times. In particular, of interest is that the model with Holm-Bonferroni selected variables has higher accuracy and lower sensitivity. Otherwise, more variables included in the model results in higher accuracy and lower sensitivity.

We investigate further by considering the confusion matrices for the models at various conditioning times, see Figure 21. The number of False Negatives, that is the number of incorrect predictions of recovery within 24 hours, increases with the order in which models were presented. The high accuracy but low sensitivity is a reflection of the unequal number of classes. Albeit the model with Holm-Bonferroni selected variables has the highest accuracy, this comes at a cost of notably more False Negatives, which for this application is a downside.

The Holm-Bonferroni model has too low of a sensitivity, whereas the full model has accuracy near equivalent to that of the model with just cluster and energy, with a lower sensitivity. Moreover, the cluster and energy model exhibits the lowest MAE for conditioning times past 30 minutes. With this in mind, we choose the model with cluster and logarithm of energy produced in the month before the failure as the optimal model.

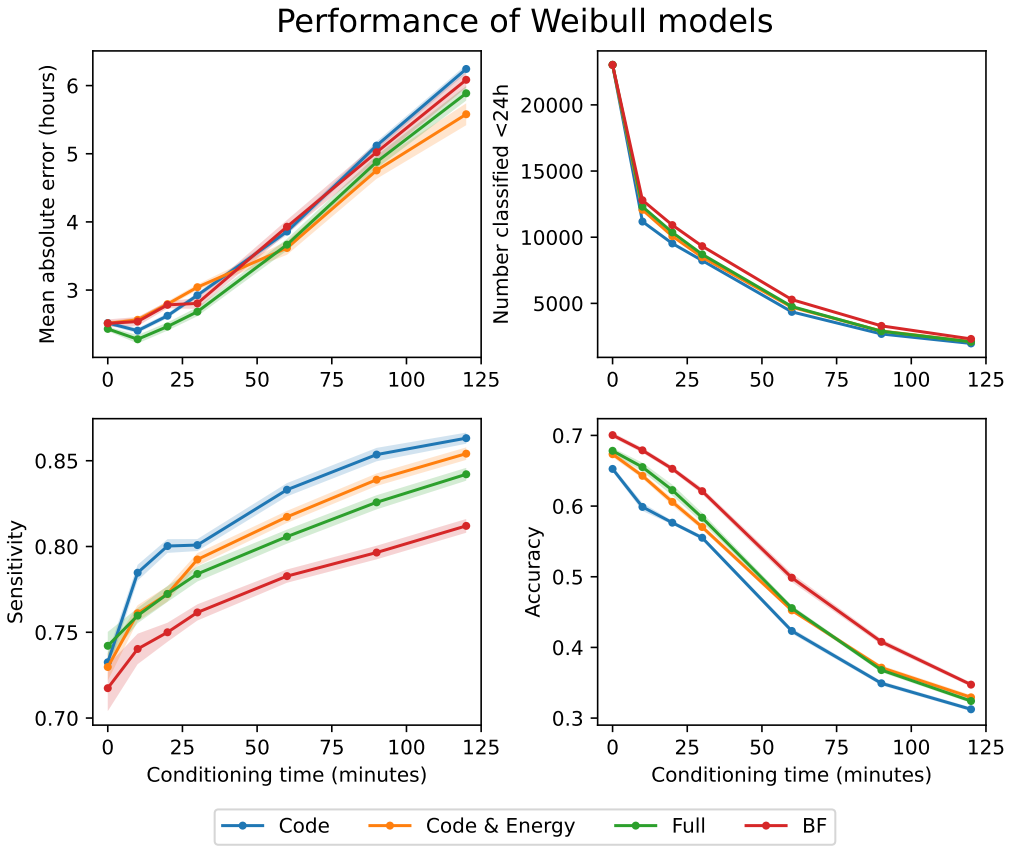


Figure 20: Performance of the Weibull models considered for different conditioning times.

