KU LEUVEN

FACULTY OF SCIENCE

Reliability Analysis of Mechanical Equipment in a Cement Production Plant

Carl ELGIN

Supervisor: Prof. Dr. ir. Liliane PINTELON
Mentor: *James Mutuota Wakiru*

Thesis presented in fulfillment of the requirements for the degree of Master of Science in Statistics

Academic year 2018-2019

Preface

This thesis represents the culmination of a period of tremendous challenge, growth, and learning. Through this research I have had the opportunity to demonstrate the methods I've learned in the classroom and expand upon them in order to solve a real-world problem.

I would like to thank Prof. Dr. ir. Liliane Pintelon for providing the opportunity to conduct this research under her supervision. She has been tremendous resource throughout the process and has provided invaluable criticism, support, and insight.

Additionally, I would like to thank my mentor James Mutuota Wakiru, without whom this research would not have been possible. James have provided more support and encouragement than I could have hoped for during this process. Be it through weekly meetings, Skype meetings, or email, James has consistently made his supportive role in my research a top priority. His domain expertise and analytical mind have truly been a guiding force throughout this research.

Carl Elgin

Abstract

This thesis details the use of a multivariate approach to study the reliability of mechanical equipment in a cement production plant. The cement plant, located in Kenya, contains a large number of mechanical equipment for which stoppage records, production totals, and vibration measurements have been recorded. The primary objective of this research is to build integrated predictive models, incorporating all three data sources, that can be used to predict the reliability and behavior of the equipment. The motivation for this research stems from the cement plant, which has a desire to improve their existing maintenance strategies based on insights about equipment reliability. This thesis will describe the process of building an analysis framework, and demonstrate the use of this framework to generate maintenance decision support.

The methodology of this thesis is tailored to the specific objectives of the research problem, but can be applied to similar problems in future research. The data collection and pre-processing steps are the foundation of the analysis, during which the structure of the data is assembled in preparation for analysis. During the descriptive analysis, preliminary insights are acquired from each of the data sources, and critical elements of the plant are selected as the focus of later models. After critical equipment is identified, the data are integrated to provide a complete record of all stoppage events, production rate, and vibrations observations. The integrated data provide a vast opportunity for reliability modelling, which is explored through the use of non-parametric, semi-parametric, and fully parametric survival analysis techniques. Additionally, several classification models are used to identify the extent to which the integrated data is able to predict future maintenance actions following failure.

The analysis and results are presented with the intent to demonstrate the implications of each model with respect to maintenance decision support. Although each model is estimated based on a specific subset of the plant, the analysis framework can be repeated for any equipment for which data is available.

List of abbreviations

AIC Akaike's Information Criterion

AFT Accelerated failure Time

ANN Artificial Neural Network

ARA Arithmetic Reduction of Age

ARI Arithmetic Reduction of Intensity

CBM Condition Based Maintenance

CDF Cumulative Distribution function

CE Cross-entropy

CM Cement Mill

FN Fan

GLM Generalized Linear Model

HPP Homogeneous Poisson Process

HTML Hypertext Markup Language

IQR Interquartile Range

ISO International Standards Organization

KDP Knowledge Discovery Process

KM Kaplan-Meier Estimate

LOWESS Locally Weighted Scatterplot Smoothing

MLP Multi-layer Perceptron

MTBF Mean Time Before Failure

MTTR Mean Time to Repair

NA Nelson-Aelen Estimate

NHPP Non-Homogeneous Poisson Process

NLP Natural Language Processing

PCA Principal Component analysis

PDF Probability Density Function

PH Proportional Hazards

PHM Proportional Hazards Model

SSE Sum of Squared Error

TRP Trend Renewal Process

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the current manufacturing environment exhibits market and price competitiveness, there is an ever-increasing need to produce quality products at a lower cost to meet the market demands. While addressing these demands, manufacturing plants are faced with significant challenges. One of these challenges is the increased cost of production resulting from high maintenance costs due to frequent and costly failures of equipment. However, throughout the cycle of production and maintenance, manufacturing plants generate and collect large amounts of data that could be leveraged to motivate decisions that add value to their maintenance procedures and operations.

The subject of this research is a plant that is faced with similar equipment availability challenges, yet generates and collects several different types of data. However, the plant is unable to use all the different data sources to motivate comprehensive maintenance decision.

1.1 Plant Background

The demand for the analysis stems from a Portland cement production facility located in Kenya, which for the past three years, has built a collection of equipment stoppage records and vibration readings for several sections of the plant. Portland cement, which is the basis of concrete, is produced through a cyclic closely-controlled process of crushing, mixing, and heating combinations of mined materials. As the production process is both resource intensive and dependent upon material availability, each stage of the production cycle is particularly vulnerable to stoppages. The company under study operates a cement production plant that is organized into 8 sections, each corresponding to a different phase of the production process. Each section is further comprised of physical equipment, each uniquely identifiable.

A high-level overview of the cement plant layout is shown in Figure 1.1, which identifies the 7 plant sections(Section 1, where raw materials are mined from a local quarry, is not part of the manufacturing process and is not included in the research) involved in the cement manufacturing process. The plant is organized into these sections according to the respective manufacturing function, as identified in Table 1.1, with each section containing equipment specialized for the specific function.

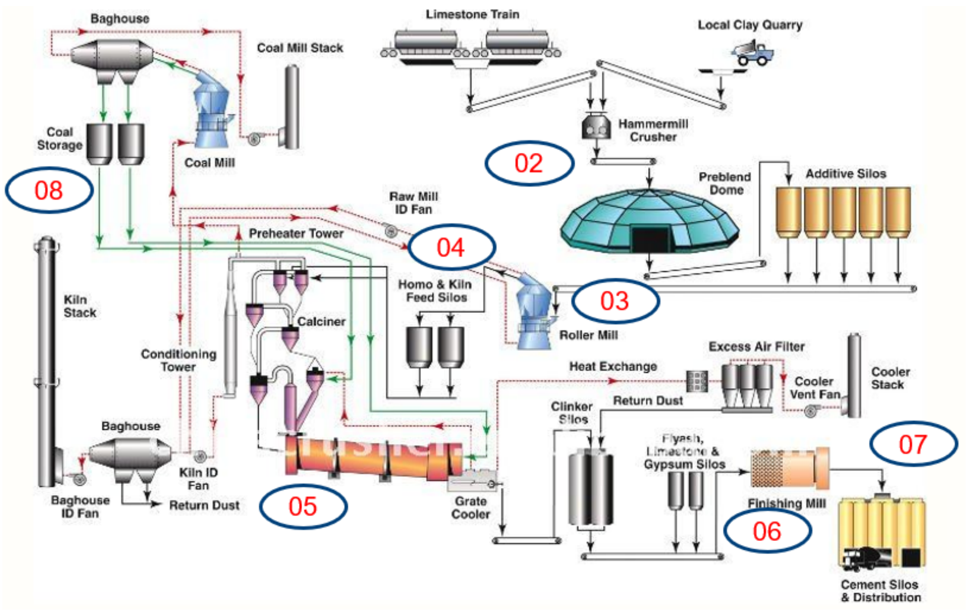

Figure 1.1: High-level overview of cement plant layout

Table 1.1: Cement plant sections and functions

Section	Function	Section	Function
02	Raw material preparation	06	Cement grinding
03	Raw meal grinding	07	Cement dispatch
04	Raw meal homogenizing	08	General plant services
05	Clinker manufacturing		

1.2 Problem environment and statement

In its current operational state, the cement plant faces several challenges which this research will aim to address. The plant keeps records of all equipment stoppage events(both failure and non-failure events) and equipment vibration readings, in addition to tracking monthly production totals. Despite historical records, the data is only presented in weekly or monthly reports for the purpose of calculating overall plant availability. As the stoppage event records are segmented into weekly reports, the historical information is distributed between a large number a files. This structure prevents the stoppage data from being used for analysis to facilitate maintenance decision support.

Despite the use of corrective, condition-based, and preventive maintenance, the plant continues to experience frequent failures of the equipment. Additionally, the plant experiences a considerable number of non-failure related stoppages, such as lack of raw materials, lack of power, fuel, and other process related consumables, which may adversely effect the maintenance and maintainability of the equipment. Unfortunately, the organization lacks a framework to establish how the different stoppage events(failure and non-failure) impact the maintenance of the plant.

Since each data source is generated from a different process, the data is collected independently and at different levels of abstraction, and is analyzed independently using different methods. For example, stoppage events are tracked to a specific equipment, but the broken part or component is not uniquely identifiable. Additionally, production output is tracked at the plant section or sub-section level, but not the equipment level. One equipment may be monitored for vibrations on several different parts, and another equipment on a different set of parts. As the different data sources may be relevant to different levels of the plant, it prevents the data from being readily used when necessary.

Although historical records pertaining to individual equipment items are maintained, the organization only derives availability measured at the overall plant level. Stoppage events are not analyzed to identify critical sections of equipment or to model long-term reliability characteristics. Despite maintaining the historical records, the plant has no decision support framework for using this information to guide maintenance decisions.

Additionally, the historical production data and vibration readings are stored in structures that neither facilitate inference nor integration with other sources. As the data is largely unstructured, the plant is unable to integrate and utilize all the data sets, and requires extensive work to devise and implement a cohesive structure. However, once integrated, it will be possible to perform knowledge extraction on the data while in a combined state. After integrating the data sources and extracting all available information, it can be used to develop predictive models to derive maintenance decision support.

In acknowledgement of these challenges, the organization has a desire to improve and optimize their maintenance programs through the use of data-driven decision support. Based on these challenges, we derive our research objectives as described in the next section.

1.3 Objectives

The primary objective of this thesis is to develop an integrated predictive model, incorporating failure event records, production output records, and vibration observations, that can be used to predict the reliability and behavior of the mechanical equipment of the plant. In order to achieve the primary objective, the research must accomplish several specific objectives.

The first specific objective is to undertake data pre-processing to prepare raw data for the current analysis, as well as future analysis. This will involve aggregating the data from each source into a single repository, cleaning the repository and removing non-informative formatting, transforming the repositories into functional tables, and standardizing the data according to relevant standards. This objective includes repeated consultations with domain experts at the cement facility to obtain clarifications regarding data structures and relevant terminology.

The second specific objective is to perform a descriptive analysis to identify important characteristics regarding the scope of the data collected. Additionally, a criticality analysis will be performed to identify key sections and equipment items within the plant on which to focus for subsequent analysis.

After identifying several critical elements of the plant, the third specific objective is to build reliability models that can be used by a maintenance engineer for maintenance decision support. Such a model would identify specific reliability characteristics of a given equipment, demonstrating a methodology that can be applied to any area of the plant in future work.

The final specific objective involves an integration of the failure event records, the production output records, and the condition monitoring(vibration measurements) records into a predictive model. The integrated data will represent the maximally available knowledge regarding the condition and behavior of an equipment with which to predict future events, and prescribe future maintenance.

1.4 Scope

As mentioned previously, the cement plant maintains records detailing all stoppage events occurring within the plant, a record of monthly production totals according to plant subsection, and a record of vibration readings measured via a handheld probe.

The equipment stoppages are recorded on a daily basis (for weekly reporting), with each stoppage linked to a uniquely identifiable equipment code. The stop time, start time, and total duration are recorded for each stoppage event, along with a brief root cause description. Additionally, each stoppage is further categorized by a code indicating the type of stoppage that occurred(e.g., planned maintenance, mechanical failure, etc.). Furthermore, each entry contains a free-form field in which the maintenance engineer can include a comment explaining what may have occurred and what action was taken to resolve the stoppage. As the stoppage events have been recorded over a three year period(2015, 2016, and 2017) the plant has experienced a total of 30,380 stoppage events.

In addition, production figures are recorded in the form of total number of tons of material processed per month. For sections(e.g., cement grinding) which are made of up multiple sub-section running in parallel, the production figures are recorded for each sub-section. For comparison, the plant has also provided the maximum production capacity rate(in tons per hour) for each of the respective sub-sections.

Finally, the plant also maintains a record of vibration measurements which are taken directly at the equipment level with a handheld proprietary probe. Vibrations are not monitored continuously, and are done so at irregular intervals, as deemed important by the plant, and not all equipment items are monitored. In total, there are 1,634 available vibration observations taken throughout the same three year period.

1.5 Research direction and structure

This thesis will detail the research of a multidimensional reliability analysis of mechanical equipment being used in a cement production facility. The methodology and subsequent results will focus on providing decision support for maintenance actions performed on the equipment, and to characterize the effect of these actions on reliability. This paper will detail each step of the research process, from cleaning and transforming the observed data, to comparing statistical models, and interpreting results.

Following the introduction of the objectives and scope of the research, the second chapter will provide a summary of academic works and techniques that may provide background information, justification, or context for the following research. Next, a chapter describing the methodology of the research will detail all of the steps to be taken during the process. After applying the methodology, the resulting analysis will be described, with results and insights explained as they are uncovered by the research. Following the results chapter, a conclusion chapter will summarize the results of the analysis with respect to the research objectives, providing a cohesive assessment of the knowledge that has been gained. Finally, a post-analysis discussion will provide a critical assessment, from the perspective of the researcher, regarding points of improvement for the methodology or research subject, in addition to recommendations for future work.

Chapter 2

Literature Review

In order to address the objectives of the thesis, adequate knowledge in the domains of maintenance engineering, research, knowledge discovery, and statistical modeling will be required. This literature review will introduce and summarize some of the academic works regarding these domains in order to provide a solid foundation for analysis. Figure 2.1 provides a brief overview of some of the topics to be covered in order to achieve the research objective. The following sections of this chapter will give a synthesis of popular maintenance strategies, as discussed in the literature, and later, an overview of the statistical methods used with these strategies.

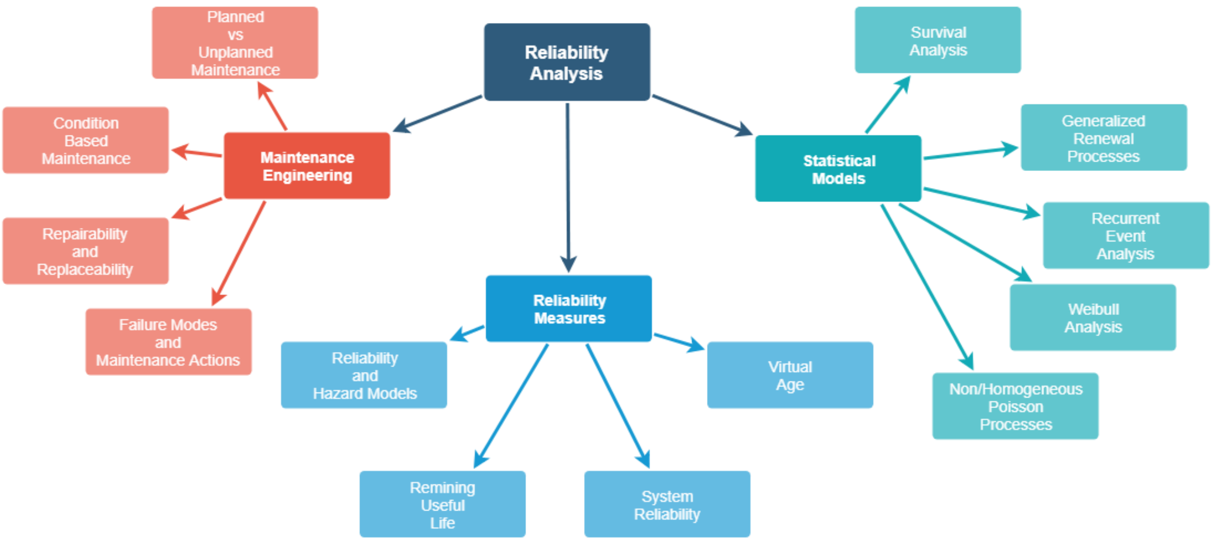

Figure 2.1: Overview of literature review topics

2.1 Maintenance strategies and related topics

As the impact of maintenance on mechanical equipment has been studied for many decades, there is myriad literature available, detailing **corrective maintenance**, **preventive maintenance**, **reliability centered maintenance**, and other strategies(Muchiri et al. (2013), Jardine and Tsang (2013)). The choice of maintenance strategy can vary between different equipment or entire systems, depending upon what is known about the life cycle, reliability, and performance of the unit. Pintelon and Parodi-Herz (2008) describe the evolution of maintenance engineering and the implications on *Maintenance Actions*(the task performed), *Maintenance Policies*(the mechanism which triggers the action), and *Maintenance Concepts*(the logic and decision structure).

2.1.1 Unplanned/Corrective Maintenance

Unplanned maintenance, also known as **corrective maintenance**, refers to a strategy such that maintenance only occurs in reaction to an event, such as a failure or stoppage. In such instances where maintenance is performed on an as-needed basis only, equipment is used until it fails, with a large degree of uncertainty regarding the expected lifetime of the equipment.

Additionally, corrective maintenance often begins with an investigation period, at the cost of the operating organization, where the cause of the failure must be diagnosed, and an appropriate solution identified. It follows that such unanticipated failures yield additional risk in that the corrective decision (whether to repair, replace, etc), despite being appropriate based on available knowledge and resources, may not be the optimal decision.

2.1.2 Planned/Preventive Maintenance

On the other hand, planned maintenance(PM) refers to a strategy in which maintenance actions are scheduled based on known characteristics of the equipment. In planned maintenance, value is placed upon recording historical event data related to performance and stoppages. This information can provide immediate insight into the **operational equipment effectiveness**, and can identify equipment whose operational cost has increased or whose performance has decreased. Furthermore, modelling equipment lifetimes and failure occurrences, increases the ability to plan maintenance such that **remaining useful life** of equipment is not wasted(Nystad and Rasmussen 2010). Most importantly, historical event data allows maintenance engineers to model equipment lifetimes, cost, and performance, to ensure that maintenance actions are taken at optimal times, as opposed to waiting until failure.

2.1.3 Condition-Based Maintenance

As detailed in Wakiru et al. (2019), **condition-based maintenance**(CBM) is a maintenance strategy that uses the information obtained while monitoring the condition of a physical asset to recommend maintenance actions for it(Wakiru et al. 2019). Using a condition-based maintenance strategy removes significant uncertainty regarding maintenance actions, as the timing and type of action is driven by the observed condition of the equipment itself. Wakiru et al. (2019) also provides an extensive review of condition-based maintenance practices in the context of a lubrication condition monitoring strategy, whereby the physical and chemical properties of lubricant can be used to assess mechanical faults of wear(Wakiru et al. 2019).

There is also extensive literature regarding several other strategies of condition monitoring such as vibration analysis. Yang et al. (2016) details a case study in which the frequency of vibrations is used to assess and model the condition of bearings in wind turbines. Additionally, Barszcz (2019) provides an extensive guide to condition monitoring of wind turbines using vibration analysis, detailing measurement collection, signal processing, and analysis methods. Vibration analysis data can include velocity, displacement, or the general condition of an equipment such as a bearing. Furthermore, a review of several other condition monitoring techniques including acoustic emission analysis, machine current signature analysis, and supervisory control and data acquisition is written by Salameh et al. (2018). Other condition monitoring techniques commonly applied in the industry include, lubricant condition monitoring, infrared thermography, acoustic emission and ultrasound as also corroborated by Wakiru et al. (2019).

CBM programs are beneficial in both prognostic and diagnostic capacities, by identifying opportunities to prevent equipment failures, or identifying the cause of faults when prognostics fails(Jardine, Lin, and Banjevic 2006). Although prognostics is efficient for achieving zero-downtime maintenance, CBM programs require both condition indicators and event data. The combination of both data sources are used to build a statistical model that explains the causal mechanism of failure events. Furthermore, future event data can be used to assess the quality of the model, and in turn, the condition indicators themselves(Jardine, Lin, and Banjevic 2006).

2.1.4 Other related topics

When considering historical records in order to plan maintenance activity, the study of all events contributing to downtime or in some way affecting the operation of an equipment is referred to as **event analysis**. In addition to the classification of different maintenance strategies, there is considerable literature regarding the classification of the stoppage events that mechanical equipment experience. Rausand and Høyland (2004) provides an extensive overview of strategies for failure event classification, depending upon the nature of the mechanical system. In addition to providing background into the purpose of studying failure events, the text makes a distinction between *failures*, *faults*, and *errors*, which are sometimes used interchangeably when describing mechanical equipment.

Furthermore, Rausand and Høyland (2004) details a classification scheme for mechanical failure events as either *intermittent failures* or *extended failures*, with extended failures having a further sub-classification. In the literature, an intermittent failure refers to failures that cause a temporary lack of *function* of a mechanical equipment, which reverts to full operational condition after the failure. Conversely, extended failures cause a lack of function that persists until a *functional block* of the equipment or system receives a maintenance action.

