# Survival analysis and Predictive Maintenance Models for non-sensored Assets in Facilities Management

Genevieve Moat
School of Computing Science
Newcastle University
Newcastle upon Tyne, NE1 7RU
g.j.moat@ncl.ac.uk

Shirley Coleman
School of Mathematics, Statistics and Physics
Newcastle University
Newcastle upon Tyne, NE1 7RU
shirley.coleman@ncl.ac.uk

***Abstract*—Maintenance is a critical component of Facilities Management (FM), and with the proliferation of big data, Internet of Things (IoT), and Industry 4.0, predictive maintenance (PdM) has emerged as a critical maintenance technique. However, modern data-driven PdM tactics are based on sensor data, but there is no obvious way to imply PdM on older buildings that lack sensors. EQUANS is a company seeking recommendations for implying PdM in the management of a historic building. This paper demonstrates the potential of survival analysis with data-driven PdM using EQUANS's non-sensored data, explicitly using the Kaplan-Meier method, parametric methods, Cox proportional hazard model, and accelerated failure time models. The boiler was chosen as the asset to focus on in this project, and the results indicated that the boiler's survival might not be related to the frequency of service but the boiler's age. The research findings propose a further step toward PdM for assets without sensors, and data collection and preventive maintenance can be improved.**

***Index Terms*—Predictive maintenance, Non-sensored data, Survival analysis, Data science**

## I. Introduction

Maintenance is essential for providing facilities management (FM) services for public buildings and schools assets, as maintenance costs are a significant component of annual FM costs. However, good maintenance practices can minimise building maintenance costs and lengthen component life [1]. There are three main types of maintenance:

- Reactive maintenance (Unplanned maintenance)
- Planned preventive maintenance (Time-based maintenance)
- Predictive maintenance (Condition-based maintenance)

Building maintenance management currently uses both preventive and reactive maintenance. Reactive maintenance is maintenance performed after an asset has failed. Planned preventive maintenance (PPM) is a time-based strategy that uses a predetermined interval of time or operating hours to prevent failure. Companies gradually realised how costly and inefficient these two approaches were, especially reactive maintenance, which usually costs more and takes longer to repair than PPM. Also, PPM cannot predict an asset's future condition and cannot prevent certain failures. In comparison, predictive maintenance (PdM) predicts impending breakdown by monitoring assets' condition. It allows the FM team to perform necessary maintenance, avoid failures, and reduce parts and labour costs [2], [3].

This project involves EQUANS (used to be ENGIE) in collaboration, and the company desire to transfer their FM service from reactive and preventive maintenance to more PdM approaches. Places & Communities north at EQUANS delivers several FM services to local councils. The concern is that EQUANS takes over long-standing FM contracts with other companies, then mobilises the contracts and sets up the system but never thinks about how they gather the data; the result is that unpractical data is collected. Consequently, they seek advice on improving data collecting, preventative maintenance, and predictive maintenance methodologies.

The fundamental objective of this project is to develop a method that applies to any asset that feeds into the PdM concept and is based on EQUANS's FM system data. Thus, at the very least, the method can help the decision-making for the PPM service to prevent failure.

This paper is organised as follows. In the next section, the background of PdM and details of EQUANS's datasets will be provided, along with problems and challenges encountered in this project. Section III reviews the literature of predictive maintenance based on the statistical method. The data cleaning and extraction will be provided in section IV, then the methodologies will be discussed in section V and appropriate approaches will be defined, followed by results based on the methodologies in section VI, then a discussion will be given in section VII. Finally, the paper finishes with conclusions and recommendations for EQUANS in section VIII.

## II. Background

### A. Background of PdM

A PdM program's three critical processes are data collecting, data processing, and maintenance decision-making [4]. PdM can be implemented as knowledge-based, which means enlisting domain experts' assistance to construct PdM models, or data-driven, which entails learning from existing data [5].

However, in this project, we only evaluate the data-driven method with no or minimal domain knowledge.

There are two ways to assess asset conditions: continuous surveillance (using sensors) and periodic inspection [1]. Techniques of data-driven PdM include process parameter measurements, vibration analysis, oil analysis, thermal analysis, acoustic analysis, and others [3]. Moreover, interpreting those techniques often includes creating machine learning or deep learning models to estimate fault events [6], [7], remaining useful life [8], [9], and plant conditions to make PdM decisions [1].

In 2001, Breiman [10] claimed that machine learning aims to predict future observations, whereas the statistical approach aims to extract associations between the response variables and the covariates.In this project, the datasets provided by EQUANS do not support machine learning due to the data content and quality (see the following subsection), typically having datasets that are too small to train and test. Thus, statistical methods seem to be most helpful and interesting in our situation.

Therefore, as an alternative to the machine learning aspect, statistical models can help in PdM. [11] proposes a generalised cost-effective condition-based model (CBM) that uses Markov decision processes to give an optimal policy associated with inspections schedule, maintenance action, and cost factor. [12] described a Bayesian-based system for learning and associating failure signals with the possibility of failure occurrences. It first identifies failure patterns and probabilities using Bayesian networks to create a more effective maintenance schedule and then tests on event-based data acquired from a reputable semiconductor manufacturer with promising results. In a more knowledge-based approach, [13] used a survival model to determine the optimal maintenance schedule in order to minimise the long-run average maintenance cost per replacement cycle. This was accomplished by developing a Weibull proportional hazard model to jointly model degradation and failure time data to forecast remaining useful life (RUL) and determine the optimal maintenance schedule.