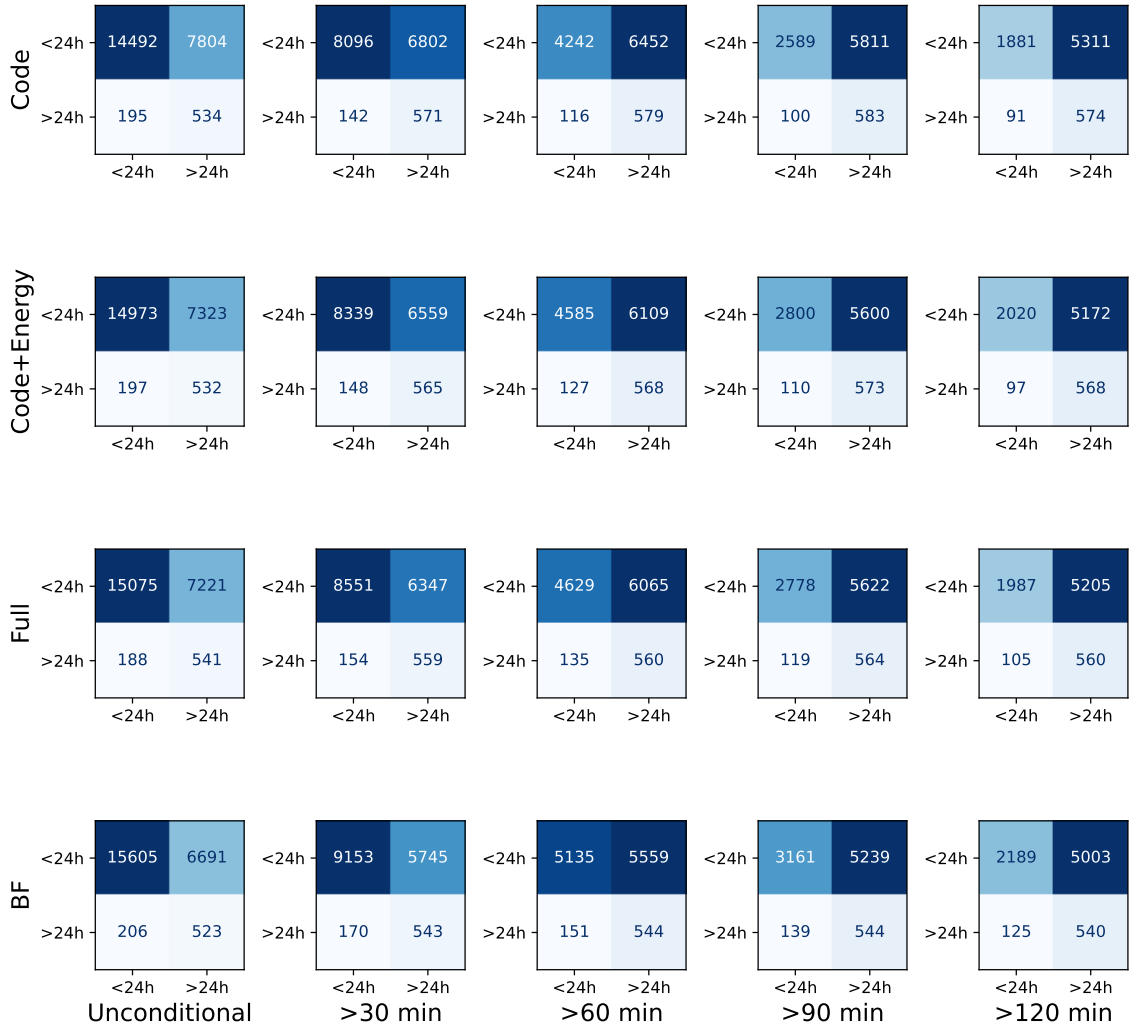


Figure 21: Confusion matrices for the Weibull models considered for various conditioning times.