When the number of different equipment in a production plant is very high, addressing challenges such as reducing downtime or the number of failure events may be difficult without the ability to compare downtime or failure occurrence between equipment. Identifying specific equipment that is responsible for the largest contribution to the challenges to be overcome is called a **criticality analysis**. The concept of a criticality analysis represents the process of selecting an aspect of focus(e.g., equipment or system) based on specific behavioral attributes. In the context of mechanical equipment and maintenance practices, this behavior may represent equipment availability, operational cost, failure rate, or any other attribute prioritized by the analyst. In Jardine and Tsang (2013), numerous methods of prioritization are detailed, such as the usage of Pareto charts, jackknife diagrams, and trend plots.

A Pareto chart is used to prioritize individual equipment or systems based on specific attribute. An example of a Pareto chart used in a case study to identify optimal change-out times for major components of mobile equipment, as detailed in Jardine and Tsang (2013). In this particular example, the Pareto chart is used to differentiate between a large number of equipment items, according to the total downtime of each component. The approach also details that Pareto charts, and similar plots, are also useful for prioritizing equipment based on other important aspects.

The jackknife diagram is an additional method that typically involves differentiating between critical equipment by comparing both the failure frequency and mean downtime. Such a plot is supplementary to the Pareto chart, which does not identify whether the critical equipment has experienced multiple failures, or only one. Additionally, the same case study detailed in Jardine and Tsang (2013) demonstrates further usage of the jackknife plot by classifying failure profiles for equipment as either acute, chronic, or both.

Furthermore, Jardine and Tsang (2013) also mention the capacity to identify critical equipment or systems based on the trend in failure rate. In this case, the failure events of an equipment are profiled such that the engineer can determine whether failures are occurring with a positive trend(i.e., increasing failure rate), no trend(i.e., failures are independent and identically distributed), or negative trend(i.e., failures are occurring less frequently). Identifying positive trends in failure rate may be particularly useful for prioritizing components which exhibit an accumulation of wear, or deterioration, over time.

Additionally, the **mean time to repair**(MTTR) or **mean time before failure**(MTBF) are reliability measures that may be considered when identifying critical equipment. As a maintenance engineer seeks to reduce equipment downtime, the MTTR indicates the expected downtime for a given equipment following failures. If the MTTR gets larger, it indicates an equipment for which failures result in longer downtimes. Similarly, the MTBF represents the expected duration of time for which the equipment will be operational before experiencing a failure event. If the MTBF changes over time, it represents an equipment that is experiencing more or less frequent failures.

Both MTTR and MTBF are reliability measures used to better monitor the lifespan of mechanical equipment. When the maintenance is performed timely (i.e. inspections, repairs, etc.), the lifespan and reliability of systems can be significantly improved(Ruijters et al. 2016).

Reliability is the probability that a component/equipment will perform its intended function, without failure, under specific condition, for a specific period of time (Choudhary, Tripathi, and Shankar 2019). For a given exponential time before failure distribution, the reliability is computed by:

$$R(t) = e^{-\lambda t} \tag{2.1}$$

For the given Weibull time before failure distribution, the reliability is computed by:

$$R(t) = e^{-\left(\frac{t}{k}\right)^{\alpha}} \tag{2.2}$$

From the above equations, we can compute reliability of a subsystem for given distribution and parameters at any point of time.

While undertaking maintenance of mechanical equipment using the aforementioned maintenance strategies, divergent maintenance data is collected describing the equipment failures, its condition, actions taken, spares used and other aspects. Accumulation of such historical data embeds important information about the equipment and may be extracted using a knowledge discovery process as discussed in the next section.

2.2 Knowledge Discovery Process

A Knowledge Discovery Process(KDP) represents an iterative framework for identifying characteristics or patterns in data and understanding how to apply them with domain knowledge. Swiniarski, Pedrycz, and Kurgan (2007) provides a wealth of information regarding KDPs, including supporting arguments for structuring a KDP as a standardized process model. In summary, the text advocates for a KDP that is ultimately useful to the user, is logical in approach and structure, follows established domain principles, and fosters standardization in data and procedures.

In addition to providing examples of KDPs suited for research or industry, Swiniarski, Pedrycz, and Kurgan (2007) provides an example of a hybrid model that can be applicable in a broad range of domains. The 6 steps for this KDP are:

1. *Understanding of the problem domain.* This first step includes familiarization with the problem as well as the domain experts, terminology, standards, and restrictions.
2. *Understanding of the data.* The second step encompasses all aspects of data collection based on the domain understanding established in step one.
3. *Preparation of the data.* Preparation is arguably the most intensive step of the KDP as it provides a strong foundation for a successful and thorough analysis. This step involves, cleaning and formatting the data, in addition to corrections for noise or missing values. This step may include further methods such as dimensionality reduction, feature selection, or summarization in order to satisfy the input requirements of the problem,
4. *Data mining.* The data mining step involves using functions relevant to the problem or domain to extract knowledge or insights from the cleaned and prepared data.
5. *Evaluation of the discoverable knowledge.* The fifth step involves a thorough evaluation of the extracted information and an assessment of its value and contribution to the analysis. This step may include additional consultations with domain experts in order to assess the validity or novelty of the information.
6. *Use of the discoverable knowledge.* The last step includes a detailed plan of how the extracted information is to be put to use in the current domain.

A similar process is introduced as the *data processing pipeline* by Aggarwal (2015), as represented in Figure 2.2. Although this approach is structured slightly differently from the hybrid KDP, it also places strong emphasis on a systematic approach to solving knowledge-intensive problems.

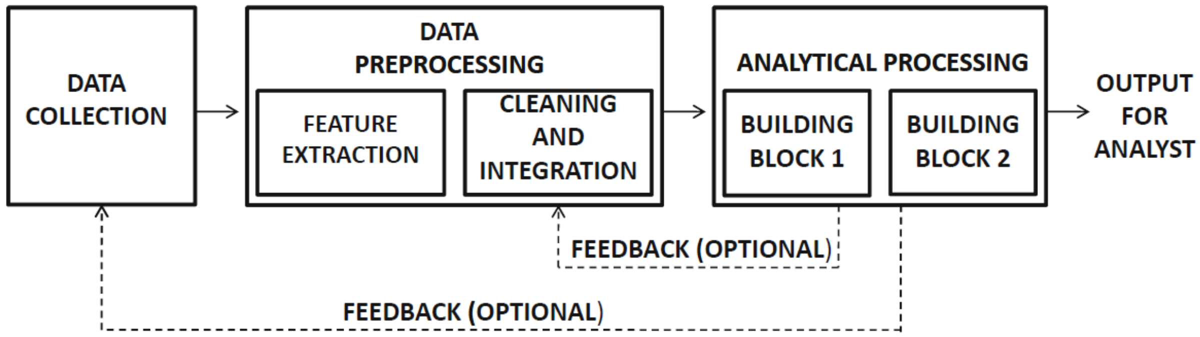

Figure 2.2: The data processing pipeline from Aggarwal(2015)

In addition to their inclusions in structured KDPs, most of the topics are frequently studied and expounded upon in their own capacity.

2.2.1 Data Collection

Although data collection is mostly driven or governed by the demands of the problem at hand, there are many relevant aspects that are common in all use cases. As detailed in Swiniarski, Pedrycz, and Kurgan (2007), an individual responsible for data collection must identify the types, techniques, amount, and quality of the data necessary to solve the problem. During this process, domain knowledge is key for understanding the requirements of the problem and identifying attributes of the information necessary for an insightful solution.

2.2.2 Data preparation

There exists a vast amount of literature regarding the concepts of data processing or data cleaning, as they are most often required for analysis of real-world data. Aggarwal (2015) details key topics of data preparation such as data cleaning, data structuring, and also integration of data coming from multiple sources. As many real-world data sources may contain errors, formatting, or other aspects that prevent it from being ready for analysis, data cleaning is often required as an initial step. Additionally, data may need additional structuring, including character string manipulations or variable transformations in order to satisfy the requirements of the analysis. The desired data may also be distributed between multiple data structures, requiring the researcher to integrate, or merge, the sources my a common reference value.

Swiniarski, Pedrycz, and Kurgan (2007) also provide an introduction into several popular methods for imputation of missing values when removing incomplete observations would not be a favorable solution. Additionally, Josse, Tierney, and Vialaneix (n.d.) provides an extensive repository detailing the available packages in R for exploring and resolving missing data of myriad forms. The repository provides guidance for identifying the scope of the missing data and choosing an appropriate imputation method based on the level of complexity and requirements of the problem.

2.2.3 Text mining

As mentioned in the KDP, data or text mining is the process of extracting additional information from clean and processed data that may otherwise not be available for analysis. The information may be domain-specific or in some way provide additional explanatory power during analysis.

As it has become increasingly popular, there is extensive literature detailing the concepts and methods used for data and text mining. An introduction to text mining is provided by Swiniarski, Pedrycz, and Kurgan (2007), detailing the ability of text mining methods to extract a large number of descriptive features from **semi-structured**(e.g., data table) or **unstructured**(e.g., free-form text) text. The concept of an *Information Retrieval System* is introduced as a system of methods for extracting and characterizing information about a subject.

Robinson and Silge (2017) provides an intuitive guide to text mining in R using the *tidytext* package and demonstrate capabilities for analysis of data in the *tidy* format. This text covers topics such as **tokenization**, the concept of breaking raw text into individual *tokens*, or terms, following the *tidy* structure, to allow for compatibility with other *tidy* structures and functions. In text mining, a dataframe of token is typically called a *corpus* and represents the core data source of intended analysis. In text mining practice, functions to remove *stop-words*, frequent yet irrelevant terms, are commonly used to greatly reduce the magnitude of text to process. Additionally, the text contains examples and demonstrations of text mining tools such as **frequency analysis**, **sentiment analysis**, **correlations analysis**, and **n-gram analysis**.

Frequency analysis represents a high-level summary of a corpus, and refers to parsing a corpus and identifying the frequency with which each token, or terms, occurs. Similarly, sentiment analysis refers to a Natural Language Processing(NLP) method involving parsing a corpus and comparing each token to a lexicon, or a large index of colloquial vocabulary, and identifying the *sentiment* of the token. In such methods, the sentiment typically represents the scale of emotion, positive or negative, associated with the term, based on common usage. Combining frequency analysis and sentiment analysis provides methods for identifying underlying sentiments of entire articles or books.

Furthermore, correlation analysis identifies the correlation coefficient between respective pairs of terms occurring together in the same comment. Although a high correlation for a pair of words does not imply that the words occur very frequently, it does imply that words occur together, or not at all. The correlation coefficient is defined in Equation 4.1 using the components defined in Table 2.1.

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}} \tag{2.3}$$

Table 2.1: Pairwise frequency combinations used to calculate correlation

	Has word 2	No word 2	Total
Has word 1	n_{11}	n_{10}	$n_{1.}$
No word 1	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Additionally, n-grams refer to consecutive sequences of n words that frequently occur together in a corpus. This methods combines elements of both frequency and correlation analysis, as it identifies recurring structures within the corpus. The result of n-gram analysis is a network of term structures that represent the relationship between the most frequent terms in a corpus. The n-gram provides a high-level summary of the text that is best understood with visual representation.

Another NLP method for text mining is **part-of-speech tagging**, whereby each token is processed by an NLP model to identify the root of the word and identify the respective part-of-speech. The *UDPipe* R package performs several NLP functions using including part-of-speech tagging, and allows for the usage of pre-trained or customized models.

Fridolin (2019) also maintain a detailed repository of R packages intended for use in the natural language processing of both *tidy* and *non-tidy* data structures.

2.2.4 Data standardization

In addition to the practices of numerical standardization, there are many resources regarding standardization in terms of the structure and format of data sources, such as those offered by the International Organization for Standardization(ISO). Such standards include ISO 14224:2016 (Collection and exchange of reliability and maintenance data for equipment)and ISO 13306:2010 (Maintenance-Maintenance terminology) in the maintenance field.

Although intended for use in petrochemical industry, ("ISO 14224:2016 (Collection and Exchange of Reliability and Maintenance Data for Equipment)" 2016) provides a detailed framework reliability and maintenance(RM) data that is generally applicable to any maintenance-intensive industry. Specifically, the text provides a guide for the process and methods involved in data collection, placing a strong emphasis on the quality of data. In addition to outlining the taxonomy and subdivision of data collection regarding safety, reliability, maintenance, and business processes, the document also provides detailed recommendations for the structure and format of stoppage event information, such as *Failure Mechanism* and *Maintenance Action* for different types of equipment.

("ISO 13373-1:2002 (Vibration Condition Monitoring - Part 1:General Procedures)" 2002) provides an introductory overview of suggested standard procedures regarding the use of vibration measurements for the purpose of condition monitoring. This text provides a summary of the concept of condition monitoring, along with recommendations regarding data collection, types of measurements, transducer types, as well as data analysis. ("ISO 13373-2:2016 (Vibration Condition Monitoring - Part 2:Processing, Analysis and Presentation of Vibration Data)" 2016) provides a a follow-up to the first document, delving deeper into the specific methods available for analysis of vibration measurements.

2.3 Statistical Methods for Analyzing Event Data

Following preparation of the data to ensure it is ready for analysis, one of the popular techniques for analyzing event data is survival analysis or reliability analysis. The terms survival analysis or reliability analysis, refer to the study of event occurrence along a time scale. The events can be single or recurring events, with the subject of interest being the rate of occurrence, event count, or time-to-event measure. In terms of survival, this event might represent the onset of disease, and for reliability, the event may be equipment failure, threshold of degradation, or stoppage. In other words, the reliability of a component represents the probability of a component surviving(not failing) at least until a specific point in time. For the purpose of this analysis, the terms *survival* and *reliability* will be used interchangeably, as they are equivalent. In any case, reliability analysis involves probabilistically modelling the durations of observed events, in order to predict the time until a future event occurs.

There is myriad literature available regarding the statistical methods available for the analysis of event data. In providing an overview of such methods, Lawless (2007) makes a primary distinction between methods aimed at modelling **counting processes** and those for modelling **gap times**. In the text, a counting process, defined as $N(s,t)$, representing the cumulative number of events occurring during the time interval $(s,t]$. Furthermore, counting processes are most often the outcome of interest when the underlying event process is such that the recurring events do not effectively change the event process itself. In such situations, the recurring events are not marked by an associated intervention, which would change the process itself. An example of a counting process in which interventions are not required following events is the count or rate of occurrence of epileptic seizures.

Gap times are defined as $W_j = T_j - T_{j-1}$, where W_j represents the time between the $(j-1)$st and jth event. Conversely, gap times are typically the outcome of interest, when the recurrence of an event is relatively rare, and is marked by an intervention, having an effect on the underlying process. Although the models used for both event counts and gap times are very similar, the distinction between methods is typically motivated by the objectives of the analysis and the characteristics of the underlying event process(Lawless 2007).

Additionally, Lawless (2007) mentions that regardless of the type of event process under study, two features of the process are typical of interest, namely **time trends** and **event clustering**. In the text, a time trend is defined as a systematic change to the event process that occurs over time. In the case of mechanical equipment, time trends may manifest in the form of an increase in the intensity, or rate, or failure occurrence, or in the duration of gap times between failures. Such a time trend may be representative of a change in the behavior of an equipment item as a result of an accumulation of wear, or in response to a change in maintenance policies or the effectiveness of maintenance actions.

On the other hand, Lawless (2007) defines event clustering simply as "the tendency for events to cluster together in time." Similar to the notion of a time trend, event clustering represents a potential change in the underlying event mechanism in response to the occurrence of other events in nearby time. When studying the recurrence of events of multiple types, event clustering considers the possibility that the close proximity or frequency of one type of event may influence the occurrence of another type.

An additional consideration in terms of the methods employed for analyzing a process of recurrent events is the type of **covariates** available for study. As in Lawless (2007), covariates are most commonly classified as *fixed* or *time-varying* covariates based on the distinct relationship between the covariate value and time. In other words, *fixed* covariates are also referred to as time-independent covariates, as their value does not change during the event process. Examples of time-independent covariates are birth year, identification number, or treatment group.

On the other hand, *time-varying* covariates are also referred to as time-dependent variables, as their values typically change throughout the event process. Examples of time-dependent covariates include age, weight, or presence of infection. An additional distinction is frequently made between *internal* and *external* time-varying covariates, or between *internal* and *ancillary* time-varying covariates, which are equivalent distinctions. Time-dependent variables are typically "internal" variables, which represent values corresponding to the intrinsic properties of the focus of study, or observational unit. Internal variables are typically the product of a stochastic process. Conversely, *ancillary* or *external* variables are those that change value as a result of an external influence which may effect more than a single observational unit(Kleinbaum, Klein, and Samet 2006). As Lawless (2007) describes, external variables are typically determined independently of the underlying stochastic event process, although they may still be time-dependent.

Cox and Oakes (1998) introduces an additional category of time-dependent covariates called *evolutionary covariates*, which depend only on the history, $\mathscr{H}_t$, of the event process. In the example, the history of the event process refers to "the history of failures, censoring and of any other random features of the problem all up to time t(Cox and Oakes 1998)."

2.3.1 Models for Event Counts

As summarized in Lawless (2007), event counts are often represented by a **Poisson process**, "which describes situation where events occur randomly in such a way that the numbers of events in non-overlapping time intervals are statistically independent." Additionally, the text also mentions that Poisson processes are often used to model events that are considered "incidental", or caused by random external factors. In the context of modelling event counts or occurrence rate, the Poisson process is defined by the intensity function:

$$\lambda_i \left(t|H_i(t)\right) = \lim_{\Delta t \downarrow 0} \frac{\Pr\left\{\Delta N_i(t) = 1\right\}}{\Delta t} = \rho(t), \quad t > 0 \tag{2.4}$$

where $\rho(t)$ is a non-negative integrable function. When the function $\rho(t)$ is constant, the process is called a **Homogeneous Poisson Process(HPP)**, otherwise it is a **Non-Homogeneous Poisson Process(NHPP)**. Poisson process models may be non-parametric, semi-parametric, or fully parametric and include covariates.

Both the homogeneous Poisson and renewal process models are based on the assumption that the time between events are independent and identically distributed. In the context of reliability, these assumptions are indicative of equipment that is entirely replaced following failure. In this regard, both models are considered perfect repair models, such that when operation resumes, the equipment is “as good as new”(Wu and Scarf 2017, @lindqvist2006). The homogeneous Poisson process typically models reliability in terms of the MTBF, which is represented by a Poisson distribution. The more general renewal process, is equivalent to the HPP when specifying a Poisson distribution, but can also incorporate other distributions, such as the Weibull(Yanez, Joglar, and Modarres 2002).

In contrast to the HPP, the non-homogeneous Poisson process is suited to describe “as bad as old” restoration following failures. The NHPP follows from the HPP model, but includes an intensity function that varies with time(Wu and Scarf 2017). In this aspect, NHPP models have the ability to describe long term trends in reliability, such as “wearing in”(growth) or “wearing out”(degradation)(Tanwar, Rai, and Bolia 2014). Additionally, as shown by Coetzee (1997), NHPP models are capable of describing the reliability of repairable systems(Hartler (1989), Coetzee (1997)).

2.3.2 Models for Gap Times

In contrast to studying the count or rate of event occurrence, *gap times* are often of interest when events are rare enough that the occurrence of a single one. Although sometimes denoted separately, the methodology for survival times and gap times is equivalent.

The **survival function** is commonly denoted as:

$$S(t) = 1 - F(t) = \Pr\{T > t\}, \quad 0 \leq t < \infty \tag{2.5}$$

The function $F(t)$ is the **Cumulative Distribution Function(CDF)** which represents the probability that the observed duration T will be less than or equal to T. With $f(x)$ representing the **Probability Density Function(PDF)**, the CDF is denoted as:

$$F(t) = \Pr\{T \leq t\} = \int_0^t f(x)dx, \quad 0 < t < \infty \tag{2.6}$$

Given a sample of observed failure or survival times the *Empirical survivor function* can be derived as(Collett 2003):

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times } \geqslant t}{\text{Number of individuals in the data set}} \tag{2.7}$$

Another facet of reliability analysis in the concept of the hazard function, which represents the instantaneous hazard rate. This hazard rate is interpreted as the probability of failure by a future time point, given that the equipment has been operational until the present time(Zacks 2012). The Hazard function is denoted as:

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\delta \to 0} \frac{pr(t < T < t + \delta | T > t)}{\delta} \tag{2.8}$$

Non-parametric Estimation

The **Kaplan-Meier**(KM) estimator provides a non-parametric estimate for $\hat{S}(t)$, where n_j is the number of observations still alive at time t_j and d_j is the number of deaths at t_j, is defined as:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{2.9}$$

Collett (2003) provides a description of the **Nelson-Aelen**(NA) estimator, which provides an alternate estimate of the survival function S(t). Although the NA estimate may perform better than the KM estimate for small samples, the KM estimator can be considered an approximation of the NA estimate, especially at short survival times. The NA estimate for $\tilde{S}(t)$, where n_j is the number of observations still alive at time t_j and d_j is the number of deaths at t_j, is defined as:

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp\left(-d_j/n_j\right) \tag{2.10}$$

Semi-parametric Estimation

The **Cox Proportional Hazards Model**(PH), introduced by Sir David Cox, designates a model in which the hazard function, $h(t)$, is a product of a *baseline hazard* function $h_0(t)$ and an exponent term, $\exp\left(z'\beta\right)$. The baseline hazard function only depends upon time t, and represents the hazard when all covariate values are zero. The second term in the product, $\exp\left(z'\beta\right)$, only depends upon the value of the covariates z', and does not depend upon time. As such, the main characteristic of of PH models is that for any given time t, a change in covariate values represents a proportional change of the respective hazard function. As summarized by Lengerich (n.d.), "his assumption[PH] means if a covariate doubles the risk[hazard] of the event on day one, it also doubles the risk of the event on any other day."