### B. Datasets from EQUANS

EQUANS provided seven datasets from three contracts. All the details are included for completeness, although only a subset was used in this project. The datasets are:

Table I
LIST OF DATASETS PROVIDED BY EQUANS

| Contract A | | |
|---|---|---|
| **Datasets** | **Size** | **Description** |
| CA_con | 23948x96 | Conditions of each asset |
| CA_labour | 81936x55 | Records of PPM and reactive job |
| **Contract B** | | |
| CB_labour | 42952x22 | Records of PPM and reactive jobs |
| **Contract C** | | |
| CC_ppm | 6188x17 | Records of PPM |
| CC_labour | 10000x9 | Records of reactive job |
| CC_con | 8961x28 | Conditions of each asset |
| CC_cost | 76x2 | Summary cost |

Table II
INFORMATION INCLUDED IN EACH DATASET

| Datasets | Example information |
|---|---|
| CA_con | Condition, code, type description, site ID, assessment date. |
| CA_labour | Job type (reactive or PPM), job status, description of the type of PPM or failure of the asset, times and dates of when the job is created and the estimate/actual respond/complete time and date of the job. |
| CB_labour | Same as described for CA_labour, but the operatives and priority of the job are also included. |
| CC_ppm | Description of the PPM, PPM period (when it is planned to start and finish) and the cost of service. |
| CC_labour | Concatenated job descriptions, full job descriptions, and work completion dates. |
| CC_con | Service and Facilities Group (SFG) code, condition rating and approximate age of the asset, and whether the asset replaces within five years. |
| CC_cost | Total cost of the corresponding job of all sites under contract C. |

The challenge with this project is that there are no sensors in those older buildings, and the FM system simply records labour data. As a result, we must design a data-driven model with no sensor data and no expert knowledge of the assets. Note that in this project, asset means the facility of a building, such as air conditioning, boiler and lift, and site represent building(s) of an organisation.

**Problem 1:** The PPM/reactive jobs in the labour databases do not identify which building asset was inspected. For example, a building could have five boilers, and an engineer may have visited to repair one of them, but the only information details recorded are the building and the fault. Hence we cannot determine which asset was inspected.

**Result:** Initially, sites with a single asset type are isolated. For instance, site A only has one asset for maintenance. Thus the condition and the PPM/reactive jobs records of the asset can be defined precisely. However, there were only 13 assets that satisfied the idea.

**Problem 2:** The receptionist edits the description of each PPM/reactive job, and occasionally the description could be an update of the asset rather than describing the job that needs to be completed.

**Result:** Therefore, manual data cleaning is required by going through each row of data to categorise the jobs and determine whether it is appropriate to use.

## III. RELATED LITERATURE REVIEW

Although the various statistical methods have been described in section II. A, most are reliant on sensor data. Due to the lack of sensor-based condition monitoring data, the statistical technique based on historical data is necessitated [14].

“The reliability of a product (system) is the probability that the product (system) will perform its intended function for a specified time period when operating under normal (or stated) environmental conditions.” [15]. The Weibull distribution has been widely employed in various fields, and in reliability theory, it is mostly used to characterise degradation data and the various effects of failure expectancies, such as decreasing, constant, and increasing.

In [16], a data-driven prognostics model was constructed by fitting a Weibull distribution using the least square method (LSM) and the maximum likelihood estimator (MLE). The model was developed using a dataset comprising the failure times (cycles) and scrapping quality of the components in the assembly. The model predicts the failure time of the assembly's components and outputs the number of failures during their lifetimes.

Similarly, [17] demonstrated how the Weibull distribution could be used to model the time between failures (MTBF) (multiple failures of a machine are counted) of a machining centre (MC), but with type I censored lifetime data. The data utilised in this paper consisted of a single column of 30 times to failures, and the fitted Weibull model established the basis of calculation of MTBF for MC. The standard Weibull distribution with two parameters is utilised in the preceding two papers.

However, [6] demonstrated that a three-parameter Weibull distribution could also be used to model the distribution of equipment failure times. This paper demonstrates how the model may aid in using PdM in hospital heating and air conditioning facilities by optimising the maintenance plan to extend the asset's RUL and considering associated maintenance costs.

In contrast to the (parametric) Weibull model, the Cox proportional hazard (PH) model is a semiparametric model that does not model the failure time but identifies factors related to the hazard. In 2017, [18] addressed PdM in a non-sensored approach, utilising the Cox PH model to determine which factors influence the risk of wastewater pipe blockage. The dataset is constructed using records of water pipe blockages and contextual data for 44800 vitrified clay wastewater pipes. The location of pipes with varying degrees of blockage risk was visualised on a map in this paper, allowing the asset management to target inspections and schedule maintenance and replacement programmes appropriately. Similarly, [14] demonstrated the use of Bayesian Weibull PH model analysis for the same wastewater pipe dataset. They explored the Bayesian and frequentist perspectives on the PH model. The paper argued that uncertainty measures such as confidence intervals (CI) and p-values could be easily misinterpreted, and they imply that the probability of hazard ratio (HR) has a better representation for non-mathematicians. The point estimate is mostly similar, such that CI and credible intervals are almost identical amongst the two approaches.