Comparison to Naive Approaches

As is the case with many time-to-event data sets, the data examined in this work is very unbalanced. This complicates the interpretation of results, since it is not trivial to say if a given accuracy or sensitivity metric is 'good'. Within the models considered, the Weibull AFT model with the clusters and logarithm of energy produced in the month before failure performed the best. However, this result is of little importance if some naive approach to the problem can perform better. We thus evaluate the best model found to two naive approaches.

The first naive 'model' will always classify events as resolving within 24 hours. The second will classify recovery into under 24 hours and over 24 hours randomly. In the case that the prediction is that the turbine will recover within 24 hours, both models will predict recovery in 12 hours, as this is the average recovery in the 24 hour window. As can be seen in Figure 22, the Weibull AFT model has a considerably better MAE while striking a more favorable balance between sensitivity and accuracy. We can thus fairly confidently say that the Weibull AFT model found to be best is considerably better than a truly naive approach.

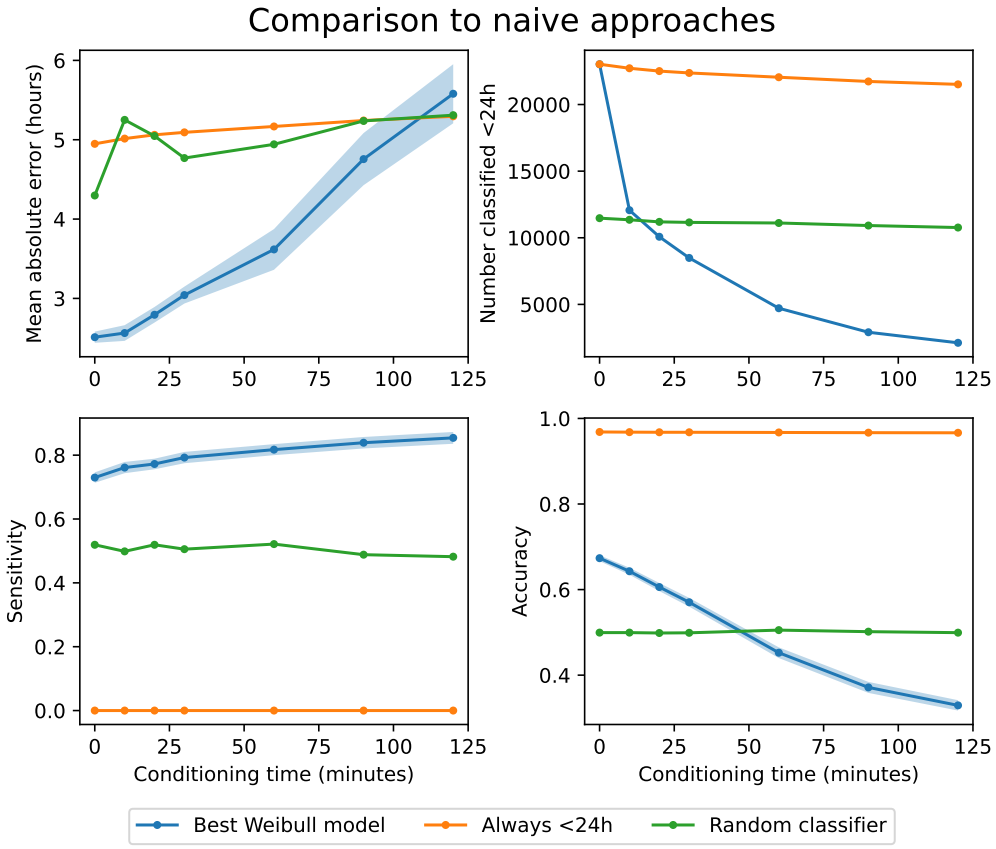


Figure 22: Comparison of the chosen model with naive approaches.

Conclusion

In this thesis we have attempted to present a picture of some models and methods used for wind turbine availability forecasting. The clustering approach and the use of Aalen's additive hazards model necessitated an overview of martingales and variation processes. To our knowledge, the clustering method presented is novel.