Proportional hazards models are very commonly used to solve regression problems in survival analysis, but is also applicable in engineering reliability. The PH model is a technique used to quantify the effect of covariates(environmental factors, maintenance actions, etc.) on a baseline hazard function(the instantaneous failure rate)(Moore 2016). In contrast to other methods, the Cox proportional hazards model is a semi-parametric model in that although the regression parameters are estimated, the baseline hazard function is never specified and remains unknown(Kleinbaum, Klein, and Samet 2006).

In a Cox model, estimating the beta coefficients is all that is necessary for the purpose of inferring the effect of covariates on the hazard function. As summarized by Cleves et al. (2010) "in a proportional hazards model the effect of covariates is multiplicative(proportional) with respect to the hazard."

The Cox PH model has many equivalent representations in the literature, but is often defined as:

$$h(w|z) = h_0(w) \exp\left(x'\beta\right) \tag{2.11}$$

where w_j are the gap times between events, $h_0(w)$ is the baseline hazard function, and x' is the covariate vector.

As explained in Therneau (2000), the **Extended Cox Model** allows for stratification according to covariates, such that the observations are divided into disjoint strata or groups. Each strata has its own baseline hazard function, but common coefficient values for the coefficient vector β. Thus, the hazard for interfailure duration i in stratum k has the form $h_k(t)e^{X_i\beta}$.

Stratification is useful because it allows for adjustment of confounding covariates, or covariates which do not satisfy the proportional hazards assumptions. An unfortunate aspect of stratification in the extended Cox model is that as the baseline hazard function is not estimated, the effect or importance of the strata is not estimated(Therneau 2000). Extended Cox models may include the interaction between strata and covariates, which identifies whether the effect of covariates differs by strata. Including each covariate by strata interaction is equivalent to modeling each strata separately(Therneau 2000).

Additionally, the extended Cox model

An extended Cox model including both time-independent and time-dependent covariates takes the form:

$$h_k(t, \mathbf{X}(t)) = h_k(t) \exp\left[\textstyle\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right] \tag{2.12}$$

With $X_1, X_2, \ldots X_{p_1}$ the time-independent covariates and $X_1(t), X_2(t), \ldots X_{p_2}(t)$ the time-dependent covariates of interest(Kleinbaum, Klein, and Samet 2006). However, when the extended Cox model includes time-dependent covariates, it may no longer satisfy the proportional hazards assumption, as both the baseline hazard and the covariates depend upon time.

Fully Parametric Estimation

Although non and semi-parametric methods yield conclusions about survival times and the effect of covariates, fully parametric models may be an ideal solution. Given an assumption about the underlying distribution, parameters can be estimated to full specify the survival and hazard functions, allowing for a complete model capable of simulation(Kleinbaum, Klein, and Samet 2006).

The **Weibull Proportional Hazards Model**, sometimes referred to as a *Weibull analysis*, is a fully-parametric extension of the proportional hazards model, in that it assumes a Weibull distribution for the failure times(Collett 2003). An example of a Weibull proportional hazards model is given in Jardine, Anderson, and Mann (n.d.), where it is used to assess the effect of oil composition on aircraft and marine engines(Jardine, Anderson, and Mann, n.d.). The Weibull proportional hazards model is a unique case of the PHM, that is equivalent to the accelerated failure time model(Moore 2016).

Additionally, the **Accelerated Failure Time Model**(AFT) is a fully parametric technique in which the survival function is assumed to follow a specific parametric distribution function. The presence of covariates in the model contributes to an acceleration factor, which represents the extend to which the time until failure is shortened or lengthened.

As summarized by Cleves et al. (2010) "in an AFT model the effect of covariates is multiplicative(proportional) with respect to the survival time". The Weibull distribution is frequently chosen in AFT models and can be parameterized as:

$$S(t) = \exp\left(-\left(\frac{t}{\mu}\right)^{\alpha}\right), \quad \log(\mu) = x'\beta \tag{2.13}$$

Cox and Oakes (1998) details various methodologies for comparing distributional families for the purpose of parametric survival models, mentioning consideration of the convenience for statistical inference, comparison behavior and fit at different time durations, and evaluation log-transformations of hazard and time, among others.

2.3.3 General Intensity-Based Models

Lawless (2007) provides an in-depth summary of broad classes of "hybrid" intensity-based models that allow for the inclusion of both calendar time trends and gap times. As illustrated in the text, such models may be applicable either when either changes in the underlying event process occur or when a subject's propensity for event occurrence changes over time. In the context of equipment failures, such models may be able to reflect time trends, such as equipment degradation, or changes to the equipment itself, such as repairs performed following failures.

The **Trend Renewal Process(TRP)** model is similar to the NHPP in that it features a trend function, similar to the intensity function of the NHPP. However, the TRP model is unique in that it has the capability to describe trend in failure occurrences in addition accommodating different types of repair. In Gamiz and Lindqvist (2016), which details thorough use of the TRP, it is described as "the least common multiple of the RP and the NHPP". Another example of a TRP model used on engine failure data, including comparison with NHPP and RP models, is given in Elvebakk, Lindqvist, and Heggl (1999). A description of the usefulness of the TRP model by Lindqvist (2006) notes that it is capable of illustrating the three dimensions of repairable systems: quality of repair, existence of trend, and heterogeneity between systems.

The **Generalized Renewal Process**, detailed by Kijima (1989), incorporates two sub-models which involve the concept of virtual age. Both of the models originally introduced by Kijima involve a stochastic term on the unit interval, representing the quality of repair(effectiveness to reduce age), used to determine the virtual age following each subsequent repair. In the context of generalized renewal theory, the concept of virtual age seeks to differentiate between the operational age of a component and the actual health in relation to a new component. Extensions of the original Kijima models, namely arithmetic reduction of intensity(ARI) and arithmetic reduction of age(ARA) models are described by Tanwar, Rai, and Bolia(Tanwar, Rai, and Bolia 2014). While the ARA models follow directly from the virtual age models, the ARI model describes the repair effectiveness in terms of the change in reliability(failure intensity) immediately prior to, and following, failure. A further description of the GRP models and their usage in repairable systems is provided by Yanez, Joglar, and Modarres (2002).

2.4 Multivariate Techniques

In addition to statistical techniques for modelling recurrent events, several other multivariate methods may prove useful in accommodating numerous covariates or building predictive models for classification. Dimensionality reduction techniques such as principal component analysis facilitate the reduction of the number of model inputs while still attempting to explain maximum variation in the data. Supervised learning techniques such as artificial neural networks and logistic regression are highly flexible techniques in which classification models can be trained using multivariate inputs.

Principal component analysis(PCA) is a multivariate technique primarily used for the purposes of dimension reduction capabilities or in cases of strong correlation between predictors. The resulting product of PCA is a set of new variables, the principal components, which are linear combinations of the original variables. The results of PCA may sometimes be the desired objective of analysis, but most often the resulting principal components are used in further analysis. As described in Everitt and Hothorn (2011), the principal components have an ordering such that the first component "explains" the largest amount of variation in the data, with each subsequent component explaining a smaller amount of variation.

As outlined in Sharma (1996), when there are p original predictor variables, $x_1, x_2, \cdots, x_p$, PCA is intended to identify p linear combinations of these predictor as defined below:

$$\begin{aligned}
\xi_1 &= w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p \\
\xi_2 &= w_{21}x_1 + w_{22}x_2 + \cdots + w_{2p}x_p \\
&\vdots \\
\xi_p &= w_{r1}x_1 + w_{p2}x_2 + \cdots + w_{pp}x_p
\end{aligned} \tag{2.14}$$

Where ξ_p is the pth principle component and w_{ij} is the weight of the jth variable on the ith principal component, such that the ξ_p are uncorrelated and the w_{ij} are orthogonal. Although the total variation in the original variables can only be captured by using all p principal components, a subset of the identified principal components may be used to explain a desired amount of the original variation.

Rencher (2002) provides a detailed summary of the procedures and consequences for selecting a number of principal components, and notes that the greatest risk is retaining components that are *sample specific* or *variable specific*. In the text, sample specific refers to components that do not generalize to the population under study, and variable specific refers to components that represent only a single variable rather than a combination of variables. The presence of variable specific principal components can typically be identified by comparing the values for the component loadings, which are useful for interpretation of the components. Common methods for choosing the number of principal components to retain includes picking a cutoff for the cumulative amount of variation to be explained, typically 70-90%, picking components based on a comparison of eigenvalues to average eigenvalue, as well as comparing such values using scree plots.

Principal components are typically interpreted using the respective component loadings, which represent the the correlation between the original variables and the new variables.

As summarized by Sharma (1996), the loadings "give an indication of the extent to which the original variables are influential or important in forming new variables".

An **Artificial Neural Network(ANN)**, refers to a function intending to mimic the behavior of a biological neuron(Blockeel 2016). A biological neuron is a cell that "fires" in response to an accumulation of biological inputs, according to a specific activation threshold.

The most basic representation of an ANN is the single-layer perceptron defined as:

$$y = f(\sum_{i=1} w_i \cdot x_i + b) \tag{2.15}$$

where f is the *transfer function*, or *activation function* with w_i the weights, x_i the inputs, and b the bias or activation term(Blockeel 2016). In practice, the transfer function may take several different forms, such as a logistic function or a hyperbolic tangent functions, among others.

As described in Blockeel (2016), **Multi-layer Perceptrons(MLP)**, commonly referred to as *feed-forward* neural networks, are ANNs consisting of multiple layers such that the output from one layer is the input of the next layer. *Hidden layers* are the layers of neurons that exist between the input layer and the output layer. The number of hidden layers, in addition to the number of neurons in each layer, has an effect on the complexity of the derived approximation.

As illustrated in Bishop (1996), there are many algorithms available for training ANNs, with the most common being *backpropagation*. In short, this algorithm represents a propagation of errors back through the network, in order to minimize a specific error function. According to Bishop (1996), "most training algorithms involve an iterative procedure for minimization of an error function, with adjustments to the weights being made in a sequence of steps." Additional popular training algorithms include the Newton method, the Levenberg-Marquardt method, and the Quasi-Newton method.

Kutner et al. (2005) provides a detailed summary of the advantages and disadvantages of ANN models as compared to traditional statistical modelling. The primary disadvantages of ANN usage are that model parameters are generally uninterpretable, and covariate effects must be identified through the use of conditional effects plots, among other tools. Additionally, tools for traditional model diagnostics for outliers, lack-of-fit testing, and covariate significance testing are less established.

On the other hand, ANNs are not contingent upon many of the common independence and distributional assumptions of traditional statistical models. Furthermore, usage of ANNs allows for modelling of complex response surfaces when using large samples. Additionally, Kutner et al. (2005) notes that the usage of bounded logistic activation functions makes ANNs more robust to the influence of extreme outliers.

Logistic regression is a member of the Generalized Linear Models(GLM) family, and represents useful technique for applying a linear regression model to predict a binary outcome. As detailed in Kutner et al. (2005), a multiple logistic regression can be used to predict Bernoulli random variables, Y_i, with expected values $E\{Y_i\} = \pi_i$, where:

$$E\{Y_i\} = \pi_i = \frac{\exp(\mathbf{X}_i'\beta)}{1+\exp(\mathbf{X}_i'\beta)} \tag{2.16}$$

Logistic regression is defined by the *logit* canonical link function, such that:

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathbf{X}'\boldsymbol{\beta} \tag{2.17}$$

where $\mathbf{X}'\boldsymbol{\beta}$ represents the linear predictor.

Although logistic regression is contingent upon traditional distribution assumptions, this allows for the usage for classical diagnostics tools for significance testing, residual analysis, and outlier detection. Furthermore, the use of the logit link functions allows for straightforward interpretation of respective covariate effects.

In addition to standard logistic regression, methods of ordinal and multinomial logistic may prove useful for classification problems. Other classification models include random forest, deep learning, decision trees, and support vector machines(Blockeel 2016).

In summary, given divergent data sets collected from corrective maintenance actions, preventive and condition-based maintenance strategies, it is necessary to process the data prior to analysis. This procedure will take into consideration the different structures, formats and information contained within each source. Through the use of a KDP tailored for the characteristics of the different data sets, the data will be prepared for analysis, and later integration, for future analysis. Eventually, the data will be analyzed using the various techniques in order to derive a decision support framework.

Chapter 3

Methodology

The following section outlines the methodology of the research project, detailing the work to be done at each step, the underlying motivation, as well as the contribution to the research objectives. As highlighted in Figure 3.1, the research methodology will have a linear flow, with each procedure building upon the work done in the previous. The first step will provide an introduction into the data and respective methods of collection. The next step will detail the steps to be involved in pre-processing the data in order to prepare it for analysis. After pre-processing, a descriptive analysis will be carried out with each respective data set. The fourth step will detail the procedures to be used to build models evaluating the reliability of the equipment. The final step will detail the steps involved in creating a predictive model that integrates data from all existing sources.

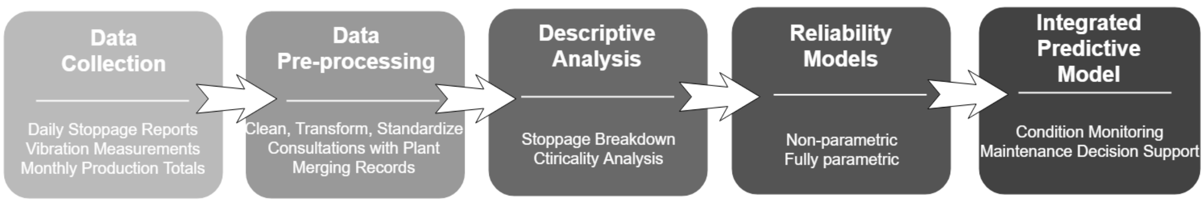

Figure 3.1: Overview of research methodology

3.1 Data collection

Although the data to be used in the coming analysis was not collected by the researcher, the following section will detail the methods and motivation used for collection, in addition to providing a description of the data sources themselves.

3.1.1 Plant stoppage event data

The available stoppage data consists of a repository of 157 Microsoft Excel documents,collected from 2015,2016, and 2017, listing all of the stoppage events that occurred during the week. As displayed in Figure 3.2, as a stoppage occurs within the plant, a maintenance engineer records the date, time, equipment, and plant section, along with a brief description of the stoppage cause or solution. Additionally, the group responsible for resolving the stoppage is also noted.

The stoppages for each day are arranged in a list under a common heading, and additional stoppages are appended to the list as they occur. Using this method, any stoppage occurring within the plant is cataloged and can be reviewed by opening the historical report for the respective week. In this current format, the reports provide a summary of the stoppage events for a given week, but are not intended for statistical analysis.

WEEK 01 STARTING FROM (MON) 29th December 2014 Ending (SUN) 04th January 2014 Today's Date 30th December 2014

DATE/ DAY	SECTION	STOP HRS	START HRS	HRS	RESP	EQUIP' CODE	COMMENTS	M
MON	RAW MATERIAL PREP.							
29/12/2014	(a) Limestone processing	B/F	10:45	4:45	PROC	P02A	Lack of front end loader - Shifting material at the C/mills & Clinker shed	
	(17.0hrs)	13:10	17:59	4:49	PROC	P02A	Lack of front end loader - Shifting material at the C/mills	
		20:43	0:28	3:45	PROC	P02A	Lack of front end loader - Shifting material at the C/mills	
		2:20	C/F	3:40	PROC	P02A	Lack of front end loader - Shifting material at the C/mills	
	(b) Kunkur stacker line	B/F	C/F	24:00	WKS	W02B	Lack of processing room - Pile full	
	(24.0hrs)							
	(c) Double hammer mill	B/F	7:30	1:30	PROC	P02C	Lack of front end loader - Shifting material at the mills / impact crusher	
	(19.7hrs)	9:30	15:13	5:43	PROC	P02C	Lack of front end loader - Shifting material at the mills / impact crusher	
		17:30	C/F	12:30	MECH	02CR31	Down to replace broken safety pin	B

Figure 3.2: A sample of one of the weekly stoppage event reports

3.1.2 Vibration measurement data

As mentioned previously, the vibration observations were recorded for a subset of all equipment items as deemed vital by the plant organization. The measurements were observed using a handheld probe, without causing interruption or downtime to the machine, and input into an electronic record. The purpose for collecting vibration measurements is to provide an indication as to the health status of the equipment, referred to as *condition monitoring*.

The vibration source data consisted of a repository of tables and graphs, displayed in an HTML file, as depicted in Figure 3.3. Each recorded vibration measurement contained the individual equipment code, the component being monitored(fan, motor, etc.), the type of vibration, the amount of vibration, and the percentage change in vibration since the previous measurement. With this given structure, each piece of equipment may consist of several components, with different types of vibrations being recorded from each component.

MOTOR

Last Measurement

Set Name	POINT name	Date/Time	Last value	Units	% change	Alarm status
05FN02 New	VDE	9/15/2017 9:28:39 AM	1.584	mm/s	199	---
---	---	11/27/2016 10:54:39 AM	0.529	mm/s	-47.7	---
---	---	11/26/2016 3:54:38 PM	1.012	mm/s	-19.4	---
---	---	10/22/2016 9:15:32 AM	1.255	mm/s	15.3	---
---	---	10/9/2016 10:49:12 AM	1.089	mm/s	2.39	---
---	---	9/21/2016 11:01:31 AM	1.063	mm/s	9.77	---
---	---	9/12/2016 9:01:08 AM	0.969	mm/s	-0.431	---
---	---	8/7/2016 10:31:26 AM	0.973	mm/s	10.9	---
---	---	8/6/2016 4:45:35 PM	0.877	mm/s	2.16	---

Figure 3.3: A sample of the observed vibration measurement data reports

3.1.3 Monthly production data

In addition to stoppage event records, the plant also maintains a record of the amount of production or processing by each area of the plant. This figure is recorded as the total production for each month, measured in tons. The plant has also provided information regarding the specific maximum throughput, or capacity rate for each plant section measured in tons of material per hour.

3.2 Data pre-processing

As each data source was primarily recorded for the purpose of obtaining a high-level overview of the operational status of the plant, the format of each source make it unsuitable for statistical analysis. The following section will detail the necessary procedures for aggregating, cleaning, and transforming each dataset in preparation for analysis.

3.2.1 Plant stoppage event data

Designed to provide a high-level synopsis of plant availability during each week, there is significant value to be realized from aggregating the reports, and structuring the data, such that management can understand how the reliability of the system, subsystems, and components has evolved over the past three years.

In order to begin pre-processing the stoppage event data, the records must first be collected from each of the 157 files, and aggregated into one worksheet The *pandas* library in the Python language will be used extensively to aggregate all of the files into a single dataframe, and systematically strip out stylistic formatting and non-informational headings.

In addition to equipment code, plant section and interval time stamps, each stoppage record contains a *Comments* field in which the person who recorded the stoppage could insert additional information about the event. This would allow the employee to provide information otherwise not recorded in the other fields, potentially describing the *failure mechanism* of the equipment or the *maintenance action* performed in response to the failure, among other possibilities.

The ability to extract additional information about each failure event will allow for further classification of stoppage events, such as identifying whether an equipment failed because of a faulty bearing or leaking seal. More detailed information about the *failure mode* of an equipment can improve the effectiveness of prescribing specific maintenance interventions, a maintenance engineers will have a narrower scope of inspection, hopefully resulting in shorter maintenance durations and less equipment downtime.

Furthermore, insight into the *maintenance action* taken in response to equipment failure can be used to identify the effect of the action on the equipment reliability. For example, a complete history of *failure mechanisms* and *maintenance actions* may reveal a difference in effectiveness of equipment replacement versus equipment repair in response to failure. Additionally, this history may identify the extent to which a specific maintenance action is effective in response to a specific failure mode, or matched to certain equipment. Perhaps repair, or other maintenance action, is largely ineffective on certain equipment, or perhaps only a finite number of repairs are effective, after which point a replacement must occur.