Additionally, [18] demonstrated the PH model's use though it did not validate the models. As a result, [19] highlight the lack of validation for Cox PHs models in order to build trust in their ability and outline a way for overcoming it, including prognostic index comparison for training and test sets and prediction calibration via cross-validation. Notably, [19] also mentions that a forest plot of the hazard ratios from the Cox PH model is beneficial, as is a nomogram for presenting a predictive model to non-mathematicians.

## IV. Data Analysis and Cleaning

The most pertinent analysis can be done by combining the literature reviews and modelling EQUANS data by the time between operation and failure (when the asset is not operated) of an asset. In this context, knowing when reactive jobs are expected to occur can help EQUANS make maintenance decisions.

### *A. Asset Determination*

In each contract, EQUANS maintains hundreds of assets, and deciding which to focus on is crucial. It was concluded with EQUANS that the boiler would be an acceptable asset to study because heating and hot water reactive jobs account for a significant amount of the total yearly cost of reactive jobs. In this manner, boiler data can illustrate a potential model which can be applied apply to any asset.

### *B. Data Cleaning*

The first step is to extract the data relating to the boilers of each site. For dataset(s) of each contract, we aim to filter each site's PPM and reactive jobs of boilers. However, several problems were encountered, such as individual boiler cannot be determined in contracts B and C's datasets, and no exact date and time in contract C's dataset indicate when the PPM and reactive job are taking place. Therefore, actions have been taken and assumptions have been made on those data to facilitate the model building:

- CA_con was used to find the list of sites with one boiler, then extracted and manually cleaned the CA_labour dataset based on the list to ensure there was only one boiler on site.
- For contract B, the list of assets was not provided. Therefore the most appropriate thing to do is to extract and manually cleaned the CB_labour dataset based on intuition to ensure there was only one boiler on site.
- The CC_labour dataset contains the duration of the PPM rather than the specific day and time. For example, 01/01/2021 to 01/02/2021. In this scenario, we will suppose the period's first date is PPM. Similarly, no dates for the reactive jobs were supplied, but the work completion dates were. Thus, we assume the finished date as when the reactive job arose.

The second step is to extract the period between the boiler being active and reacting, and the period can work as a time-to-failure or a time-to-event concept. The boiler is presumed to be operational after PPM and repair, and only sites with more than one job will calculate the time difference, then the site with no such time difference is removed. Each site of the labour dataset has had the following extracted:

1) The time (in days) between the PPM and the nearest later reactive job.
2) The time (in days) between the PPM jobs.
3) The time (in days) between reactive jobs.

For example:

Table III
EXAMPLE OF TIME-TO-EVENT CALCULATION

| | Job Type | Raised date | Time-To-Event |
|---|---|---|---|
| Site 1 | PPM | 01/01/2019 | Nan[a] |
| Site 1 | Reactive | 15/01/2019 | 14 |
| Site 1 | Reactive | 25/01/2019 | 10 |
| Site 2 | Reactive | 01/01/2019 | Nan |
| Site 2 | Reactive | 02/02/2019 | 32 |
| Site 2 | Reactive | 14/08/2019 | 193 |
| Site 3 | PPM | 01/01/2019 | Nan |
| Site 3 | PPM | 01/01/2020 | 365 |

[a] Nan is not available.

## V. METHODOLOGY

Overall, we aim to use non-sensored data to build statistical models that predict the time-to-failure and probability of the boiler breaking at a given time point and analyse the impact of variables on operating time. The practical approach is to use survival analysis.

### A. Survival Analysis

Survival analysis is a collection of approaches for analysing time-to-event data. It is also known as reliability theory (in engineering), event history analysis (in social sciences), or duration analysis (in economics). The term 'survival' is equivalent to probabilities in survival analysis, implying that survival can be 'measured' [20].

In this project, each boiler at each site is an 'observation', the boiler's reactive job occurrence is an 'event', and the duration between the boiler's operation and reactive is the 'time-to-event'. In mathematics terms, this can be represented as $i \in \{1, ..., N\}$, where $i$ indicates the $N$ observations and $t_i$ is the time-to-event for the observation $i$.

By utilising survival analysis, we can:

- Estimate survival in a parametric or non-parametric manner.
- Estimate the average survival time of the boilers (average time-to-event) and compare the average amongst different contracts or maintenance strategies.
- Estimate the probability that boilers will fault at a particular time point (survival at a given time).
- Handle censoring data (see the following subsection).

### B. Censoring

Censoring measures whether the event occurs or not, and censoring occurs when the time-to-event is unknown for an included observation. Perhaps observations are observed in a period between time 0 to $t$ and $t_i \in \{0, ..., t\}$, the three censoring types are [21, pp.1-5]:

- Left censoring: The event of interest occurred before $t$, but the precise time of occurrence is unknown, such that $t_i < x$.
- Right censoring: The event of interest occurs after $t$, but the precise time of occurrence is unknown, such that $t_i > x$ or some positive integer $x$.
- Interval censoring: The event of interest occurred between times $t_1$ and $t_2$, such that $t_1 < x < t_2$ for some positive integer $t_1, t_2$ and $t_2 > t_1$.