Challenges were presented by the different nature of the models, and as a consequence the evaluation of models relied on quite simple metrics. It is of great practical benefit that the best model is fully parametric and does not have too many parameters. Cox regression could be viewed as a good alternative, but fitting and prediction takes a considerably longer time. Aalen's additive hazards model turned out to be quite troublesome to work with, mainly due to its inability to produce an estimate of the tail of recovery time for many failure events. This prompted us to use a naive classification threshold, equivalent to using the median. Of interest is that despite this choice, unconditional predictions are on par with the other models. Albeit these results were not discussed above to limit the scope, Cox regression and AFT models performed significantly worse with the naive classification threshold, even for unconditional prediction.

There is a notable number of limitations associated with our work. Focus was placed on turbine failures deemed interesting enough to study, which reduced the size of our data set by a magnitude. This was in part a computational consideration – semi- and nonparametric models exhibited issues with memory allocation for data sets larger than the one used in the thesis. Furthermore, turbines of only one customer were studied, mainly due to complexities associated with merging data from many sources. This exposes the risk that our models do not translate to other customers well. A given customer could be seen as a proxy for many important factors like maintenance budgets, accessibility of turbines, and general weather effects their turbines are exposed to.

Plenty of further research can be done into the topic. As our work set out to compare five very different models, we were limited by having to compare them on common ground. Taking the Weibull AFT model as the only candidate, there are several immediate improvements that can be made. Note that we computed ROC curves and resulting optimal thresholds for classification from unconditional survival functions, that were then used to classify and predict from conditional survival functions. It is reasonable to suspect that the optimal threshold would increase with higher conditioning time, as the classes become more balanced. A way to address this could be to find optimal thresholds for a select number of conditioning times, and restricting prediction to those times. Alternatively, the change in optimal thresholds with conditioning time could be studied, and perhaps interpolated. This would allow for prediction for any conditional time with a somewhat optimal threshold.

The significance of many other variables could be studied. For example, seasonality could

affect the ease of access to turbines and hence recovery times. A turbine might take longer to fix if it breaks over the weekend. The use of many of these additional categorical variables was attempted, but unfortunately convergence was an issue. It could also be of interest to study if there is any spatial dependence of recovery times, that is if the location of the turbine experiencing a failure has much impact on recovery. A simpler approach in a similar vein would be to attempt to quantify ease of access to and environmental impact on turbines and use those quantities as variables in the model.

In this work, apart from a quick check on whether the proportional hazards assumption holds in our data, we completely omitted the study of residuals. This was a conscious choice done in part due to the broad scope associated with evaluating five different models. That said, the study of residuals could further justify including certain variables, and would be much easier in the context of studying only the Weibull AFT model. A further criticism could be the use of survival analysis approaches overall. As our filtered data set contains one censored and over $100\,000$ uncensored observations, negligible bias would be introduced by discarding the censored observation, and analyzing the data with more standard methods. However, as discussed before, our approach allows us to translate the methodology to a different data set, where the censoring rate may be considerably higher.

Stepwise selection procedures could be attempted with the Weibull AFT model. An attempt on our behalf was made to perform a forward BIC selection with the Weibull model, but unfortunately in one of the steps a model failed to converge. In our case, the process was also extremely time-consuming due to the sheer number of variables involved and the overall size of the data set.

Before fitting any models we clustered error codes based on their distributions. As we did not wish to impose any assumptions on the recovery time distributions in this step, we chose a fully nonparametric approach. Of further interest could be an evaluation of our clustering method as compared to other more common approaches, e.g. K-means or K-medians. These were not considered in our work because it was deemed fairly reductive to cluster the error codes based on one point of the survival curve. Furthermore, more study into optimal choices for the number of iterations and the cutoff p-value is necessary, as our choices were a heuristic decision justified by what seemed 'good enough'.