It is advantageous to attempt to extract as much meaningful information about each stoppage from the *Comments* field as possible. Doing so will greatly improve the modelling capabilities and allow for more accurate predictions regarding the timing and specifics of maintenance interventions. As the *Comments* field is entirely unstructured, and may vary depending upon the stoppage category, plant section, individual employee, etc. text mining will be useful to understand the meaning, value, and patterns of the available information. The *tidytext* package in the R software provides extensive functionality for text mining based on the concepts outlined by Robinson and Silge (2017).

3.2.2 Vibration measurement data

The vibration data was provided in a large HTML document, in which the observed readings are presented in a combination of data tables and graphical plots. In order to transform the data into a usable format, the Python package *BeautifulSoup* will be used to parse the HTML file and extract the source data from consecutive table elements. As the graphical plots represent the same data from the tables, they will be ignored. After extracting the observed vibration readings into a single data frame, human input errors will be corrected for the date, equipment code, and measurement type.

In addition to being recorded at irregular intervals, measurements were taken by hand causing observations made at essentially the same time to have differing time stamps. In order to combine readings from the same equipment and sub-component taken on the same day, only the date will be retained, ignoring the exact time of day.

3.2.3 Monthly production data

Plant production values were recorded for 10 sub-sections of the plant, and presented in monthly summary reports. Aggregating this data source will involve manually opening each report, copying the monthly total, and entering it into a new table. Once aggregated, the final table will contain one row for each sub-section, with the columns representing the total production tonnage for each month.

3.3 Descriptive analysis

Following the pre-processing of both data sets, an exploratory statistical analysis will be performed for each. This data exploration will yield some preliminary insights into what kind of information the datasets contain, what information can possibly be extracted, the completeness of the data, as well as any limitations for future modeling. In addition to describing the data sets themselves, the exploratory phase will serve to identify several critical sections and items of equipment that will be examined more closely in the subsequent analyses.

3.3.1 Plant stoppage event data

After processing and combining the weekly stoppage reports into a single data source, spanning the three year period, valuable insights can be realized from the historical record. This can allow maintenance engineering to track downtime and availability over a much longer time duration. Attempting to improve reliability and availability of production equipment when only viewing a weeks' worth of information makes identifying trends very difficult. In addition, it is difficult to separate between more reliable and less reliable systems, that is, systems needing maintenance and those that do not, which may cause maintenance engineering to misidentify equipment or entire systems.

As the stoppage event data provides the largest source of information regarding the performance of the cement plant, it will serve as the basis for a criticality analysis. This is because each stoppage event is classified by respective event category, equipment, and plant section; thus, visualizing the distribution of failure and non-failure events throughout the plant will aid in identification of critical areas of focus. Plant sections or specific equipment in which failure events are large contributors to overall downtime will be identified as the areas in which the largest improvement in downtime reduction can be made. Similarly, the total number of failure events per section or equipment provides a similar endpoint for measuring criticality.

3.3.2 Vibration measurement data

In addition to the value provided by stoppage and production records, the vibration readings may serve as an adequate indicator of the overall mechanical health of an item of equipment prior to experiencing a failure. A descriptive analysis of this data should facilitate a better understanding of the extent to which these variables represent equipment health, and how best to incorporate them into future models. Given that the objective is to build models using an integration of all data sources, and not all plant equipment is monitored for vibrations, the descriptive analysis will also indicate which equipment is most frequently monitored, identifying candidates for future integrated models.

A descriptive analysis of this data may include boxplots to assess and compare the range of values across variables, time series plots, and correlation analysis. Additionally, given the 9 possible vibration variables, dimensionality reduction techniques such as principal component analysis may provide useful.

3.3.3 Monthly production data

A descriptive analysis of the monthly production data will aim to provide insight into the distribution of workload between subsections of the plant. Such knowledge may provide significant explanatory power for reliability models in the forthcoming analysis. As the data consists of monthly figures, time series plots may be the most useful method for descriptive analysis.

3.4 Data integration

Following the criticality analysis of the stoppage event records, a section of the plant will be identified as the focus of further analysis. Ideally, the criticality analysis will reveal a specific type of equipment about which sufficient data has been recorded such that it will be possible to model the recurrent failure events, while accounting for the condition of the equipment between failures, using other data, such as vibration measurements.

In order to facilitate attaining the main objective of the study, using predictive modelling, it is imperative that the three types of data are integrated or fused. Given a specific critical equipment, the integrated data will contain a record of all stoppage events, production figures, and vibration measurement the equipment has experienced.

This means that in addition to identifying the gaps between failure events, the integrated data will also account for maintenance interventions and non-failure stoppages that occur during these gap times. Furthermore, the data will contain the full production rate history, such that the predictive models can use the changes in production to better estimate equipment reliability. Finally, the integrated data will also include all observed vibration measurements in order to assess the condition of the equipment between failures.

There are no established procedure for data integration, as it entirely depends on the structure of the available data as well as the intended use. However, the aim of this integration is to use both the production and vibration observations to *augment* the stoppage records. Each data set contains both an equipment identifier and a time variable, which will serve as the reference values for the integration. Given the failure records for a certain equipment, the timestamps of the failure events can be referenced against the production records to create time-dependent covariates for production. The procedure will be repeated for the vibration data, using timestamps to identify all vibration measurements that were observed between failure events, and incorporating them into the integrated data as additional time-dependent covariates. In this form, the integrated data will represents all available information about the performance and condition of an equipment in between failure events.

3.5 Reliability models

After identifying several equipment items through the criticality analysis, several basic reliability models, using only the failure event history will be built using survival analysis techniques. Although not yet incorporating all of the data sources, the basic models will still provide some preliminary insights into the reliability characteristics of the critical equipment. Tests for trend, distributional assessments, and model estimation using non-parametric, semi-parametric- and fully parametric methods will be performed.

3.6 Integrated predictive models

The final element of the analysis will involve an extension of the previously established models to identify the potential for a model with the ability to predict future failure events from all available data sources. Some of the survival analysis methods to be used include extended Cox models and parametric accelerated failure time models. Additionally, the use of these models for the purpose of predicting equipment reliability and MTBF will be demonstrated.

In addition to predicting equipment reliability, classification models such as artificial neural networks and linear regression may be used to assess the capability to predict future maintenance actions, using the integrated dataset. ANN models are highly flexible classifiers, while logistic regression provides simple interpretation of covariate effects, both of which may prove insightful.

In short, the integrated predictive models will assess the extent to which the available data can be used to improve the understanding of the reliability characteristics of specific equipment. The integrated predictive models will be used to make inference about reliability in addition to providing examples of how the models can be used in practice.

Chapter 4

Results and discussion

The following chapter will present the results of the research after applying the previously outlined methodology. As all data sets were collected prior to the start of the research, this chapter will begin with a detailed description of the steps undertaken for pre-processing the data. The pre-processing section will include an explanation of the information contained within each data source and what value it will contribute to the analysis. After pre-processing, the data is now ready for analysis, which will begin with a descriptive analysis of each data set In addition to providing a high-level summary of each data, the descriptive analysis will highlight any nuances or insights that may already be appearing. Along with the descriptive analysis, the process of performing a criticality analysis, in which important plants sections and equipment are identified, will be presented. Following the identification of critical equipment, the process of data integration will be expounded, while detailing how the integrated data can be used for predictive modelling.

Next, basic reliability models will be estimated using the failure event history of the critical equipment. The basic analysis will identify and present reliability characteristics of the critical equipment that can be extended in the final predictive models. Finally, the basic models will be extended using the entirety of the integrated dataset in order to generate models to predict reliability based on past events, production level, and vibration condition monitoring.

4.1 Data pre-processing

4.1.1 Plant stoppage event data

After aggregating all of the stoppage event records into a single data frame, duplicate headings and formatting were removed to leave only the raw data. As illustrated in Figure 3.2, values such as *date* and *subsection* were only populated for the first relevant stoppage, requiring additional scripting to identify rows for which these values should be carried over. Date, time, and other manual entry errors were manually corrected as discovered, and any duplicate event records were removed.

Given that the maintenance engineers recorded stoppage events in shifts, if a piece of equipment was already stopped at the start of a new work day(6:00am), a new stoppage entry was created for the current day, and a code indicating "before the shift" was assigned to the stop time of the new event. Similarly, if an equipment remained stopped at the end of a shift(6:00am), a code indicating "continued stoppage" was assigned to the equipment start-up time. As a result, unless a stoppage occurred entirely within a single shift, the total stoppage duration was computed by tracking stoppage events across multiple days and sometimes multiple report files. In addition to inflating the number of daily events recorded, spreading information about a single stoppage event across multiple files increases the risk of human error in data entry.

To resolve this issue, the full data frame was parsed to identify stoppage events that span multiple rows, such that the beginning timestamp of the stoppage was recorded in the first row and the end timestamp of the stoppages was recorded in the last row. These rows were merged into a single event in only a single row, listing both beginning and ending timestamps of the stoppage. This correction greatly reduced the number of data rows, ensuring that each row contained the complete information for only a single event. Cleaning and merging the stoppage records reduced the number of rows from 37549 to 30380, while retaining all available information.

Finally, as it is important to be able to distinguish between different types of stoppage events(e.g., failure, planned maintenance, etc.) to achieve the objective of the analysis, several unique categories of stoppages were identified based on the noted responsible organization. The coded abbreviation and description of each of the stoppage categories is provided in Table 4.1, in which failure events are further distinguished as either *mechanical failures*(MECH) or *electrical failures*(ELE).

Table 4.1: Cement plant stoppage categories

Code	Stoppage category description
SC	Circumstantial stoppage not related to mechanical failure of breakdown
SM	Stoppage caused by lack of raw materials
BF	Stoppage due to lack of storage space
SP	Stoppage related to utilities or related projects
SD	Stoppage related to planned shutdown
SL	Stoppage in response to low sales volume
ELE	Stoppage caused by failure of electrical system
ENG	Stoppage to perform planned maintenance
MECH	Stoppage due to mechanical equipment failure
MOB	Stoppage due to lack of mobile equipment(eg. front-end loader)
ICT	Stoppage related to IT infrastructure
PROC	Stoppage caused by plant processes not classified as breakdowns
PM	Stoppage related to preparation for planned maintenance

Text mining for failure mechanism and maintenance action extraction

In addition to equipment code, plant section and interval time stamps, each stoppage record contains a *Comments* field in which the person who recorded the stoppage could insert additional information about the event. In total there were 30380 total stoppage events, each containing a short sentence describing details relevant to the event. The first step in analyzing the text was to turn this list of sentences into a list of single words, called **tokens**. In this format, the tokens could be analyzed to identify their frequency of occurrence, in which stoppage category they occur, their sentiment, and even their part-of-speech. Using the *tidytext* package, tokenizing the comments resulted in 163643 total tokens, or 1535 unique words, as depicted in Table 4.2.

Table 4.2: Number of tokens and unique tokens

	Number of Tokens	Number of Unique Tokens
After tokenizing comments	163643	1535
After removing stopwords	122296	1461
After removing numerics and punctuation	119676	1122

Note:
Any token containing numeric characters was removed
Punctuation was removed from each token while preserving the remaining characters

Although tokenizing the stoppage comments separates each word from the rest of the sentence, the identifier of the originating sentence, as well as the stoppage category can be retained. Grouping the tokens by stoppage category makes it possible to identify the most frequently used words to describe stoppages for each category. Figure 4.1 shows the 10(unless fewer than 10) most frequently used words for each respective stoppage category. This comparison is informative as it not only describes how categories of stoppages differ from one another, it begins to describe the spectrum of stoppages that occur within each category. As shown in the figure, the most frequently used terms under the *Planned Maintenance*(PM) category are “Planned”, “maintenance”, “preparation”, and nothing else. In contrast, the *Mechanical*(MECH) category shows a variety of common terms to describe stoppages, namely “tripped”, “repaired”, “leaking”, etc..

As the concepts of *failure mechanism* and *maintenance action* are only applicable for failure events, further insights will focus of the events categorized under *MECH* and *ELE*.Considering only failure events consists of 4837 stoppages and 24749 tokens (835 unique tokens). In addition to identifying the frequency of word occurrence, **Sentiment Analysis** is a popular technique for text mining, which involves assigning a sentiment, or subjective emotional association, to each token(Robinson and Silge 2017). In the tidytext package, the sentiment of a word is assigned based on either of three lexicons; the **AFINN** lexicon assigns a sentiment in terms of emotions(joy, fear, anger, etc.), the **bing** lexicon assigns a sentiment as either positive or negative, and the **nrc** assigns an integer score between -5 and 5. For the purpose of extracting additional information about equipment stoppage events, associating words with emotions may not be useful, but binary classification of words as either negative or positive may provide a step towards separating words describing the failure mode from those describing the related maintenance action. Furthermore, using a range of integer values to represent the sentiment may demonstrate the the potential to describe the severity of failures modes and maintenance actions.

Figure 4.1: 10 most frequently used terms per stoppage category

The results of a sentiment analysis on the failure stoppage comments using the *Bing* lexicon are shown in Figure 4.2. Each graph indicates the frequency of the 10 most frequently used terms for both positive and negative sentiments. As evident by the figure, the initial classification of words by sentiment appears intuitive, since words related to failure modes, “alarm”, “fault”, “leaking”, etc., are classified as negative, and words associated with maintenance actions, “rectify”, “cleared”, “restored”, etc., are classified as positive. However, words typically associated with a negative sentiment(by the nrc lexicon) are used far more frequently than words associated with a positive sentiment. This could indicate that the majority of the information contained in the *Comments* section of each stoppage event refers to the failure mode.

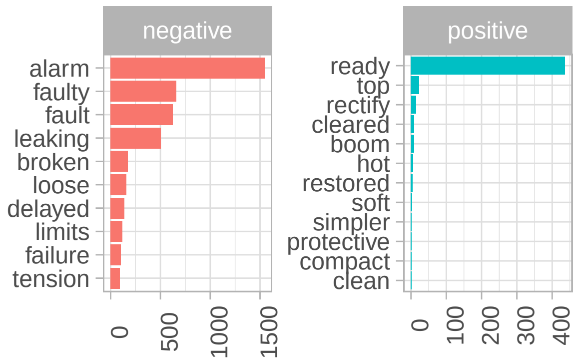

Figure 4.2: Most frequently used terms to describe failures according to sentiment

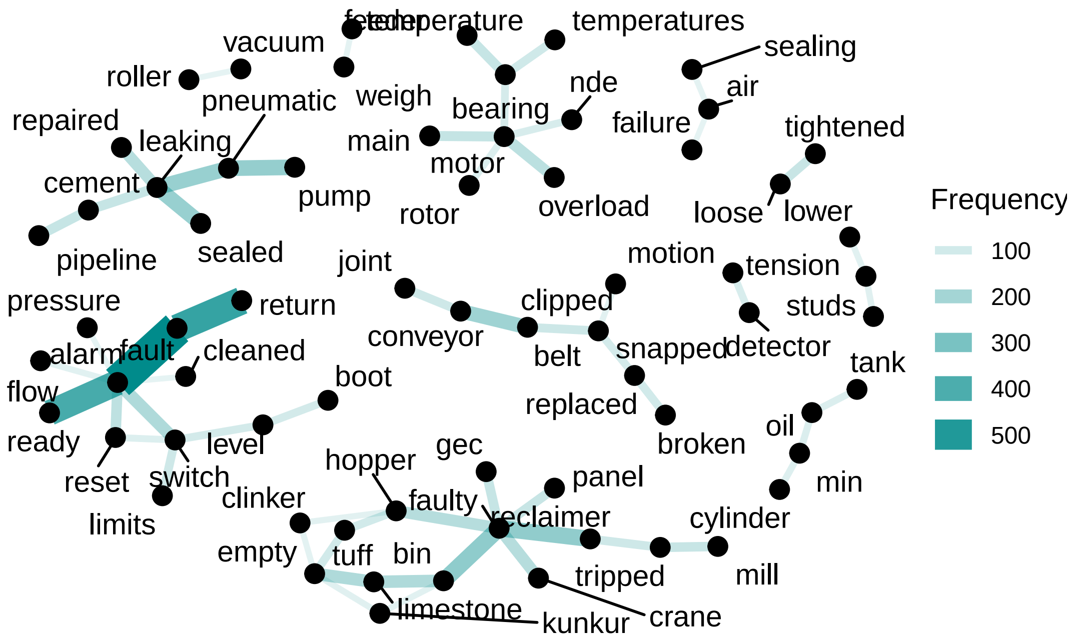

Figure 4.3: Graph of the most frequently occurring bigrams for failures

In addition to identifying the frequency with which unique words are used to describe stoppages, it is also possible to identify the most commonly used pairs of words, called **bigrams**. The network graph in Figure 4.3 depicts a portion(bigrams occurring 50 or more times) of the most frequently occurring bigrams in relation to other bigrams. In the network graph, each node represents a unique word, and nodes are connected when both words occur in the same bigram. The size and opacity of the line represents the frequency with which the bigrams occur. Nodes directly connected to many other nodes represent words that frequently appear with other words. For example, *faulty*, *alarm*, and *leaking* are nodes with many connections, indicating that they frequently appear next to many different words. This can be validated from the stoppage records, which indicate numerous failures caused by faults and alarms. These words are likely being used to describe the failure mode of the respective stoppages. This same type of representation can be performed with **n-grams**, where n is the number of words occurring in succession.

To gain more insight into the words occurring together, we employed **correlation analysis**. Correlation analysis identifies the correlation coefficient, as defined in Equation 4.1, between respective pairs of words occurring together in the same comment. Although a high correlation for a pair of words does not imply that the words occur very frequently, it does imply that words occur together, or not at all. Table 4.4 shows the 15 pairs of words with the highest respective correlations.

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}} \tag{4.1}$$

Table 4.3: Pairwise frequency combinations used to calculate correlation

	Has word 2	No word 2	Total
Has word 1	n_{11}	n_{10}	$n_{1.}$
No word 1	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Table 4.4: 15 highest correlations between pairs of words

Word 1	Word 2	Correlation
public	holiday	1.0000000
house	keeping	0.9994159
hot	gases	0.9985030
motion	detector	0.9973061
planned	maintenance	0.9970205
weighout	taking	0.9946017
physical	weighout	0.9911390
impact	crusher	0.9894996
front	loader	0.9883418
sample	analysis	0.9863776
physical	taking	0.9857886
tertiary	materials	0.9857834
touched	table	0.9845732
final	product	0.9799933
shifting	mills	0.9764481

As illustrated in Table 4.4 some pairs of words with high correlations are less informative, such as “public holiday”. On the other hand, some pairs such as “hot gases” and “tertiary materials” may imply a failure due to lack of or inability to produce, respectively.

Additional information was extracted from the comments of each stoppage by tagging each tokenised word with the respective **part-of-speech**. The *UDPipe* R package performs several Natural Language Processing functions using pre-trained or customized models. In this case, a pre-trained model was used to annotate the tokenised terms with the relevant part-of-speech. The graphs in Figure 4.4 display the 10 most commonly used terms to describe failures,divided by their part-of-speech. The full definitions of the part-of-speech abbreviations are provided in Table 5.6 in the appendix.

Despite several miss-classifications when using the pre-trained model, classification by part-of-speech proved the concept of using categorization to extract relevant information from the stoppage comments. As shown in the figure, the list of most common verbs contain words such as *tripped* and *leaking* which could be identified as specific failure modes of a stoppage. The list also contains *repaired* and *replaced* which could be identified as the respective maintenance actions performed in response to the stoppages. The lists of adjectives and nouns also show potential to identify a specific component of a unit of equipment(*motor*, *pump*, *switch*, etc.), as well as its operational status(*faulty*, *empty*, *worn*, etc.).

The use of a customized model with a standardized lexicon of failure modes and maintenance actions may make it possible to extract and classify the failure mode, unit component, and maintenance action for each individual stoppage. This could allow maintenance engineers to estimate more specific reliability models according to specific failure modes or maintenance actions.

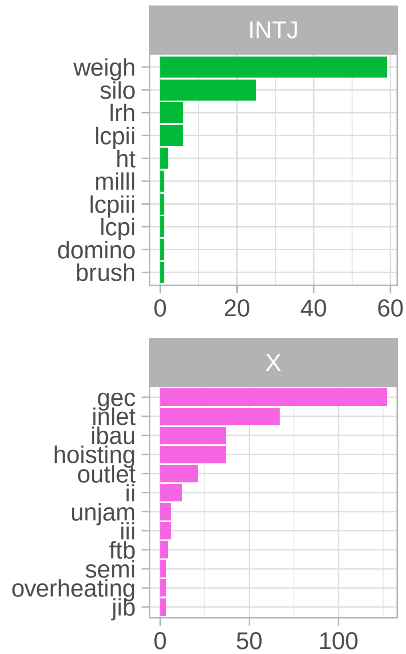

Figure 4.4: 10 most frequently used terms for failures by part-of-speech

Continuing with the subset of failures only, an additional sentiment analysis was performed on the most frequently used verbs, shown in Figure 4.5. The results of this analysis appear to confirm the notion that the stoppage comments contain far more negatively associated words than positive ones, in terms of both number of unique words and total number of words. Despite this fact, the pre-trained model appears to classify failure modes(negative verbs) and maintenance actions(positive verbs) quite accurately.