In this project, the right censoring data is the focus. For example, a site's boiler that never fails is right-censoring since the time-to-event is not observed. Importance of assuming non-informative censorship: censorship does not indicate they are likely to survive/succumb.

### C. Mathematical Intuition

Let $T$ be a non-negative continuous random variable. Assume the random variable has a probability density function (pdf) $f(t)$, and cumulative distribution function (cdf) $F(t)$.

The **Survival Function**:

$$S(t) = Pr(T > t) = 1 - F(t) \tag{1}$$

The survival function $S(t)$ is the proportion of the population with the time-to-event value(s) more than $t$. In other words, it gives the probability of surviving beyond time $t$.

The **Hazard Function (HAZ)**:

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t) | T \ge t}{\Delta t} \tag{2}$$

The hazard function (hazard rate) $h(t)$ is the rate at which the event is occurring out of the surviving population at any given time $t$. In other words, it gives the 'instant probability' of succumbing to the event, given that the individuals have survived to time $t$.

The **Hazard Ratio (HR)**:

$$HR = \frac{h_1(t)}{h_0(t)} \tag{3}$$

The HR is the ratio of HAZ for two subjects with different values of covariate $X$, such that $X \in \{0, 1\}$. For example, if $HR < 1$, it means individuals with $X = 1$ has a lower hazard compared to individuals with $X = 0$.

### D. Kaplan-Meier Methods vs Parametric Method

The two approaches, parametric (based on statistical assumptions) or non-parametric (not relying on statistical assumptions) models, can be used to model survival data, and Kaplan-Meier (KM) is a non-parametric estimator for a survival function [22]. The survival function of KM estimator defined as:

$$\hat{S}(t) = \prod_{i:t_i \le t} \frac{n_i - d_i}{n_i} \tag{4}$$

The KM estimator holds for all t $>$ 0 and depends on two variables, $n_i$, the number of observations in risk at time $t_i$, and $d_i$, which is the number of events at time $t_i$. The non-parametric KM estimator is sometimes used since it does not require prior knowledge of the time-to-event data distribution and can account for censored data. [23]. The Log-Rank test can compare two survival functions (curves) to determine if there is a difference in survival between the two groups. The null hypothesis is no difference in survival $S(t)$ between the two groups, and the alternative hypothesis is that survival $S(t)$ between two groups are different.

The parametric model for $S(t)$ can be calculated by assuming $h(t)$ follows a specific distribution. Three parametric models will be employed to model the boilers' time-to-event: Weibull, Exponential, and Log-Normal. These are the three most common probability distribution models used in predictive modelling nowadays [24]. Although making a parametric assumption of the time-to-event data allows for modelling a survival function that provides more detail and better inference, it also needs to be careful to make assumptions that are justified by the data [20].

### *E. Cox Proportional Hazard Model vs Accelerated Failure Time Model*

Comparatively, the Cox proportional hazard (PH) model and the accelerated failure time (AFT) model include the time-to-event covariates. The Cox PH model is a semiparametric regression model which has the HAZ:

$$h(t, X) = h_0(t)exp(\beta X) \tag{5}$$

Where $h_0(t)$ is the baseline hazard, $\beta$ is the coefficient, and $X$ is the covariate. $\beta$ and $X$ can be a value if there is one covariate, or they could be vectors if there is more than one covariate. Equivalently, if there is no covariate, then $h(t, X) = h_0$ which is defined in Equation (2). The Cox PH regression model does not assume $h_0$, and $h_0$ can be assumed as any functional form, which is left as unspecified. Therefore, it assumes a parametric form for the effect of the covariates on the hazard rate.

The Cox PH model has the assumptions:

- Survival times $t$ are independent.
- The hazard is proportional (HR does not change over time).
- $\ln(HAZ)$ is the linear function of the numerical covariate(s).
- Values of X's do not change over time.

In comparison, the AFT model is a parametric survival model that does not require the PH assumption, therefore overcoming the Cox PH model's violation problem. For example, unlike the Cox PH model, the AFT models account for covariate effects straight from the time-to-event. The survival function is:

$$S(t|X) = S_0[t \exp(\beta X)] \tag{6}$$

Where $S_0$ is the baseline survival function, $\beta$ is the regression coefficient, and $X$ is the covariate. $\beta$ and $X$ can be a value or a vector, depends on the number of covariates.

There are three AFT models are use in the project: Weibull, exponential and log-normal. After fitting the AFT models, Akaike Information Criterion (AIC) can select the best model. It is one of the most widely used criteria for selecting the optimal model. It computes the likelihood of a model and its parameters, with the lowest AIC being preferred, such that:

$$AIC = 2p - 2log(likelihood) \tag{7}$$

where $log(likelihood)$ is the log-likelihood function, and p is the number of parameters estimated, it is straightforward that the model with the largest likelihood with fewer parameters is favoured.

### *F. Preparation for Survival Analysis*

The labour datasets for contracts A and B are based on annual PPM and bi-annual PPM during data cleaning, respectively. We can use contract A as an example of annual service and contract B as an example of bi-annual service to examine if boiler survival rates vary. Since the contracted A and B datasets only contain the times variable, they can model the time-to-event using KM and parametric methods. The contract C's dataset was chosen for the Cox PH and AFT methods as it has certain valid factors that could be beneficial in estimating the boilers' time-to-events, such as age, condition rating and whether the asset will be replaced within five years.