In this thesis we exclusively used the median for point prediction, as this choice retains a reasonable level of interpretability. The same cannot be said of the classification, where we opted to tune thresholds by considering ROC curves for each model. Something similar could be done for the quantile used for predicting the recovery time. We decided not to pursue this idea as it was not readily apparent that it is theoretically sound to do so, unlike for classification where it is a near standard approach for cases with unbalanced classes. Furthermore, the use of expected value as a point predictor could be investigated; since the Weibull AFT model is fully parametric, there is no issue with integrating the survival function. For classification, alternate approaches to choosing the optimal threshold could be studied.

Another way to approach point prediction would be to compute expected recovery time values not from the unconditional distribution, but where the classification decision is accounted for. In practice, this would mean finding the expected value of the random variable $T|T \leq 24\text{h}$,

where as usual T denotes the recovery time.

Survival analysis packages such as Python's `lifelines`, that was used in this thesis, offer the ability to regularize the models. Regularization could help prevent overfitting and hence improve performance on new data. Furthermore, one could evaluate how an ensemble of different Weibull AFT models performs. Classification could be decided by a majority vote and point prediction by a simple average.

In conclusion, albeit plenty of further research can be carried out, the Weibull AFT model with cluster and logarithm of energy produced in the month prior to failure already delivers quite excellent results. The model is significantly better than naive approaches at predicting whether a turbine will recover in 24 hours or not, and the prediction MAE of 2.5 hours can be deemed to be quite good. To assess the usefulness of the model in a practical context, case studies need to be carried out. These would use the presented or further improved model as a step in forecasting energy export for a customer, wind farm, or region. The model's usefulness would then be readily apparent if the corrections it offers to the predicted energy export are significant.

Bibliography

[1] X. Wang, P. Guo, and X. Huang. A review of wind power forecasting models. *Energy procedia*, 12:770–778, 2011.

[2] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable energy*, 37(1):1–8, 2012.

[3] G. Sideratos and N. D. Hatziargyriou. An advanced statistical method for wind power forecasting. *IEEE Transactions on power systems*, 22(1):258–265, 2007.

[4] Y. Teng, Q. Hui, Y. Li, O. Leng, and Z. Chen. Availability estimation of wind power forecasting and optimization of day-ahead unit commitment. *Journal of Modern Power Systems and Clean Energy*, 7(6):1675–1683, Nov 2019.

[5] M. Greenwood. The first life table. *Notes and Records of the Royal Society of London*, 1(2):70–72, 1938.

[6] J. Tobacman. Paul meier, 1924 - 2011, Aug 2011. Chicago Tribune.

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

[8] Citation classic - nonparametric estimation from incomplete observations.

[9] M. Greenwood. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926.

[10] W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[11] O. Aalen. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701 – 726, 1978.

[12] P. K. Andersen. *Statistical Models Based on Counting Processes*. Principles of Pediatric Neurosurgery. Beijing World Publishing Corporation, 1993.

[13] T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analysis*. Wiley Series in Probability and Statistics. Wiley, 1991.

[14] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

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

[16] W. Nelson and G. J. Hahn. Linear estimation of a regression relationship from censored data part i—simple methods and their application. *Technometrics*, 14(2):247–269, 1972.

[17] O. Aalen, O. Borgan, and H. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer New York, 2008.

[18] J. L. Doob. *Stochastic Processes*. A Wiley-interscience publication. Wiley, 1990.

[19] M. Beiglboeck, W. Schachermayer, and B. Veliyev. A short proof of the doob-meyer theorem, 2010.

[20] L. T. Nielsen. A note on absolutely continuous processes, 2019.

[21] R. Rebolledo. Central limit theorems for local martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 51(3):269–286, Jan 1980.

[22] J. Wang. Smoothing hazard rates. 2005.

[23] H. Muller and J. Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):61–76, 1994.

[24] J. P. Klein and M. L. Moeschberger. *Survival analysis: Techniques for censored and truncated data*. Springer, 2003.

[25] D. Panchenko. Lecture notes in statistics for applications, lecture 3. 2006.

[26] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[27] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[28] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[29] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, 2004.

Master's Theses in Mathematical Sciences 2023:E67
ISSN 1404-6342

LUNFMS-3123-2023

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/