Figure 4.5: Most frequently used verbs to describe failures according to sentiment

As the ultimate goal of the text mining was to extract a failure mechanism and maintenance action according to (“ISO 14224:2016 (Collection and Exchange of Reliability and Maintenance Data for Equipment)” 2016) standards, the terms identified in Figure 4.5 were manually mapped to the failure mechanism and maintenance action categories in Tables 5.1 and 5.2 in the appendix. After an initial mapping, the text mining procedure was repeated iteratively, until all possible failures could be categorized. The remaining failure events for which the description did not provide enough information for classification were assigned to the *other* category.

The results of this text mining endeavor are summarized in Table 4.5 and Table 4.6, which show the distribution of respective failure mechanism and maintenance action categories extracted from all of the failure events.

Table 4.5: Frequency of failure mechanisms

FM	Meaning	Count
1.0	General Mechanical	2470
1.1	Mechanical Leakage	531
1.2	Abnormal Vibration	70
1.3	Alignment Failure	130
1.4	Deformation	6
1.5	Looseness	145
1.6	Sticking	26
2.4	Wear	32
2.5	Breakage	482
2.7	Overheating	426
3.0	General Instrument Failure	127
3.4	Instrument Adjustment Failure	116
4.4	Faulty Power	2
6.3	Miscellaneous Other	274

Table 4.6: Frequency of maintenance actions

MA	Meaning	Count
1	Replace	668
2	Repair	1034
4	Adjust	2933
5	Refit	19
9	Inspection	18
12	Other	165

4.1.2 Vibration measurement data

As the vibration data was provided in a large HTML document, Python was used to scrape the data from the table elements and insert the entries into a new dataframe. Whitespace, formatting, and duplicate headings were removed, and human input errors for the date, equipment code, and measurement type were manually corrected. Since vibration measurements that were recorded on the same day were observed as close together in time as possible, they were combined into single rows creating a multivariate entry.

In this format, each row represents observations taken on a unique component of a unique equipment on a given day, with up to 9 variables measured. The 9 vibration variables and descriptions of the type of vibration being measured are provided in Table 4.7.

Table 4.7: Description of vibration measurement variables

Variable	Unit	Description
ADE	mm/s	Axial drive end
ANDE	mm/s	Axial non-drive end
HDE	mm/s	Horizontal drive end
HDE.CAV	mm/s	Horizontal drive end cavitation
HDE.ENV	gE	Horizontal drive end enveloping
HNDE	mm/s	Horizontal non-drive end
HNDE.ENV	gE	Horizontal non-drive end enveloping
HNDE.FL	mm/s	Horizontal non-drive end
VDE	mm/s	Vertical drive end
VNDE	mm/s	Vertical non-drive end

As the vibration measurements were observed manually using a portable probe, some observations were incomplete such that values were recorded for some variables and not for others. A summary of the proportion of missing values, along with the respective combinations and proportions is provided in Figure 5.1 in the appendix. Approximately 75% of the vibration measurements were complete observations, with *HDE.CAV*, *ANDE*, and *ADE* being the most frequently missing, with 53, 44, and 40 missing values respectively. In order to make use of these partial observations, the missing values were imputed from the remaining vibration measurement through the use of the PMM (Predictive Mean Matching) algorithm from the *MICE*(Multivariate Imputation by Chained Equations) package in R. The PMM algorithm imputes a missing value by identifying observations with similar remaining covariates and randomly draws a value from these observed candidates to ensure a plausible outcome.

4.1.3 Monthly production data

As the plant was divided into only 10 subsection for the purpose of recording monthly production, this is the smallest data source. Aggregation of this data involved manually coping the values from each monthly report into a new dataframe. Once aggregated, the final table contained one row for each sub-section, with the columns representing the total production tonnage for each month.

Given that the maximum production rate for each sub-section of the plant was also provided, but in tons per hour, this figure was be multiplied by the number of hours in each respective month to identify the maximum monthly production. Dividing the observed monthly total by the maximum monthly total yielded a measure of *relative* production rate as compared to the maximum capacity. This provides an effective measure of the relative load that a sub-section of the plant experienced each month.

4.2 Descriptive analysis of respective data

The following sections will provide a descriptive analysis of each data set after being pre-processed.

4.2.1 Plant stoppage event data

After processing the historical stoppage data, it is useful to generate some summary measures relating the data to the structure and organization of the cement plant. Given that for each stoppage event, the time and date of the beginning and end of the event are recorded, as well as a classification of system, subsystem, and component, insights are available at each organizational level. Using the stoppage categorization and plant section outlined in the previous chapter, we compared a frequency breakdown of the stoppages. Table 4.8 provides a breakdown of the total number of stoppages by category for each respective section of the plant. As shown in the table, Section 2 observes the single largest number of stoppages, 7880, which are categorized as *PROC*. Although Section 2 observes the largest number of stoppages, Section 6 observes the largest amount of *failure*(*ELE* or *MECH*) events, 1098 and 1225 respectively. A graphical comparison of the stoppage frequency breakdown is provided in Figure 4.7.

Table 4.8: Total number of stoppages by category per plant section

	Plant Sections						
Category	2	3	4	5	6	7	8
BF	38	NA	15	548	65	NA	NA
ELE	376	559	14	326	1098	144	NA
ENG	258	62	NA	8	250	398	NA
ICT	NA	NA	NA	NA	NA	35	NA
MECH	236	551	8	185	1225	113	2
MOB	62	1	NA	NA	39	NA	NA
PM	119	63	NA	1	257	377	NA
PROC	7880	1011	12	709	4188	89	1
SC	360	207	13	750	679	2707	NA
SD	10	5	NA	112	9	8	3
SL	NA	NA	NA	NA	1	2618	NA
SM	12	NA	NA	36	60	11	1
SP	93	235	NA	401	543	183	NA

Note:
NA indicates that no stoppages of that category were observed

4.2.2 Criticality analysis

Figure 4.6 displays a comparison of the total number of stoppages, as well as the percentage contribution to overall plant downtime broken down by stoppage category and also plant section. The two left plots, in which stoppage events are grouped by the plant section in which they occur, identify plant sections 2 and 6 as having both the highest total number of stoppages(top left plot) as well as being responsible for the largest contributions to overall plant downtime(bottom left plot). As indicated by the two right plots, process(PROC) and circumstantial(SC) stoppages are the most frequently occurring categories, as well as the largest contributors to overall plant downtime. Neither circumstantial stoppages(e.g., lack of alumina or hot gases) nor process related stoppages(e.g., silo change over) are considered equipment failure events.

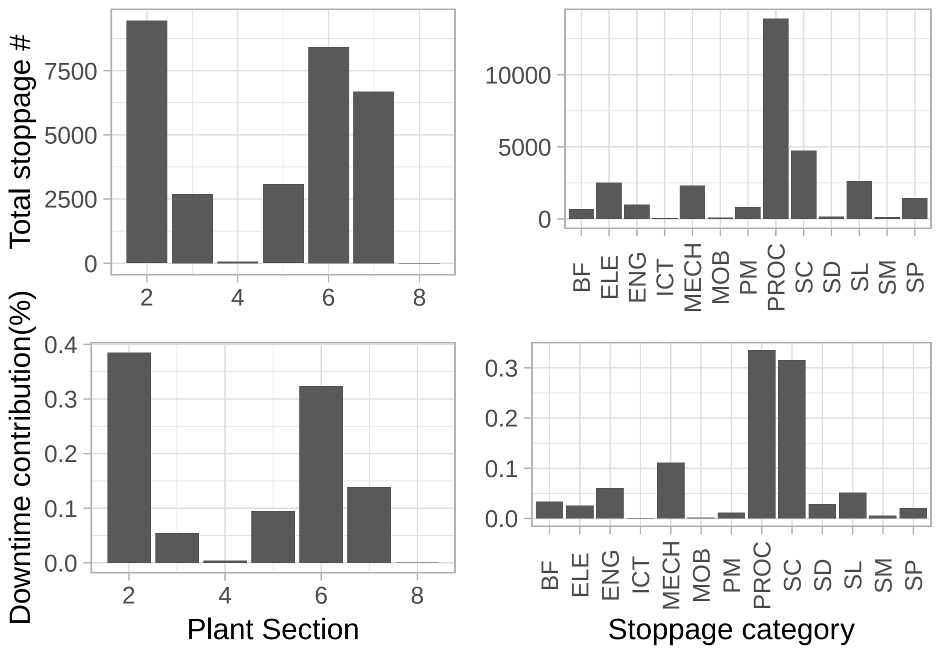

Figure 4.6: Total number of stoppages and contribution to overall downtime, by plant section and stoppage category. The top plots depict the total number of stoppages by plant section(upper left) and by stoppage category(upper right). The Lower plots depict the contribution to plant downtime by plant section(lower left) and by stoppage category(lower right).

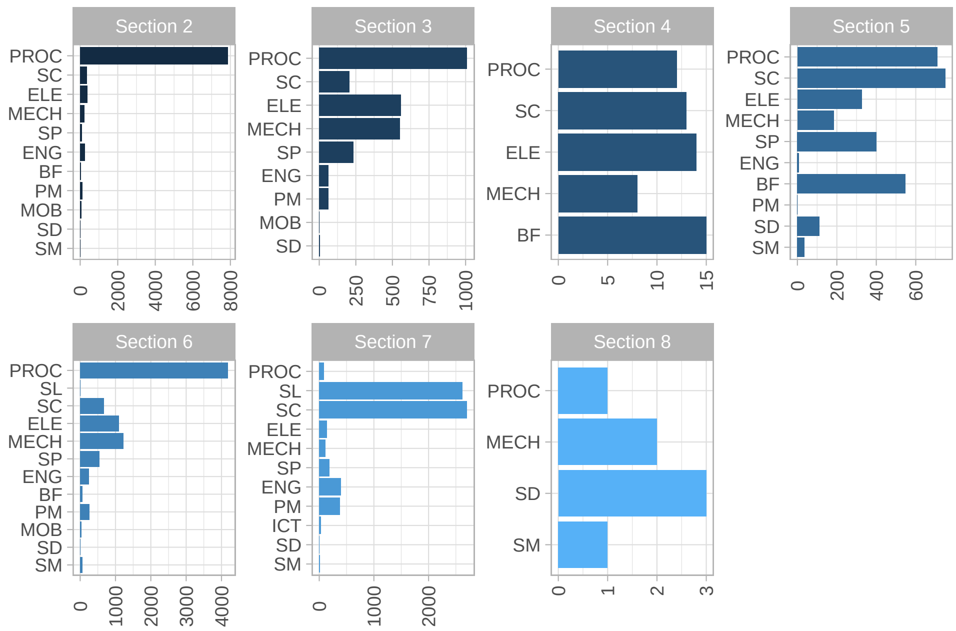

Figure 4.7: Total number of stoppages in each category, per plant section

Figure 4.7 provides a breakdown of the total number of stoppage events, per category, for each of the respective plant sections. Following the previous assessment that identified section 2 as the the most critical section in terms of all stoppage events, it is now evident from Figure 4.7 that almost all of these stoppages(nearly 8000) are process stoppages(PROC) with very few stoppages categorized as failure events. Although the majority of stoppage events in section 6 are also process related, this section has the largest number of failure events. Additionally, sections 3 and 5 are responsible for the next highest number of failure events, respectively. In light of this, section 6 is identified as a critical section based on not only the number of failures, but also the contribution to downtime.

Following from the observation that Section 6 observes the largest number of failures, the following analysis will consider only those failure events. As suggested by Jardine and Tsang (2013), a Pareto chart and jackknife plot can help identify equipment items that are prone to frequent or long stoppage events, which may be of particular interest going forward.

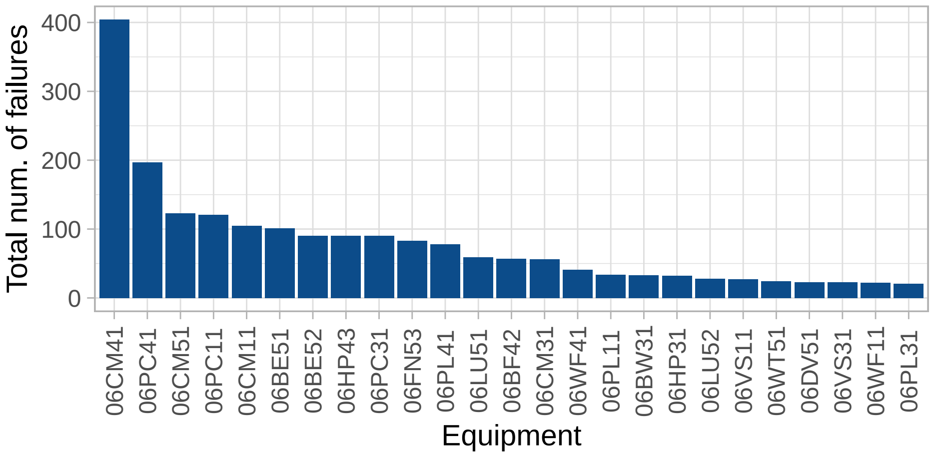

Figure 4.8: Cumulative number of failure events by equipment in section 6

The plot in Figure 4.8 depicts the total number of failure events in Section 6 according to the related equipment(only equipment with more than 20 recorded stoppages). As shown in the plot, cement mill 4 (equipment 06CM41) is responsible for approximately 400 of the stoppages in the historical record, roughly twice as many as the next equipment.Furthermore, cement mills 5 and 1(06CM51 and 06CM11) are each responsible for over 100 failure events. As such, section 6 contains an additional cement mill, cement mill 3(06CM31), responsible for over 50 stoppages. Due to the large number of failure events(688) and being responsible for the core function of section 6(cement grinding), the cement mills are identified as the most critical equipment.

4.2.3 Event clustering(All stoppage events)

Despite the occurrence of failure events being of primary interest to maintenance engineers, as identified previously, they are far outnumbered by non-failure stoppage events from the procedural and circumstantial categories. As such, it is important to understand their relevance in understanding and modelling the occurrence of failure events.

For the purpose of looking closer at non-failure events, we compare the mean stoppage duration for each category of stoppages across all plant sections, as shown in Figure 4.9.

These plots indicate that on average *SD* stoppages, which correspond to planned shutdowns, last much longer than any other category. In contrast with the frequency plots in Figure 4.7, which identified process(PROC) as the most frequent stoppage category for almost every plant section, the mean stoppage duration plots indicate that procedural stoppages have short durations, on average. Similarly, circumstantial(SC) stoppage events, which have the second highest frequency of occurrence, have extremely short durations in all plant sections.

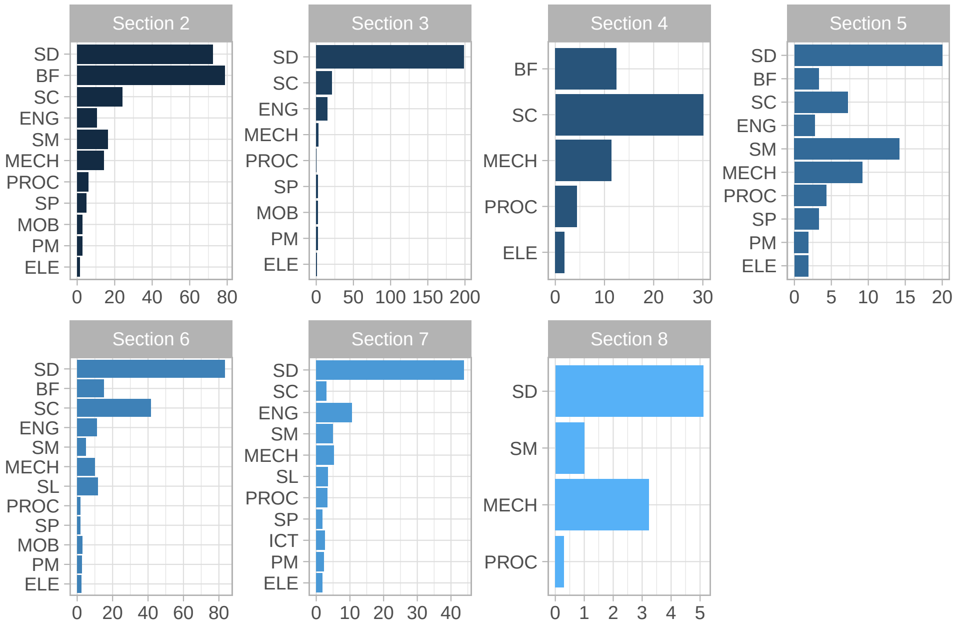

Figure 4.9: Mean stoppage duration in hours of each category, per plant section number

As all sections of the plant experience frequent and repeated stoppages of relatively short duration, it is logical to presume that these events may have a negative impact on related mechanical equipment. As mechanical equipment is prone to natural wear from normal operating conditions, it may also experience additional wear from repeatedly stopping and restarting. As the amount of wear on a mechanical equipment accumulates, the likelihood of experiencing an imminent failure increases. As such, it may be important to identify clusters, or patterns, of events that may indicate a relationship between failure and non-failure related events.

As described by Lawless (2007) when studying recurrent events of multiple states, visually identifying clusters of time-ordered events may indicate a relationship between the frequency, order, or time between events. Despite identifying the cement mills in section 6 as the most critical equipment items in terms of failure events, the total number of events makes visual representation difficult, as points will be continuously overlapping. Section 3(raw mill) will be used as an example of the type of clustering that can occur between failure and non-failure events, as the number of events is frequent enough to be observed without hindering visual observation.

As shown in Table 4.9 the raw mill has experienced 117 mechanical failures and 64 maintenance interventions(ENG and PM), in addition to 161 procedural and 205 circumstantial stoppages. As an increase in wear is more likely to have an impact on mechanical failures as opposed to electrical ones, this failure type will be the focus.

Table 4.9: Number of stoppage events per category for plant section 3(Raw Mill)

ELE	ENG	MECH	PM	PROC	SC	SD
42	62	117	2	161	205	5

Figure 4.10 contains three plots which display event occurrences, according to their time ordering, against a time scale, representing time(in days) since the first failure. In order to clearly visualize the events, each plot only shows events which occurred between 500 and 700 days since the first failure. For the top two plots, a cluster of black triangles represents a period of time in which the raw mill was repeatedly stopped and restarted for non-failure reasons, potentially causing an accumulation of wear. It is of particular interest to identify whether the presence and frequency of these non-failure stoppages has an effect of the interfailure durations(the time between red circles) of mechanical equipment, essentially causing failures to occur sooner. Conversely, events of planned maintenance,identified in the bottom plot, would theoretically have a positive impact on interfailure durations, thereby prolonging the time until the next failure event occurs. The chosen method for accounting for these non-failure events will be detailed in the section about data integration.