To properly format the datasets and satisfy the assumptions for using survival analysis methods, the first time-to-event values for each site for contracts A and B were extracted to ensure that each failure is independent. However, in section II, we were only left with data from six sites after data cleaning. As a result, the values of the time-to-event are not independent. After all, a new column 'Censored' will be added to indicate censorship, with Censored = 1 for sites having reactive tasks and Censored = 0 for sites with no reactive jobs (such that the event was never observed). We assume the censored data is non-informative in this project, such that boiler data is censored does not necessarily mean the boiler will have a longer operational life than the one that is not censored. Time-to-event data beyond 365 days for annual service data and 182 days for bi-annual service data are also counted as censored.

*1) Bootstrap:* The bootstrap sampling approach seeks to deduce an estimate of a population parameter $\theta$ from sample data by drawing repeated samples from the sample data [25]. It is a non-parametric resampling method that involves sampling independently from a sample of data with size n and inferring from the resampled data. However, in this project, we do not aim to calculate the test statistic. Nevertheless, instead, we increase the size of the datasets. After data cleaning, we have 42 rows of data for contract A, 14 censored, and 28 non-censored. Contract B has 62 rows after data cleaning, with 32 filtered and 30 non-sensored. The datasets for contracts A and B are somewhat small and unbalanced to compare. Thus, both datasets were bootstrapped to size 100. This strategy is used to increase contracts A and B labour datasets to facilitate comparison.

*2) Software used:* Python's library 'lifelines' (version 0.26.0) is utilised to perform the KM method, log-rank test, Schoenfeld's test, modelling the Cox PH model and AFT models and library 'reliability' (version 0.6.0) for parametric methods.

## VI. Results

The objective is to use survival analysis models on the datasets provided by EQUANS to address the questions in section VI. B. The datasets of contract A and B will be used to compute the KM and parametric models, and dataset of contract C will be used to compute the Cox PH and AFT models, the intention of the way the datasets and methods are allocated is explained in section V. F, and the details of the datasets will be presented in the next subsection.

Firstly, KM and parametric models have been computed to compare the probability that the boilers will be faulty between annual and bi-annual services, and the results of KM and parametric models were compared. Furthermore, the boiler attributes were assessed to determine the impact on the probability of survival by using Cox PH and AFT models.

### *A. Description of the Datasets*

Overall, there are three labour datasets used, which include the basic information such as the Site ID, time-to-event, whether the data is censored or not. For example, Tables IV, V and VI show excerpts of the data, and further details are presented in Tables VII and VIII.

Table IV
Labour dataset of contract A (size:100x4)

| Site ID | Job Type | Time-To-Event | Censored |
|---|---|---|---|
| 19 | Planned | 188 | 0 |
| 30 | Reactive | 159 | 1 |
| 13 | Planned | 292 | 0 |
| 29 | Planned | 368 | 0 |
| 23 | Reactive | 3 | 1 |

Table V
Labour dataset of contract B (size:100x4)

| Site ID | Job Type Code | Time-To-Event | Censored |
|---|---|---|---|
| 103 | Reactive | 55 | 1 |
| 219 | Reactive | 12 | 1 |
| 98 | PPM | 186 | 0 |
| 168 | PPM | 188 | 0 |
| 228 | Reactive | 20 | 1 |

Table VI
Labour dataset of contract C (size:126x7)

| Site ID | Job Type | Time To Event | Condition Rating | Approximated Age | Replace within 5 Years? | Censored |
|---|---|---|---|---|---|---|
| 53 | Reactive | 3 | 3-Poor | 5 | 0 | 1 |
| 38 | Reactive | 165 | 2-Good | 1 | 0 | 1 |
| 53 | Reactive | 78 | 3-Poor | 5 | 0 | 1 |
| 60 | Reactive | 2 | 2-Good | 2 | 0 | 1 |
| 193 | Reactive | 60 | 2-Good | 5 | 1 | 1 |

### *B. Results Plan*

The questions addressed in this analysis include:

- Is there a difference in the probability of the boilers between annual service (contract A) and bi-annual service (contract B)?

Table VII
Details of variable of labour dataset of contracts A and B

| Covariate | Type | Values/Categories |
|---|---|---|
| Job Type | Nominal | PPM/Reactive |
| Time-To-Event | Numerical | Times measured in days |
| Censored | Nominal | 0 = censored/1 = non-censored |

Table VIII
Details of variable of labour dataset of contract C

| Covariate | Type | Values/Categories |
|---|---|---|
| Job Type | Nominal | Reactive |
| Time-To-Event | Numerical | Times measured in days |
| Condition Rating | Nominal | 0 = 3-Poor/1 = 2-Good |
| Approximated Age | Numerical | 11, 26, 43, 46 |
| Replace within 5 Years? | Nominal | 0 = No/1 = Yes |
| Censored | Nominal | 0=censored/1=non-censored |

- Can we evaluate models for datasets which have covariates with the time-to-event data (contract C) and models for datasets with no covariates (contracts A and B)?.
- In a high-level view, can we help EQUANS to identify if they can collect data in a better way? Can improvement be made on their preventive service?