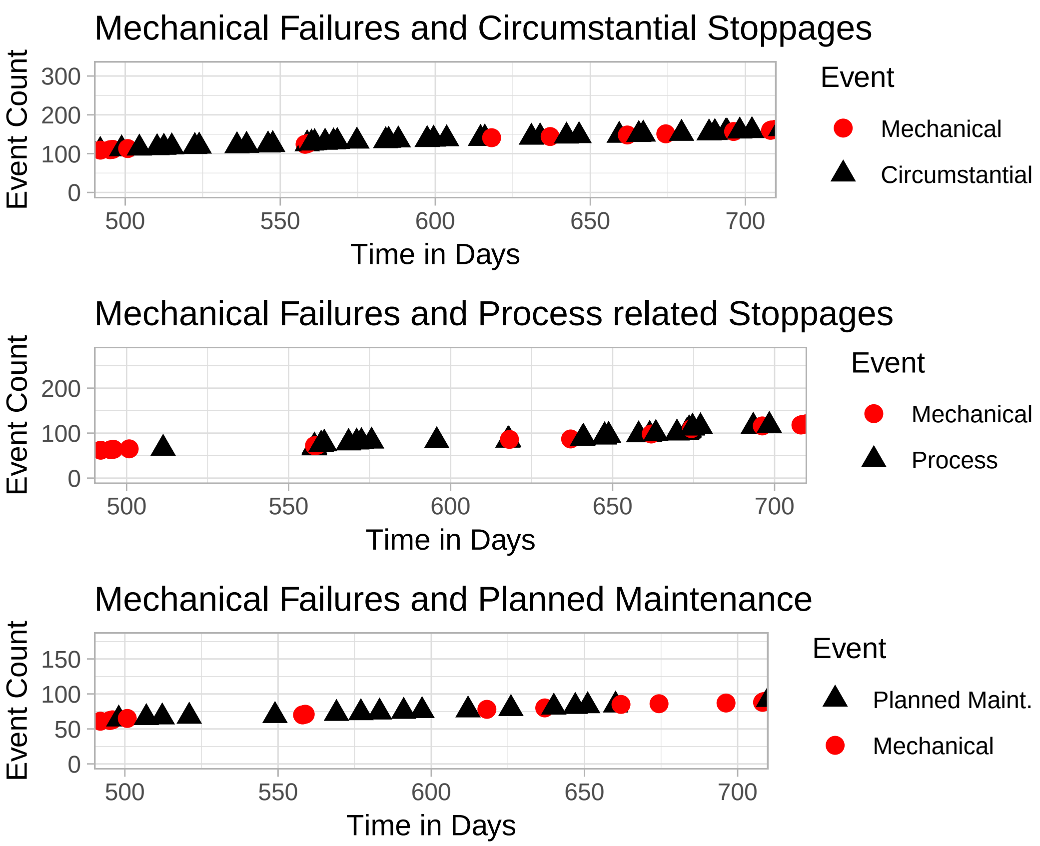

Figure 4.10: Plot of cumulative number of events against time since first event

4.2.4 Vibration measurement data

After collecting and cleaning the vibration measurements into a single data frame, a descriptive analysis was performed to provide insight regarding the scope of information the data contains. In total, the vibration data consists of 1634 observations made on different equipment throughout the plant. Figure 4.11 contains a comparison of boxplots for each of the 10 available vibration measurements, with the colored box representing the interquartile range(IQR), or 1st and 3rd quantiles, in addition to the median value. All variables are measured in units of mm/s except for *HDE.ENV* and *HNDE.ENV* which are measured in units of gE, which represents enveloping acceleration. The enveloping observations measure vibrations in a higher frequency caused by rolling element bearing or gear mesh problems.

As evident from the plots, the median values are all relatively small, indicating that machines typically operate with minimal vibration. Both plots show frequent outlying observations, with several extreme outliers(200-500mm/s range) observed which are not shown. Given the units of measurement, and the fact that these observations were recorded via a handheld probe, it is more likely that the most extreme outliers are the result of observational error, and not correct values.

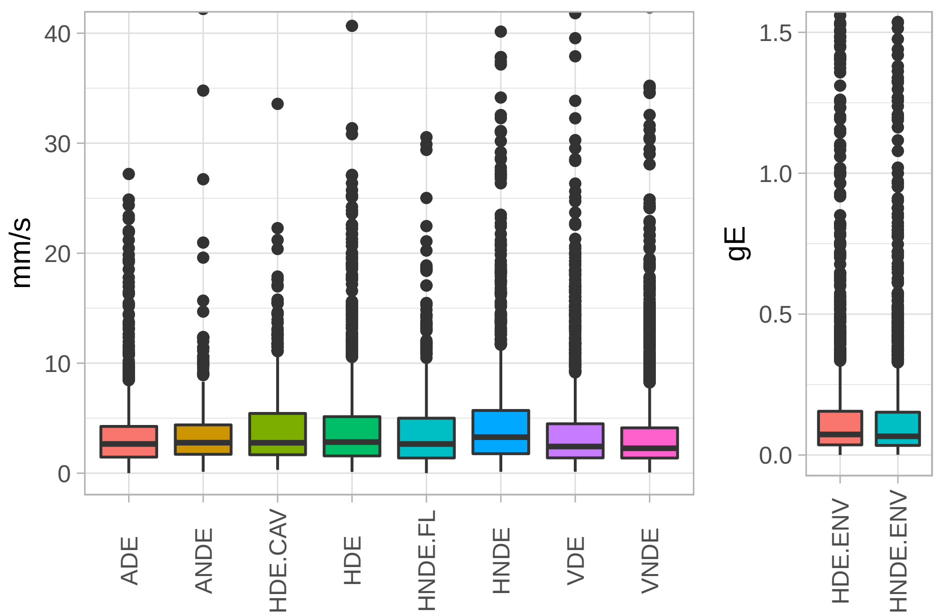

Figure 4.11: Boxplot of each respective vibration measurement

As the individual vibration measurements are observed and recorded, the probe assigns a health status of *Normal*, *Alert*, or *Danger*, which is also recorded. Figure 4.12 contains a breakdown of the range of recorded observations, per variable, according to the respective assigned status. The variable HDE.ENV is not included as all observed values were labeled a normal status.

As illustrated in the plots, the range of values corresponding to normal observations is the largest for each respective variable. Conversely, the range of values corresponding to alert status observations is significantly smaller, with values close to 5 for all variables. Additionally, the range of values for danger status observations is consistently above the IQR for alert status observations, while still occasionally having lower values than some normal observations.

This indicates that the alert status may be assigned by the probe in a proprietary manner accounting for more than just the observed value, or that there may be error or inconsistency in observation status assignment. At any rate, this insight provides context for what magnitude of vibration would be considered dangerous to the operational ability of equipment. Table 5.3 in the appendix contains a summary of the quantiles for each vibration variable and status combination.

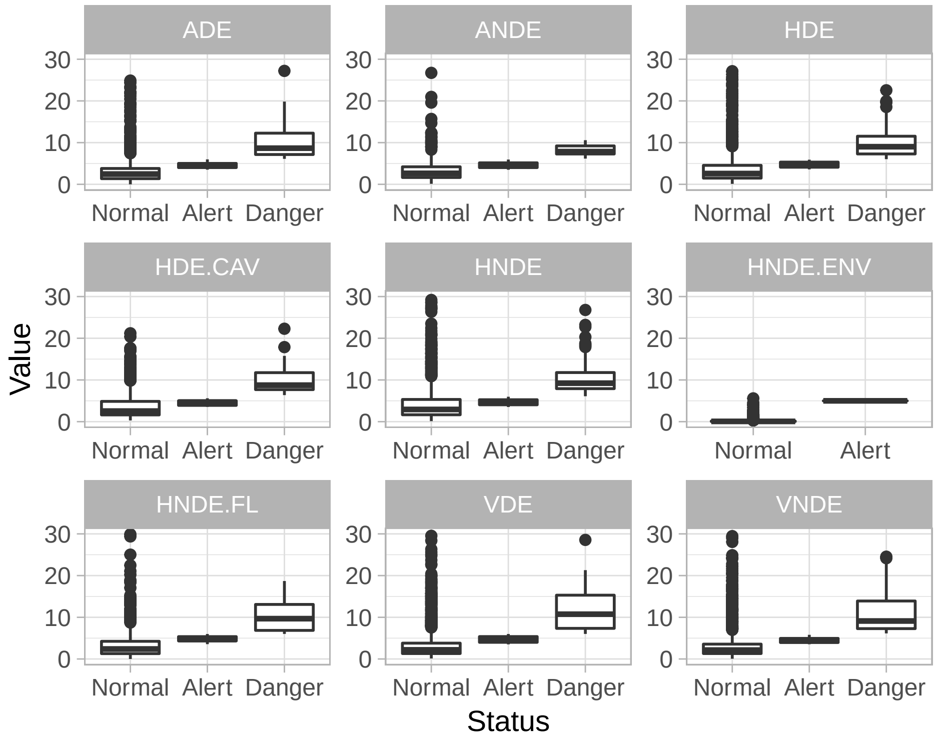

Figure 4.12: Summary of values for each health status level

As vibration measurements are neither continuously observed nor observed at regular intervals, the number of observations per equipment item is highly varied. Additionally, depending upon the equipment, vibrations may be observed at different locations on the equipment, identified by component(i.e., fan, motor, bearing, etc.). As such, in order to use these historical records to describe the health and event patterns for equipment items, it is important to understand the the quantity of available information for relevant plant sections and equipment.

Figure 4.13 provides a breakdown of the number of vibration readings by equipment and section, denoting the proportion from respective component locations. Since some equipment have relatively few vibration observations, only the top 6 equipment per section are shown in the figure.

Based on the plots, it is evident that the equipment in section 5 is responsible for the largest sample of vibration observations, with all of these equipment items being fans(xFNx equipment code). Additionally, fans appear to be some of the most frequently monitored equipment in the remaining sections of the plant. Based on the largest representative sample size, fans will be of primary interest in later condition-based models.

Figure 4.13: Number of vibration observations per equipment per component by plant section(only 6 most frequent equipment per section shown).

Despite the vibration measurements being observed at irregular intervals, observing the changes in vibrations over time may describe changes in the "health" of equipment over time, which would otherwise be unknown without such an indicator.

Figure 4.14 contains a time series plot of the available vibration measurements for fan 2 in section 5(05FN02), which tracks the progression of values throughout time. The bottom plot shown a time series of the 8 vibration variables measured in mm/s along with two horizontal dashed lines at values of 3.5 and 6, representing approximate minimum values for being classified as *alert* or *danger* by the recording probe.

As evident from the plot, during periods where observations are more frequent, the vibration levels, and health of the equipment, has a large amount variation of variation over time. The two enveloping variables in the top plot also display distinct variation, despite all measurements of these variables for fan 2 being classified as *normal*.

As shown in Table 5.3 in the appendix, the minimum observed *alert* value for the HNDE.ENV variable was 6.09gE. The near-horizontal lines between December, 2016 and September, 2017 represent a period of time in which no vibration readings were recorded for fan 2.

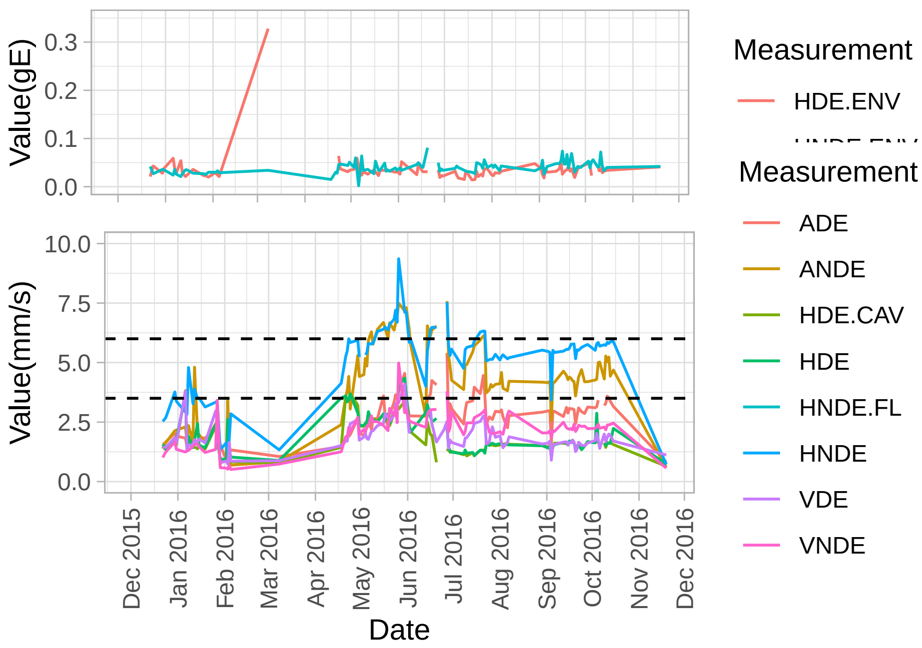

Figure 4.14: Time series ofthe vibration measurements for 05FN02

Several fans throughout the plant were identified as having been monitored via vibration measurements, and have also experienced failure events, thus having near-complete records in all data sources. These fans are referenced by equipment ID “03FN08”, “05FN01”, “05FN02”, “05FN31”, “05FN33”, “06FN11”, “06FN41”, “06FN51”, “06FN52”, “06FN53”, and “07FN01”. When building predictive models for reliability using using fan data, this subset will be used unless otherwise specified. Unfortunately, as the variable *HNDE.FL* will not be included in this subset as it is not measured on fans, but on motors.

Although each vibration variable represents velocity or acceleration in a specific direction, measured at a specific location on the equipment, different parts of the equipment do not vibrate in isolation. As shown in the time series in Figure 4.14, each of the observed variables tend to follow the same patterns in time, implying that as the equipment vibrates more, the increase is is typically reflected in all variables. However, the sparsity of observations in the time series may not adequately reflect the relationship between variables, making them appear far more similar than they are.

Figure 4.15 contains a correlation matrix for all available vibration variables, with values close to 1 indicating high positive correlation, values close to 0 indicating no correlation, and values represented by “x” indicate values that are not observed together. As shown in the figure, the variables with the highest correlation of 0.79 are ADE and ANDE, which measure vibrations in the axial direction on the drive-end(ADE) and non-drive-end(ANDE) of an equipment. Furthermore, axial drive-end(ADE) and horizontal drive-end(HDE), as well as ANDE and vertical drive-end(VDE), have strong correlations of

0.63.

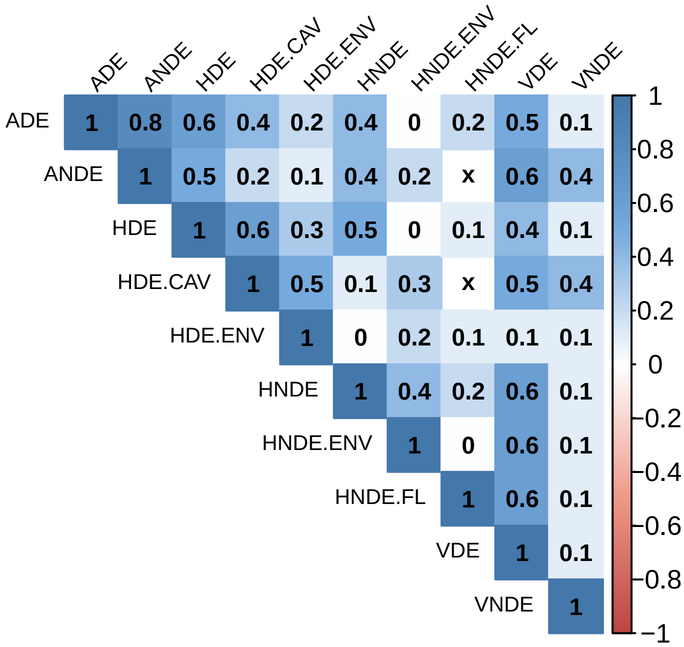

Figure 4.15: Correlation plot of each vibration measurement computed using all observations

Principal component analysis on the vibration measurements

Based on the correlation between most of the vibration measurements and the larger number of predictors, principal component analysis was performed in order to offer a reduced set of uncorrelated covariates for future condition-based reliability models using the original observations. For the purpose of this PCA, only the data from the subset of fans is used, as the results of the PCA will be used to model fan reliability.

As described in Rencher (2002), a cutoff of 80% for the cumulative amount of explained variance was used to justify retaining the first 4 principal components. Additionally, as these components will be used in later models to predict reliability, it is advantageous to retain too many components now, as excess components can simply be excluded during model comparison. Table 5.4 in the appendix contains the component loadings for the first 5 principal components using 9 of the original vibration variables.

A biplot of the PCA results, as shown in Figure 5.2 of the appendix, revealed a strong influence of the outlying values previously described. In response, values above 50mm/s for ADE, ANDE, HDE, HNDE, and VDE were reduced to the mean values for each variable. After reducing the outlying values, a second PCA was performed in order to provide additional predictors for future modeling. Similar to the first PCA, the second technique identified 4 components to explain more than 80% of the variation in the original variables. Table 5.5 in the appendix contains the component loadings for the first 4 principal components after reducing the outlying values. A biplot of the results from the second PCA is shown in Figure 4.16

Based on the loadings for the second PCA, the first principal component largely represents *HDE.CAV* and *HDE.ENV*, which are both horizontal drive-end vibrations, as their loadings have the largest absolute values. Component 2 is dominated by *ANDE* and *HNDE*, which are both non-drive-end measurements. Component 3 has the largest correlations with *HDE* and *ADE*, which are both drive-end measurements. Additionally, component 4 is correlated highest with *VDE* and *ANDE*, both of which are drive-end measurements. The relationships between the first two principal components and the original observations are best illustrated in the respective bigram.

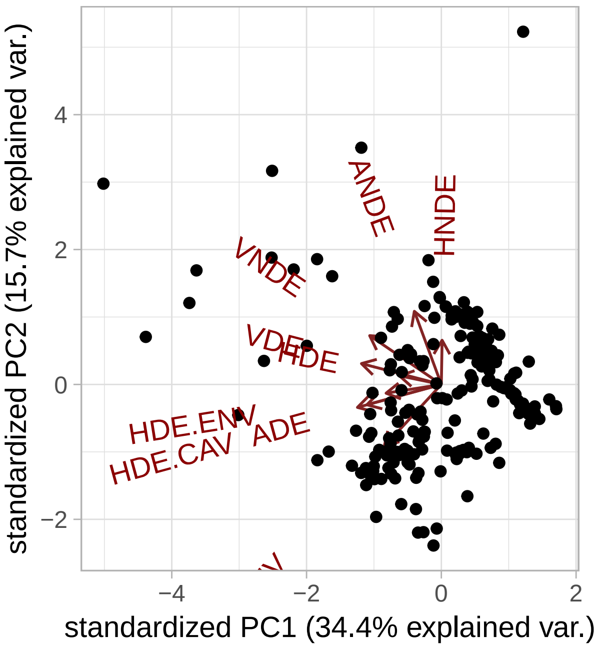

Figure 4.16: Biplot of the fan vibration secondary PCA

4.2.5 Monthly production data

In addition to historical event records, the plant also maintains a record of total monthly production output(measured in tons) for 9 subsections of the plant. The figures are reported by subsection as some plant sections contain multiple subsections which may be operating in parallel(e.g., cement mills) or processing different materials(e.g., coal mill and kiln). Section 5, for example, is made up of 4 cement mills operating in parallel subsections, with total material output being recorded for each.

As the production records were recorded during the same 3-year period as the stoppage event records, this information will provide an important indication of the production performance by vital equipment leading to failure events.

Figure 4.17 contains a comparison of the monthly relative production for each of the four cement mills, over the 3-year period. As evident in the plot, cement mills 4 and 5 maintain an average monthly production rate that fluctuates around 50% for the majority of the monitoring duration. In contrast, cement mills 1 and 3 operate at considerably lower average monthly production rates, and even experience periods of complete shutdown(these shutdowns are indicated in the stoppage event records). This information corroborates with the earlier conclusion from the descriptive analysis of the stoppage event records, that cement mills 4 and 5 experience more failure events than the remaining cement mills, which may be attributed to the fact that they are used most often.

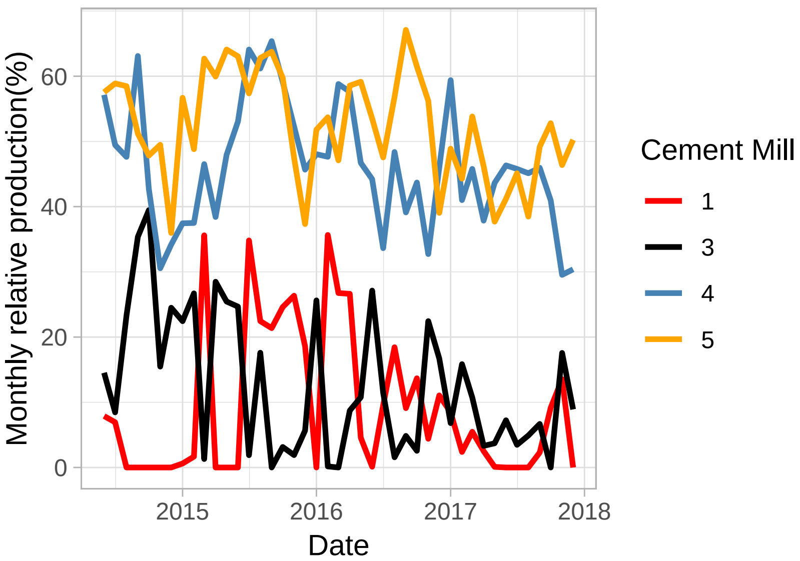

Figure 4.17: Historical monthly relative production for cement mills in section 6

4.3 Data integration

Following the pre-processing and descriptive analysis of all data sources there is a better understanding of what information each data set contains. Using the criticality analysis to compare areas of the plant based on the frequency of failure events has identified the cement mills located in section 6 as the most critical. In addition to experiencing the highest number of failure events, the cement mills also experience a large number of non-failure stoppages due to raw material or other resource shortages. Although the cement mills themselves were not monitored for vibrations, the respective stoppage event and production records will be integrated.

The descriptive analysis of the vibration data has concluded that fans are the most frequently monitored equipment throughout the plant. As the fans are responsible for performing temperature regulation for large equipment throughout the plant, such as the kiln in section 5, the condition of the fans is critical to the functionality of each section.

As both the cement mills and fans have been identified as critical equipment throughout the plant, the following section will detail the process of data integration for each.

4.3.1 Integrated cement mill data

In its prepared form, each entry in the stoppage record data represents a stoppage event, denoting the category of stoppage, beginning and ending timestamp, equipment, stoppage duration(repair time), and respective failure mechanism and maintenance action(if failure event).

In order to prepare the data for integration, the event records representing *failure durations* should be turned into records representing *interfailure durations*, or *gap times*. Whereas the original data identified when the equipment failed, and the duration of downtime, the integrated data must represent how long the equipment was operational for, prior to failure, and what occurred to the equipment during this time. The following steps were performed to integrate the data sets for the cement mills in section 6.

1. Select only the *failure* events for equipment “06CM11”, “06CM31”, “06CM41”, and “06CM51”.
2. Using these failure events, convert the failure records into *gap times*. As the first gap time corresponds to the time between the first and second failures, the sample size for gap times is two records smaller than the sample size of failure events. Since the failure mechanism, maintenance action, and repair time have been extracted earlier, they are retained such that each value represents what occurred prior to the gap time. The gap times now represent all available information regarding failure events, with each row corresponding to a single gap time.
3. To incorporate the production records into the gap times, the beginning and end of each respective gap time is referenced with the production data. If the gap time occurs within a single month, the corresponding monthly production rate is assigned to the gap time. If the gap time spans multiple months, a data row is inserted for each month for which production is recorded. As multiple rows now represent a single gap time, each data row will be identified as a *segment* for which a new variable, *stat*, is used to identify whether the segment ended with failure of the equipment. In this format, *stat* is set to 0 for all but the final segment, indicating that the equipment was continuously operational until the final segment. The production rate for each segment is considered a time-dependent covariate, as the segments are used to track the changes in production while the equipment is approaching the next failure.
4. A similar step is used to incorporate information regarding the non-failure events that occur within the gap times for each respective equipment. For each segment, the beginning and end timestamps can be referenced against the non-failure events for each equipment. In order to account for maintenance interventions(planned maintenance not initiated by failures) that occur during the gap time, a new variable *maint.* is introduced, which keeps a running total of these events. Additionally, a variable *numstop* is used to keep a running total of the stoppages that are neither failure-related nor maintenance interventions. Since both of these new variables are also time-dependent variables, new time segments are inserted when either of these values changes. When iterating through the consecutive gap times, each row corresponds to a segment of time in which only one of the time-dependent covariates change, while all other variables remain constant.

Although the integration procedure considerably increases the complexity of the data, it provides a framework for accounting for maximal information regarding the behavior of an equipment between failure events. Table 4.10 provides a sample of the first interfailure duration for cement mill 1 after integration of the data. Due to the size of the dataframe, the additional covariate values are not shown, but a description of all covariates is provided in Table 4.11

Table 4.10: Sample of the intregrated cement mill data for the first interfailure duration for cement mill 1

id	seg	stat	start	stop	tstart	tstop	equipment
1	1	0	2015-01-27 04:20:00	2015-01-30 06:00:00	0.000000	3.069444	06CM11
1	2	0	2015-01-30 06:00:00	2015-01-31 23:59:59	3.069444	4.819433	06CM11
1	3	0	2015-01-31 23:59:59	2015-02-27 08:10:00	4.819433	31.159722	06CM11
1	4	0	2015-02-27 08:10:00	2015-02-28 23:59:59	31.159722	32.819433	06CM11
1	5	0	2015-02-28 23:59:59	2015-03-01 13:40:00	32.819433	33.388889	06CM11
1	6	0	2015-03-01 13:40:00	2015-03-01 14:47:00	33.388889	33.435417	06CM11
1	7	0	2015-03-01 14:47:00	2015-03-01 22:30:00	33.435417	33.756944	06CM11
1	8	0	2015-03-01 22:30:00	2015-03-02 05:15:00	33.756944	34.038194	06CM11
1	9	1	2015-03-02 05:15:00	2015-03-02 16:30:00	34.038194	34.506944	06CM11

As shown in Table 4.10, the *id* variable indicates all the rows that correspond to the first interfailure duration. The interfailure durations are numbered only to facilitate estimation, and are not treated as ordered. The *seg* variable indicates the segment, or time interval, which is a further subdivision of the gap time. As this interfailure duration contains 9 segments, it indicates that there have been 8 changes in the time dependent variables for cement mill 1 during this gap time.

As evident in Table 4.10, the *stat* variable indicates when the equipment finally fails, and is necessary for model estimation. The *start* and *stop* variables keep track of the segment time via timestamp, and *tstart* and *tstop* track the progression of each segment by counting the days since the start of the interfailure duration.

A description of each covariate is provided in Table 4.11, which is classified into *identifiers*, *time-independent* covariates, and *time-dependent covariates*. The identifiers are used to keep track of the time, equipment, and equipment status. The time-independent covariates do not change throughout the interfailure duration, and represent what occurred to the equipment immediately prior to becoming operational. The time-dependent covariates represent the incoming information about what is occurring to the cement mills, and is accounted for based on the time it occurs.

Table 4.11: List of covariates and descriptions for integrated cement mill data

Variable Type	Variable	Description
Identifiers	id	Identifier for each interfailure duration
	seg	Identifier for time-dependent segment of duration
	stat	Status of whether the time segment ended in failure
	start	Timestamp for start of each segment
	stop	Timestamp for end of each segment
	tstart	Time(in days) since start of first segment
	tstop	Time(in days) since start of first segment
	equipment	Equipment being observed
Time-independent	ma	Previous maintenance action performed
	fm	Failure mechanism of previous failure event
	ma	Maintenance action of previous failure event
	reptime	Repair time(downtime) for previous failure event
Time-dependent	prod	Relative production rate for associated equipment
	numstop	Cumulative count of non-failure stoppages
	maint	Cumulative count of planned maintenance actions

4.3.2 Integrated fan data

Since the fans are also monitored for vibrations, the integration of data sources follows the same procedure, but with an additional step to insert the vibration measurements as time-dependent covariates, as indicated below.

1. Select only the *failure* events for equipment “03FN08”, “05FN01”, “05FN02”, “05FN31”, “05FN33”, “06FN11”, “06FN41”, “06FN51”, “06FN52”, and “06FN53”.
2. *The same step(2) is repeated as for cement mills*
3. *The same step(3) is repeated as for cement mills*
4. *The same step(4) is repeated as for cement mills*
5. A similar step is performed to integrate the vibrations measurements into the data frame. For each segment, the vibration data is queried to retrieve observations for the respective equipment. For each observation, a new segment is inserted into the interfailure duration, keeping the remaining covariates unchanged. For the first segment, representing the time interval in which the equipment just became operational, the vibration level is unknown, but assumed to be at an acceptable level. As such the values for the first time segment are imputed under this assumption.

The integration of the data sources for fans follows a similar procedure, and provides a similar framework for accounting for the changes the fans experience between failure events. Table 4.12 provides a sample of the first interfailure duration for section 3, fan 8 after integration of the data. The additional covariate values are not shown, but a description of all covariates is provided in Table 4.13.

Table 4.12: Sample of the intregrated fan data for the first interfailure duration for section 3 fan 8

id	seg	stat	start	stop	tstart	tstop	equipment
1	1	0	2016-10-30 00:28:00	2016-10-31 23:59:59	0.000000	1.980544	03FN08
1	2	0	2016-10-31 23:59:59	2016-11-30 23:59:59	1.980544	31.980544	03FN08
1	3	0	2016-11-30 23:59:59	2016-12-31 23:59:59	31.980544	62.980544	03FN08
1	4	0	2016-12-31 23:59:59	2017-01-31 23:59:59	62.980544	93.980544	03FN08
1	5	0	2017-01-31 23:59:59	2017-02-28 23:59:59	93.980544	121.980544	03FN08
1	6	0	2017-02-28 23:59:59	2017-03-31 23:59:59	121.980544	152.980544	03FN08
1	7	0	2017-03-31 23:59:59	2017-04-30 23:59:59	152.980544	182.980544	03FN08
1	8	0	2017-04-30 23:59:59	2017-05-31 23:59:59	182.980544	213.980544	03FN08
1	9	0	2017-05-31 23:59:59	2017-06-30 23:59:59	213.980544	243.980544	03FN08
1	10	0	2017-06-30 23:59:59	2017-07-14 07:22:00	243.980544	257.287500	03FN08
1	11	0	2017-07-14 07:22:00	2017-07-26 12:05:00	257.287500	269.484028	03FN08
1	12	0	2017-07-26 12:05:00	2017-07-27 03:28:00	269.484028	270.125000	03FN08
1	13	0	2017-07-27 03:28:00	2017-07-31 23:59:59	270.125000	274.980544	03FN08
1	14	0	2017-07-31 23:59:59	2017-08-31 23:59:59	274.980544	305.980544	03FN08
1	15	1	2017-08-31 23:59:59	2017-09-17 02:22:00	305.980544	322.079167	03FN08

Table 4.13: List of covariates and descriptions for integrated fan data

Variable Type	Variable	Description
Identifiers	id	Identifier for each interfailure duration
	seg	Identifier for time-dependent segment of duration
	stat	Status of whether the time segment ended in failure
	start	Timestamp for start of each segment
	stop	Timestamp for end of each segment
	tstart	Time(in days) since start of first segment
	tstop	Time(in days) since start of first segment
	equipment	Equipment being observed
Time-independent	ma	Previous maintenance action performed
	fm	Failure mechanism of previous failure event
	ma	Maintenance action of previous failure event
	reptime	Repair time(downtime) for previous failure event
Time-dependent	prod	Relative production rate for associated equipment
	numstop	Cumulative count of non-failure stoppages
	maint	Cumulative count of planned maintenance actions
	ADE	Axial drive end
	ANDE	Axial non-drive end
	HDE	Horizontal drive end
	HDE.CAV	Horizontal drive end cavitation
	HDE.ENV	Horizontal drive end enveloping
	HNDE	Horizontal non-drive end
	HNDE.ENV	Horizontal non-drive end enveloping
	HNDE.FL	Horizontal non-drive end
	VDE	Vertical drive end
	VNDE	Vertical non-drive end
	comp1	First principal component score from primary PCA
	comp2	Second principal component score from primary PCA
	comp3	Third principal component score from primary PCA
	comp4	Fourth principal component score from primary PCA
	compb1	First principal component score from secondary PCA
	compb2	Second principal component score from secondary PCA
	compb3	Third principal component score from secondary PCA
	compb4	Fourth principal component score from secondary PCA

As evident by Table 4.12, the integrated fan data follows an identical structure as for cement mills integrated data set, apart from the addition of more covariates. However, the major difference between the two is that because each vibration measurement corresponds to a new observation of several different variables(e.g., ADE, ANDE, etc.) at the same point in time, all vibration variables change at the same time. In the case of the cement mills, for each additional segment in a gap time, a maximum of 1 variable has been updated, but for the fans, a maximum of 17 variables(9 original measurements, 8 principal component scores) have been updated.

A description of each covariate is provided in Table 4.13, which is also classified into *identifiers*, *time-independent* covariates, and *time-dependent covariates*. The meaning of these covariates is consistent between data sources, with the exception of including additional time-dependent covariates to incorporate the change in vibrations being observed.

4.4 Basic reliability models

4.4.1 Basic cement mill reliability

In the following analysis, a subset of failure events for cement mills from plant section 6 will be used. Considering this data, the random variables is the time between failures, which will be referred to as *interfailure* duration.

As illustrated in Table 4.14, which contains descriptive statistics for the current subset of interfailure times, the sample mean for cement mill 4 is more than 3 times the median interfailure time. Furthermore, the maximum interfailure time is more than 10 times the sample mean value, and the sample mean is very close to the 3rd quantile value. These statistics indicate that the sample distribution is highly skewed to the right, suggesting that extremely large interfailure durations do exist, yet are far less frequent.

Table 4.14: Descriptive statistics for interfailure times(days) for section 6 cement mills

Mill	Min.	Q1	Median	Mean	Q3	Max.
ALL	0.001	0.125	0.948	5.971	4.805	223.757
1	0.003	0.141	0.967	8.849	4.454	223.757
3	0.007	0.188	4.309	19.051	12.415	171.381
4	0.001	0.086	0.768	2.673	3.146	33.939
5	0.007	0.473	2.120	8.518	8.785	116.829

4.4.1.1 Trend assessment

In the study of recurrent events, there are several common methods for identifying the presence of a *trend*, which can refer to several different aspects.

As suggested by Lindqvist (2006), the plot in Figure 4.18, represents a trend chart of the cumulative number of failures against the time since the first failure for cement mill 4. The cumulative number of failures is simply an ordering of the approximately 400 failure events that have occurred for cement mill 4 in the historical records. The failure rate is visualized against a time scale representing the approximately 1,100 days over which the cement mill was monitored.

This type of plot is one method for assessing the presence of trend failure occurrence rate, either increasing or decreasing. In the literature, this plot is derived from an adaptation of the Nelson-Aelen estimate, but is equivalent to what is shown below. In terms of this plot, a straight diagonal line would indicate no trend. In this instance, apart from a large increase near 300 days since first failure, the line appears relatively straight, indicating no obvious trend in failure rate.

Although this plot suggests that the rate of failure occurrence for cement mill 4 has remained *relatively* constant, the visual representation emphasizes the *rate* of occurrence, making it difficult to compare the actual gap times, or interfailure durations. Comparing the actual observed gap times, as presented in Figure 4.19 provides a complementary assessment.

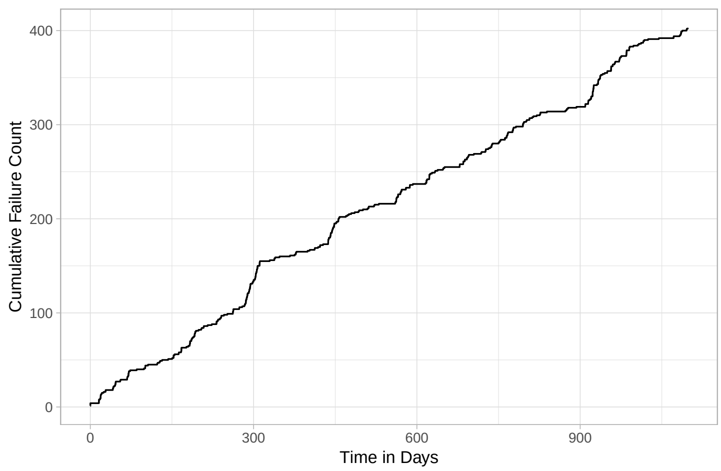

Figure 4.18: Plot of cumulative number of failures against time since first failure for cement mill 4

The plots in Figure 4.19 contain a slightly different representation of failure occurrence for each of the 4 cement mills in plant section 6. Similar to the plot in Figure 4.18, each failure event is identified by the order in which it occurred. However, in these plots, the time scale represents the interfailure duration, as opposed to the time since first failure. As the interfailure duration represents the duration of time in which an equipment was operational, or time between failures, this representation allows for easier comparison of equipment behavior.

In these plots, each point represents an interfailure duration; points on the far left indicate instances where the cement mill failed after a shorter duration, and points on the far right indicate instances when the cement mill was operational for a long time before failure occurred. A negative trend in interfailure duration, indicating possible degradation, would be represented by shrinking interfailure durations. As evident in the plots, neither cement mill appears to have a distinct negative trend. In fact, cement mill 5 experienced most of the exceptionally long interfailure durations near the end of the three-year historical record.

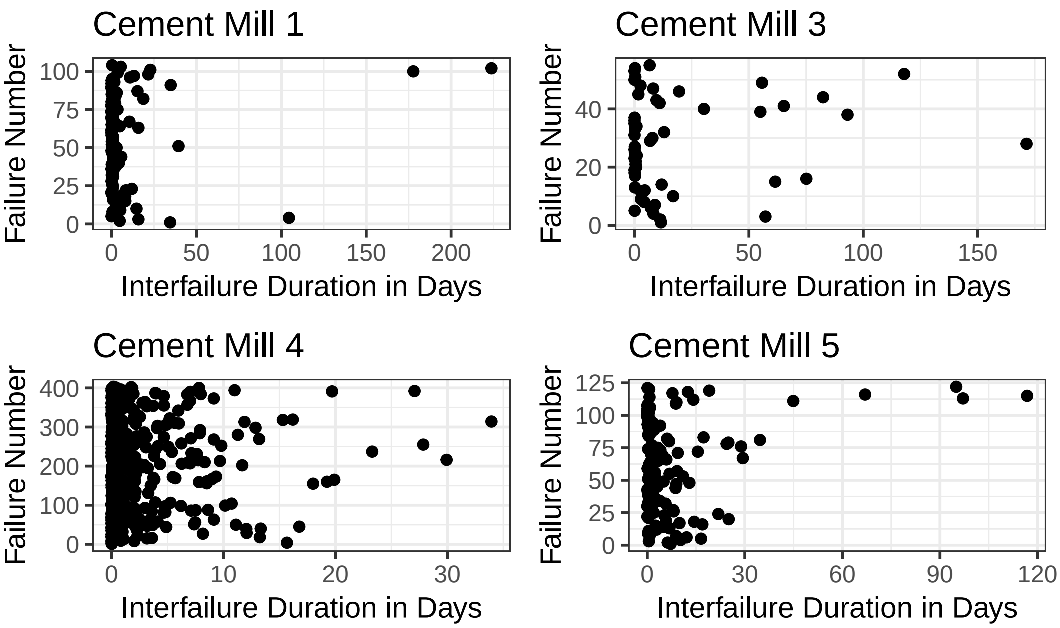

Figure 4.19: Plot of interfailure duration against ordered failure number for each cement mill

In addition to visual inspection, Lawless (2007) recommends using a parametric approach to formally test the presence of trend in failure occurrence. By employing a Cox model such that $h_{ij}(w) = h_0(w) \exp(z_{ij}\beta)$, with z_{ij} representing the jth gap time of the ith cement mill, and a test for $H_0 : \beta = 0$. In this case, the null hypothesis represents no trend, and alternative hypothesis represents a non-zero trend. When performed individually for each cement mill, the Cox model reported no trend, supporting the conclusion based on visual inspection.

Alternatively, Table 4.15 contains respective Pearson correlation coefficients of the interfailure durations and ordered failure numbers for each cement mill. The small values further support the conclusion of no trend, allowing for the usage of standard modelling procedures, with the assumption that failure events occur independently. As no trend was detected, the analysis will continue with traditional survival analysis techniques.

Table 4.15: Pearson correlations coefficients between interfailure duration and ordered failure number

Cement Mills			
1	3	4	5
0.1640296	0.1398092	0.0516574	0.256835

4.4.1.2 Non-parametric reliability estimation

Figure 4.20 shows a comparison of the Kaplan-Meier(KM) survival curves for each of the four cement mills. Each line represents the decreasing probability of a respective cement mill being functional at a particular point in time after having been operational for the specified number of days. The dashed lines represent a comparison of the median survival times for each of the cement mills. The median survival time can be interpreted as the time at which half of the previous failures have occurred.

The KM procedure provides a non-parametric estimation of the survival function $\hat{S}(t)$ for each cement mill using the observed interfailure durations.

Another facet of reliability analysis is the concept of the hazard function, which represents the instantaneous hazard rate. This hazard rate is interpreted as the probability of failure by a future time point, given that the equipment has been operational until the present time(Zacks 2012). Although not provided here, it is possible to derive the hazard function $h(t)$ from the survival function $\hat{S}(t)$, and the PDF $f(t)$, and vice versa.

$$h(t) = \frac{f(t)}{S(t)} \qquad S(t) = \exp\left\{-\int_0^t h(x)dx\right\} \tag{4.2}$$

Figure 4.20 also contains a plot of the *Cumlative Hazard Function* $H(t)$ for each of the respective cement mills. The cumulative hazard for time t is the integral of the hazard function $h(t)$ from 0 to t, and can be interpreted as the number of failures one would expect by time t, if the failure process were repeatable(Cleves et al. 2010).