### *C. KM Estimators*

According to Figure 1, there is no significant difference between the survival curves, and the confidence intervals overlap. However, after 200 days, the annual service's survival curve becomes lower than the bi-annual one. However, it does not necessarily mean that annual service boilers have a lower survival probability than bi-annual service boilers after 200 days, but because we have set the data for bi-annual service to be censored if it exceeds 182 days.

The comment above was based on the graphical method. To compare the survival curve statistically, we computed a Log-Rank test and the resulting p-value is greater than the 10% significant level. Therefore, we do not have evidence to believe that the survival functions are different for the annual and bi-annual services.

### *D. Parametric Methods*

The six distributions that have been considered fitting the time-to-event data are Weibull and Log-normal (two and three parameters) and Exponential (one and two parameter(s)). The models are fitted with an automatic function in Python's 'reliability' library, which gives a 'best' model by the lowest Log-likelihood values amongst the fitted models. As a result of the automatic function, the Weibull distribution with three parameters appears to be the best fit for the annual and biannual service datasets, as it has the lowest Log-likelihood values. The comparison of the KM curves with the Weibull distribution with three parameters fitted is shown in Figure 2.

For EQUANS, the results can be used to investigate after how many days the probability of an asset will be less than 50%, such that it might help change the time interval for PPM or replace parts to prevent the failure.

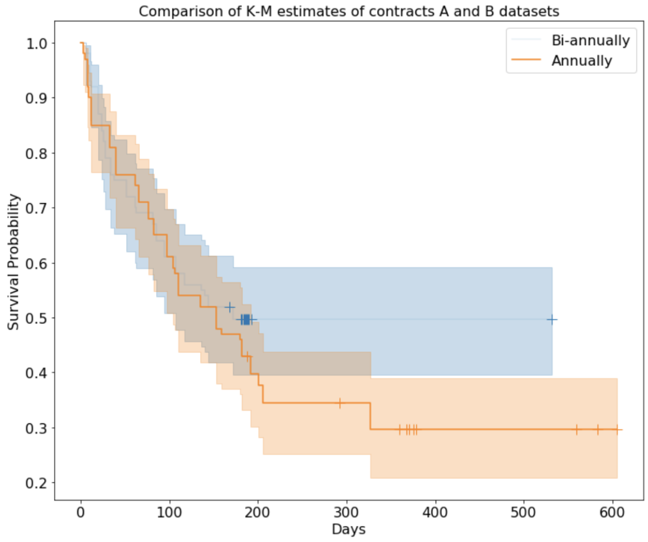
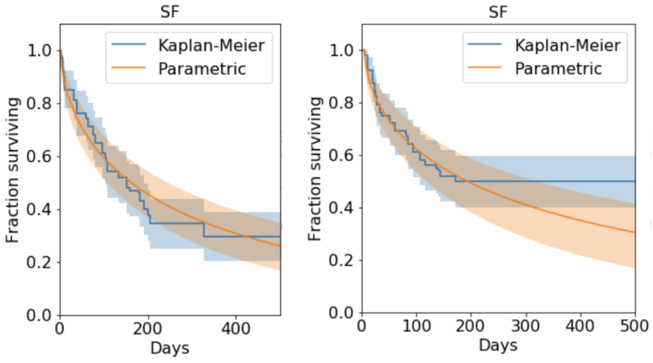


Figure 1. Example of time-to-event calculation



(a)

(b)

Figure 2. (a) Annual service: $S(t)$ of KM estimator and Weibull distribution (3P). (b) Bi-annual service: $S(t)$ of KM estimator and Weibull distribution (3P).

### E. Cox PH Model

In this subsection, contract C's labour dataset is using for fitting Cox PH and AFT models. First, we use Schoenfeld's test to verify the PH assumption. The Schoenfeld residual test is used to evaluate whether the HR varies over time. The null hypothesis is that the HAZs are proportional, and the alternative hypothesis is that the HAZs are not proportional. Python's library has an automatic function, 'lifelines', that compute statistics that check the PH assumption. By running the 'check_assumption' function with the dataset, we passed Schoenfeld's test, such that the PH assumption is valid. Alternatively, the test statistic and the p-values of each covariate can be assessed as well.

Secondly, the Cox PH model is fitted and the estimation results and p-values are shown in Table IX

Table IX
ESTIMATION RESULTS OF COX PH MODEL

| Covariate | **coef** | **HR** | **p-value** |
|---|---|---|---|
| Approx_age | 0.03 | 1.03 | 0.03 |
| Condition_Rating_2_Good | -0.41 | 0.67 | 0.16 |
| Replace_within_5_Years_or_not | -0.40 | 0.67 | 0.07 |

From Table IX, we can see that covariate '*Approx_Age*' is significant at the 5% level, indicating there is some relationship between the age of the asset and survival, but the other two covariates are not significant at the 5% level, which implies they do not impact the survival. The coefficient of '*Approx_Age*' indicates that a (unit) increase in the boiler's age will increase the estimated hazard by 0.03, assuming that all covariates remain the same.

The purpose of fitting a Cox PH model is to measure the impact of covariates on the HAZ, which would help evaluate the covariates' impact. Therefore, for EQUANS, this result can be beneficial for examining the impact of the factors that might affect the survival of the asset. For instance, it may be beneficial to assess the relationship between different values/classes of covariates and the asset's survival probability to determine which factor should be targeted and when assets/parts should be replaced to avoid failure.

### F. AFT Models

Weibull, Exponential and Log-normal AFT models are fitted using the same contract C's labour dataset.

Table X
SUMMARY RESULTS OF FITTED AFT MODELS

| | **AFT Models** | | | | | |
|---|---|---|---|---|---|---|
| | **Weibull** | | **Log-normal** | | **Exponential** | |
| | AIC:1102.56 | | AIC:1098.76 | | AIC:1101.03 | |
| | $\beta$ | **p-values** | $\beta$ | **p-values** | $\beta$ | **p-values** |
| Intercept | 3.15 | <0.005 | 2.63 | <0.005 | 3.17 | <0.005 |
| Approx_Age | -0.05 | <0.005 | -0.06 | <0.005 | -0.05 | <0.005 |
| Condition_Rating_2_Good | 0.45 | 0.07 | 0.35 | 0.2 | 0.46 | 0.05 |
| Replace_within_5_Years_or_not_1 | 0.79 | 0.05 | 1.18 | 0.01 | 0.77 | 0.05 |

The results in Table X have indicated that covariate '*Approx_Age*' is significant at the 5% level through all three AFT models, '*Condition_Rating_2_Good*' is significant at the 10% level of Weibull and Exponential models, and '*Replace_within_5_Years_or_not_1*' is significant at the 10% level through all three AFT models. Where 5% significant level represents a moderate relationship between covariate and survival, and 10% significant level indicates weak relationship. Furthermore, the survival time for the boiler with older age is decreased by a factor of 0.05 for Weibull and Exponential AFT models and 0.06 for the Log-normal AFT model. Moreover, the AIC values of the three AFT models are very similar, and the Log-normal AFT model gives the lowest AIC amongst them.

In comparison, fitting AFT models and picking the 'best' one, the model can be used to measure the impact of covariates on the survival probability. Similar to the Cox PH model, fitting the AFT model can also determine the factor affecting the asset's survival. However, AFT models are based on the intuition of the HAZ distribution, and Cox PH is not. The AFT model is not always realistic, especially the Exponential AFT model. There is no state-of-the-art method to reliably compare semiparametric and parametric models, such that we cannot compare the partial AIC value of the Cox PH model with the AIC values of the AFT models [26]. The purpose is not to compare the Cox PH and AFT models but to evaluate those inferences from the models. However, alternatively, we can assess the fit of the models visually.

## VII. Discussion

Numerous assumptions were used when extracting data from EQUANS's datasets. For example, we assume the boiler is operating when a PPM or reactive job is completed, and we assume the completion time for contract C's labour dataset is when the fault occurs. Those assumptions might lead to bias and uncertainty, but principles aim to show rather than provide accurate predictions. Similarly, we can consider the methodologies employed to paint a picture of the data, which aids in comprehending the time-to-event, but not in drawing any definitive conclusions.

In terms of the investigation of the survival probabilities based on the frequency of the service of the boilers, common sense would expect bi-annual service will have a high probability of survival. However, our results have shown that failure might not be related to the frequency of maintenance, and it can be due to many reasons, particularly closely related to failure mode and other factors, for example, the user (human factors), randomness and accident. The critical point is, the boiler is composed of numerous subcomponents, such as pumps, gas valves and water temperature sensors; hence, different subcomponents will experience different failure rates. Additionally, when the boiler gets PPM, the pumps have not been inspected. As a result, the boiler may fail due to the pump, and the PPM may benefit some components but not others. Although we demonstrated how survival analysis could aid in making maintenance decisions for buildings without sensors in this project, a valid future work task would be to:

1) Use datasets that are as accurate as possible without making too many assumptions.
2) Collect data of various subcomponents of the asset to conduct accurate data analysis and prediction on the RUL or survival of the asset.
3) Analyse data with different covariates that may affect the asset's survival probability.
4) In the future, if the company accessed the associated cost of fitting sensors and knew the maintenance cost, we could recommend the circumstances in which it would be worthwhile to fit sensors.

## VIII. Conclusion

This project was undertaken to investigate the methods to implement predictive maintenance for non-sensored data based on the datasets provided by EQUANS. The aim is to assist EQUANS to transfer from preventive to predictive maintenance by evaluating predictive maintenance in an experimental setting, recommending data collection to gain more critical insights through analysis, and improving preventative maintenance. The relationship between time-to-event (the period of an asset being operational and becoming defective) and maintenance service was demonstrated, also the relationship between time-to-event and asset information.

The KM approach, parametric method, Cox PH model, and AFT model are used to get insight into the probability that an asset may become defective within a specified period. Refering to the results plan in section VI. B , the findings of this experiment reveal that:

- The frequency of maintenance (annual or bi-annual) may not alter the time-to-event.
- By modelling the time-to-event data with covariates, the impact of the factor on survival can be evaluated compared to modelling without covariates. The age of the asset may be a relevant factor related to the survival of the asset.
- We have encountered many problems during the data cleaning and extraction process. With a more precise dataset, such as identifying the individual asset when the job occurs and having a complete, comprehensive failure description, it will be valid to use the techniques for modelling the time-to-event of an asset, and recommending whether or not to install sensors on critical asset.
- The models used in this project can assist EQUANS to predict when an asset will fail and the failure rate trend. Furthermore, it can help decide when the PPM can be taken to the minimum to reduce PPM cost.

In summary, we demonstrate proof of concept, illustrating how data science and statistics may contribute value to the maintenance of old buildings.

## References

[1] J. Cheng, W. Chen, K. Chen, and Q. Wang, "Data-driven predictive maintenance planning framework for MEP components based on BIM and IoT using machine learning algorithms," *Automation in Construction*, vol. 112, 2020.

[2] R. K. Mobley, "Impact of Maintenance," in *An Introduction to Predictive Maintenance (Second Edition)*, second edition ed., ser. Plant Engineering, R. Mobley, Ed. Burlington: Butterworth-Heinemann, 2002, pp. 1–22.

[3] S. Selcuk, "Predictive maintenance, its implementation and latest trends," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 231, no. 9, pp. 1670–1679, 2017.

[4] A. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance, journal=Mechanical Systems and Signal Processing," vol. 20, no. 7, pp. 1483–1510, 2006.

[5] M. Le Nguyen, F. Turgis, P.-E. Fayemi, and A. Bifet, "Challenges of Stream Learning for Predictive Maintenance in the Railway Sector," *Communications in Computer and Information Science*, vol. 1325, pp. 14–29, 2020.

[6] G. Sánchez-Barroso and J. Sanz-Calcedo, "Application of predictive maintenance in hospital heating, ventilation and air conditioning facilities," *Emerging Science Journal*, vol. 3, no. 5, pp. 337–343, 2019.

[7] M. Baptista, S. Sankararaman, I. de Medeiros, J. Nascimento, C., H. Prendinger, and E. Henriques, "Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling," *Computers and Industrial Engineering*, vol. 115, pp. 41–53, 2018.

[8] I. de Pater and M. Mitici, "Predictive maintenance for multi-component systems of repairables with Remaining-Useful-Life prognostics and a limited stock of spare components," *Reliability Engineering and System Safety*, vol. 214, 2021.

[9] E. Ramasso and R. Gouriveau, "Remaining useful life estimation by classification of predictions based on a neuro-fuzzy system and theory of belief functions," *IEEE Transactions on Reliability*, vol. 63, no. 2, pp. 555–566, 2014.

[10] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199 – 231, 2001.

[11] S. Amari, L. McLaughlin, and H. Pham, "Cost-effective condition-based maintenance using Markov decision processes," 2006, pp. 464–469.

[12] A. Abu-Samah, M. Shahzad, E. Zamai, and A. Ben Said, "Failure prediction methodology for improved proactive maintenance using Bayesian approach," vol. 28, no. 21, 2015, pp. 844–851, cited By 14.

[13] J. Hu and P. Chen, "Predictive maintenance of systems subject to hard failure based on proportional hazards model," vol. 196, 2020, cited By 33.

[14] P. Castle, J. Ham, M. Hodkiewicz, and A. Polpo, "Interpretable survival models for predictive maintenance," 2020, pp. 3392–3399, cited By 0.

[15] D. N. P. Murthy, M. Xie, and R. Jiang, *Weibull models*. J. Wiley Hoboken, N.J, 2004. [Online]. Available: http://www.loc.gov/catdir/toc/wiley032/2003053450.html

[16] C. Okoh, R. Roy, and J. Mehnen, "Predictive Maintenance Modelling for Through-Life Engineering Services," vol. 59, 2017, pp. 196–201, cited By 12.

[17] Y. Dai, Y.-f. Zhou, and Y.-z. Jia, "Distribution of time between failures of machining center based on type I censored data," *Reliability Engineering & System Safety - RELIAB ENG SYST SAFETY*, vol. 79, pp. 377–379, 03 2003.

[18] Q. Xie, C. Bharat, R. Nazim Khan, A. Best, and M. Hodkiewicz, "Cox proportional hazards modelling of blockage risk in vitrified clay wastewater pipes," vol. 14, no. 7, pp. 669–675, 2017, cited By 12.

[19] C. Bharat, K. Murray, E. Cripps, and M. Hodkiewicz, "Methods for displaying and calibration of Cox proportional hazards models," vol. 232, no. 1, pp. 105–115, 2018, cited By 4.

[20] F. Emmert-Streib and M. Dehmer, "Introduction to Survival Analysis in Practice," *Machine Learning and Knowledge Extraction*, vol. 1, pp. 1013–1038, 09 2019.

[21] E. Lee and J. Wang, *Statistical Methods for Survival Data Analysis*, ser. Wiley Series in Probability and Statistics. Wiley, 2003. [Online]. Available: https://books.google.co.uk/books?id=kcrcpglZregC

[22] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. [Online]. Available: http://www.jstor.org/stable/2281868

[23] C. Smith, "Analysing censored data using Kaplan-Meier methods; survival analysis," vol. 26, no. 4, pp. 173–174, 2011, cited By 4.

[24] A. Chyad and O. Abudayyeh, "Impact of Environmental Factors on the Condition Rating of Concrete Bridge Decks Using Statistical-Distribution Methods," vol. 26, no. 3, 2021.

[25] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. CRC Press, 1994.

[26] M. Jullum and N. Hjort, "What price semiparametric Cox regression?" *Lifetime Data Analysis*, vol. 25, p. 409, 07 2019.