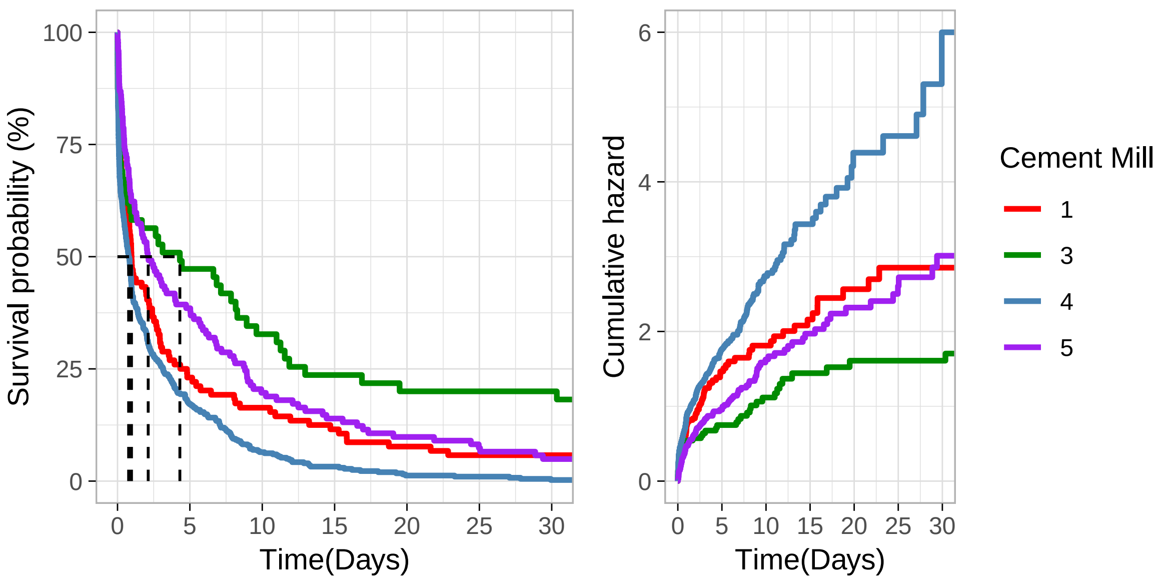

Figure 4.20: Non-parametric comparison of cement mill reliability

Figure 4.20 shows that median survival time is largest for cement mill 3(close to 5 days), and second largest for cement mill 5(close to 2.5 days). The median survival times for cement mills 1 and 4 appear approximately equal from the plotted KM estimates. Additionally, cement mill 3 appears to have the highest survival probability for most time durations.

In addition to visual comparison, the KM estimates for each cement mill may be compared using the log-rank test, which will identify whether the cement mills have differing reliability estimates. A p-value of 0 for the χ^2_3 estimate of the log-rank test suggests that the survival functions of the four cement mills are not identical. Based on this result, we will incorporate the cement mill ID as a categorical covariate in the forthcoming parametric models.

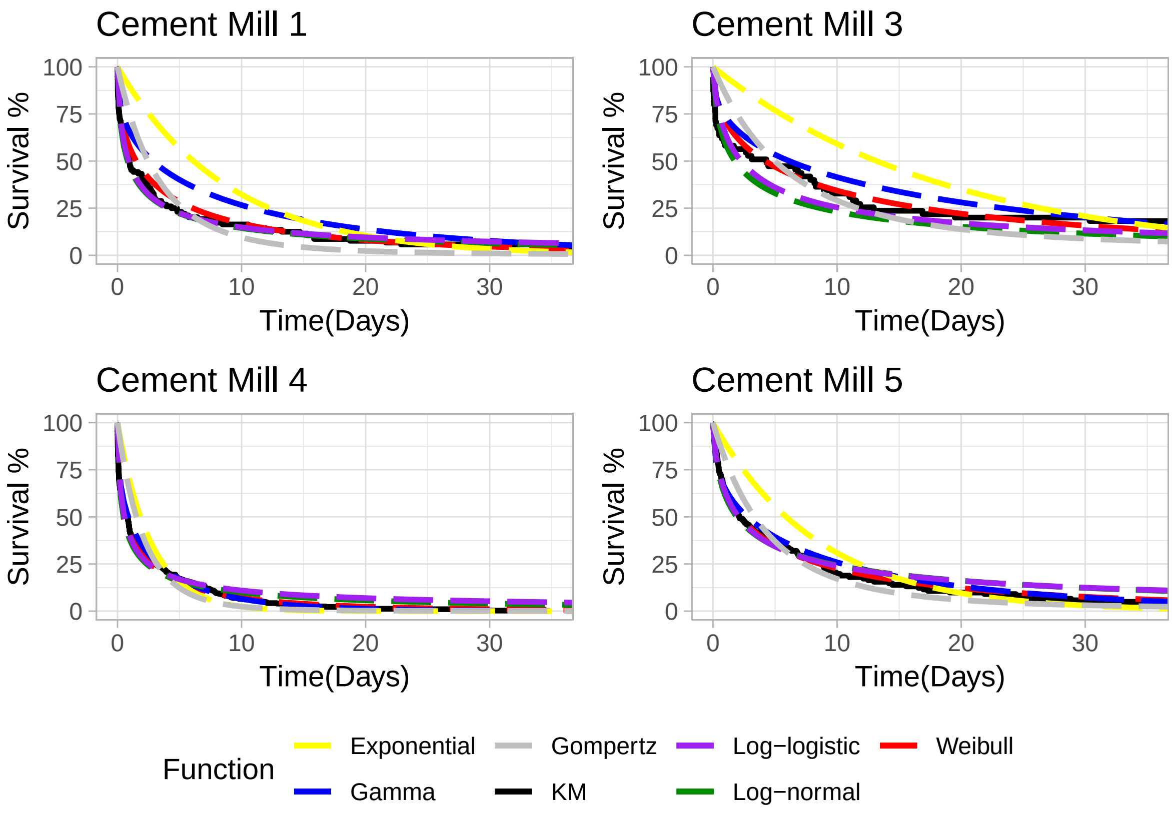

Figure 4.21: Comparison of Kaplan-Meier survival estimate with other model distributions

The plots in Figure 4.21 shows a comparison of the KM survival estimate for each cement mill against fitted Weibull, Log-normal, Gamma, Exponential, Log-logistic, and Gompertz distributions using the *flexsurvreg* function from the *flexsurv* package in R. As shown in the plots, the regression model using a Weibull distribution provides a reasonable approximation of the KM estimate for each of the cement mills.

Furthermore, Table 4.16 contains a comparison of the Akike's Information Criterion(AIC) from each of the respective fitted models. The smallest AIC value of 2681.79 supports the notion that a Weibull model provides the closest approximation of the survival function for each cement mill. As such, the Weibull model will be used for the basis of a parametric model of the reliability characteristics of each cement mill.

Table 4.16: AIC comparison of model distribution options

Weibull	Log-normal	Gamma	Exponential	Log-logistic	Gompertz
2681.794	2712.833	2731.35	3468.762	2740.371	3211.258

4.4.1.3 Parametric reliability estimation

Based on the conclusions that the reliability of the four cement mills is not equivalent but is well approximated by a Weibull distribution, parametric models can be estimated to model the behavior of each cement mill. The Weibull function is extremely popular for modelling the hazard function for reliability analysis, and has unique implications for **Accelerated Failure Time(AFT)** and **Proportional Hazards(PH)** estimation(Kleinbaum, Klein, and Samet 2006).

As outlined in Kleinbaum, Klein, and Samet (2006), the AFT assumption implies that the effect of covariates is multiplicative(proportional) with respect to the *survival time* $S(t)$, whereas the PH assumption implies that the effect of covariates is multiplicative(proportional) with respect to the *hazard function* $h(t)$. When the process is well represented by a Weibull function, each assumption implies the other, however, the distinction between the two formulations determines the interpretation of the estimated effect of the covariates.

The appropriateness of a Weibull fit, as well as the AFT and PH assumptions, can be assessed graphically by plotting the relationship between $log(-log[\hat{S}_c(t_i)])$ and $log(t_i)$, where $\hat{S}_c(t_i)$ is the Kaplan-Meier survival estimate for each cement mill c. Figure 4.22 contains a plot of the complementary log-log transformation for each of the respective cement mills, along with a fitted least squares line to assess linearity, which would suggest that a Weibull hazard function would be appropriate for modelling.

If each of the relationships appear linear, then parallel lines would imply that both PH and AFT assumptions appear valid, given the properties of the Weibull hazard function. In this case, the lines largely appear straight, supporting the use of a Weibull distribution, but clearly intersect indicating that both the PH and AFT assumptions may not be valid.

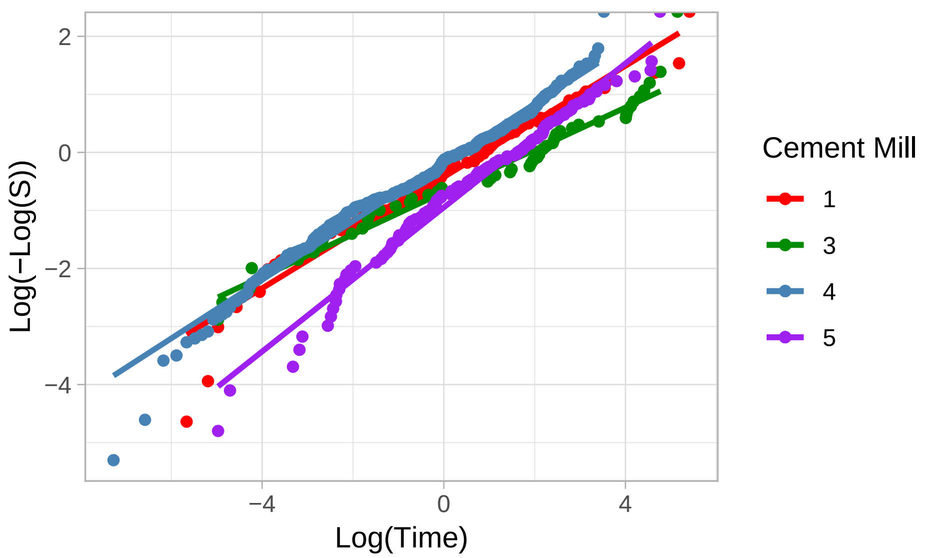

Figure 4.22: Assessment of Weibull fit for each of the cement mills

Assuming that the reliability of each cement mill is not equivalent, but each can be adequately represented by a Weibull function, we can begin to estimate respective model parameters. Using the *survreg* function from the *Survival* package, we can specify a Weibull distribution and obtain parameter estimates using maximum likelihood estimation. Figure 4.23 shows a comparison of the output of an estimated parametric regression model against the KM survival estimate for each cement mill.

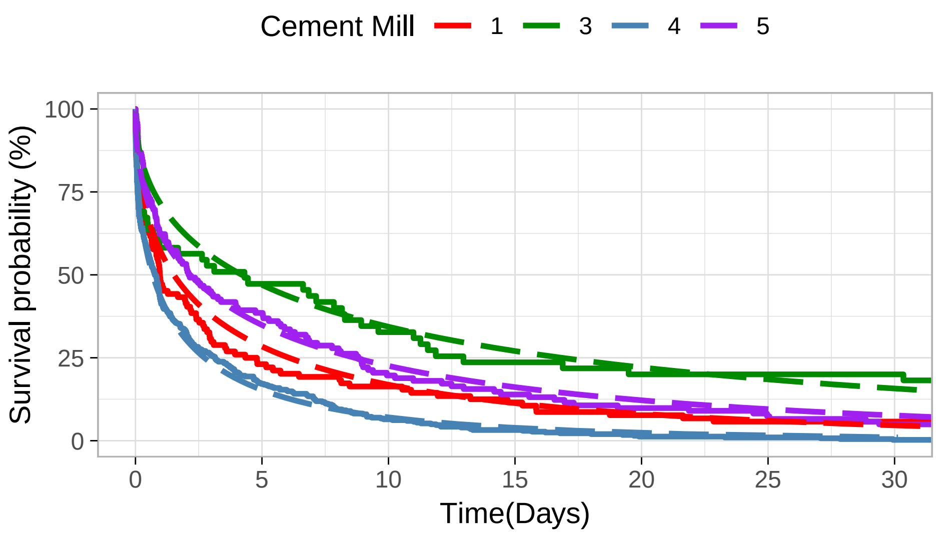

Figure 4.23: Comparison of Kaplan-Meier survival estimate with Weibull model fit

The fitted Weibull model uses the parameterization of α and μ as the shape and scale terms respectively. The shape parameter is held constant, while the cement mill identifier is used as a categorical covariate with a multiplicative effect on the scale parameter. In this parameterization, cement mill 1 is used as the baseline, with covariates x_1, x_2, and x_3 representing dummy variables for cement mills 3, 4, and 5.

$$S(t) = exp(-(\frac{t}{\mu})^{\alpha}), \qquad log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{4.3}$$

Table 4.17 contains the estimates of the intercept β_0 and β_1, β_2, and β_3 coefficients. Individually, each coefficient represents the logarithm of the ratio of survival times of each mill with the baseline mill(mill 1). As such, positive coefficients such as β_1 and β_3 represent longer survival times for mills 3 and 5, compared with mill 1. Conversely, the negative coefficient for β_2 suggests shorter survival times for mill 4.

Table 4.17: Beta coefficient estimates

Coefficients			
β_0	β_1	β_2	β_3
1.149169	1.020933	-0.8008054	0.3558786

Table 4.18 contains the respective shape and scale parameters for each of the cement mills derived from the estimated coefficients. As the cement mill ID is included in the model as a categorical covariate, only one model is being estimated, and thus the shape parameter

Table 4.18: Parameter estimates for each cement mill

	Cement Mills			
Parameter	1	3	4	5
Shape: α	0.4995278	0.4995278	0.4995278	0.4995278
Scale: μ	3.1555706	8.7591758	1.4167478	4.5043695

α remains constant. The dummy variables for each cement mill have a linear effect on the the log of the scale parameter μ, with each respective value reported in the table.

4.4.2 Basic fan reliability

The analysis of fan reliability will follow a similar procedure as for the cement mills in plant section 6, as discussed in the previous section. However, since the plant contains a large number of fans operating throughout different sections of the plant, this analysis will use two subsets of data. As identified in the descriptive analysis of the vibration dataset, the first and more general set of equipment consists of 10 fans for which complete records of stoppage events, vibration measurements, and plant subsection production rate, are recorded. The second dataset is a subset of the first, consisting of only the fans from plant section 5(05FN01, 05FN02, 05FN31, and 05FN31), which will be used for parametric estimation. For each respective part of the analysis, the dataset being used will be noted.

4.4.2.1 Non-parametric reliability estimation

Figure 4.24 provides the respective KM estimates of the reliability function $\hat{S}(t)$ and cumulative hazard function $\hat{H}(t)$ for each of the fans. Each plot has been restricted to show only the first 30 days days of each interfailure time in order to better visualize differences in the KM estimate at earlier durations. Although not visually depicted in the figures, the interfailure durations of the fans are much longer than those previously observed for the cement mills. As the cement mills are large equipment providing the majority of the core function of processing raw materials, and the fans provide supportive functionality through cooling, this difference in interfailure durations is intuitive.

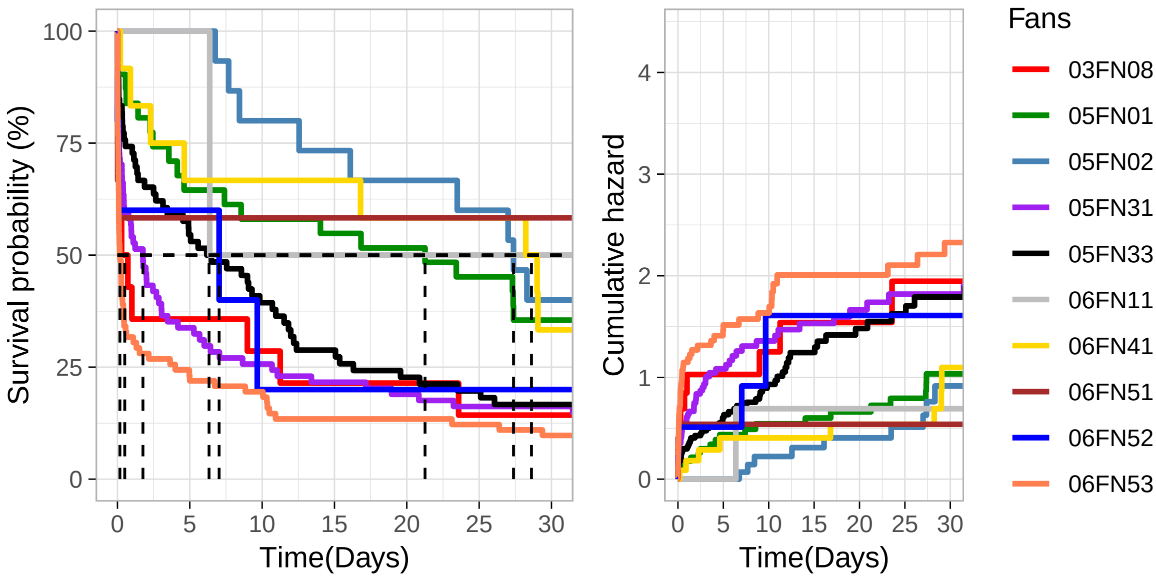

Figure 4.24: Non-parametric comparison of fan reliability

4.4.2.2 Trend assessment

In following with the methodology of the analysis, the subset of fans in plant section 5 was assessed in order to evaluate the possibility of trend in the interfailure durations. Figure 4.25 contains a plot of all failures for each fan, according to their ordered failure number and interfailure duration.

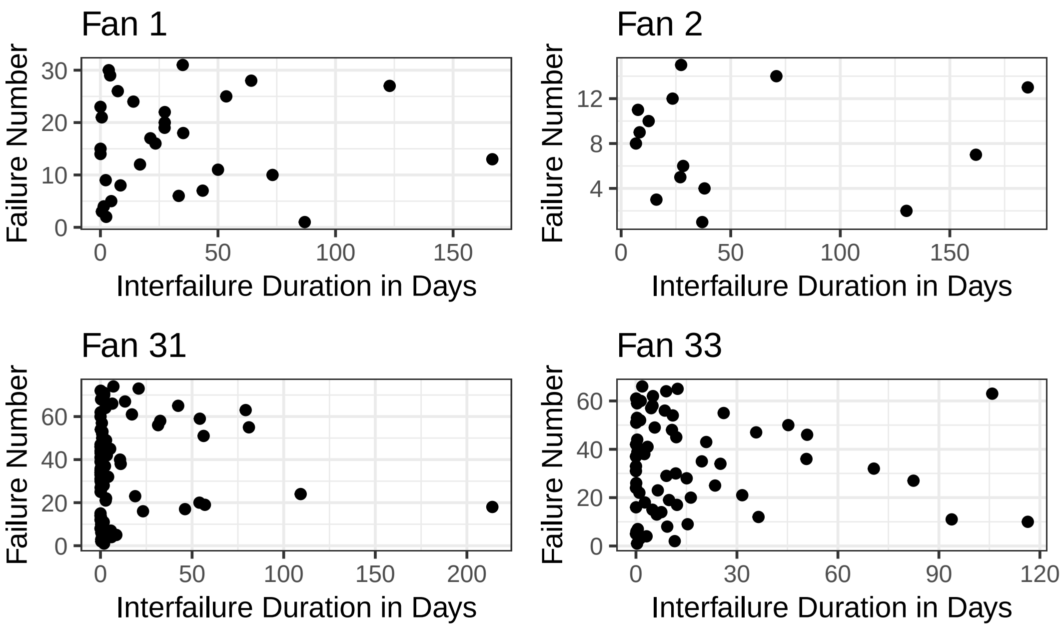

Figure 4.25: Plot of interfailure duration against ordered failure number for each fan in section 5

Similar to the conclusion drawn from the cement mills, there appears to be no indication of a trend in the interfailure durations for any fan in section 5. Furthermore, estimating a Cox model for each fan using the ordered failure number as the only covariate, as suggested by Lawless (2007), further indicated no presence of trend. Furthermore, the small values for the Pearson correlation coefficients provided in Table 4.19 further support this conclusion.

In addition to providing evidence against the presence of trend, the plots in Figure 4.25 provide further context regarding the history of failures for each fan. Based on the highest ordered failure numbers, it is clear than fans 31 and 33 experience far more failure events than fans 1 and 2.

Table 4.19: Pearson correlations coefficients between interfailure duration and ordered failure number

Fan			
1	2	31	33
0.0369481	0.0333637	-0.0033797	-0.0394446

4.4.2.3 Parametric reliability estimation

In further analysis with the subset of fans from plant section 5, the plots in Figure 4.26 show a comparison of the KM survival estimate for each fan against fitted Weibull, Log-normal, Gamma, Exponential, Log-logistic, and Gompertz distributions using the *flexsurvreg* function from the *flexsurv* package in R. As shown in the plots, the regression model using a Weibull distribution provides a reasonable approximation of the KM estimate for for both fans 31 and 33.

In contrast, for fans 1 and 2, the Weibull model appears to underestimate reliability during some short time durations, as compared to the KM estimate. Despite the poor fit of the Weibull model for short durations, it still provides a visually similar approximation for longer interfailure durations. Furthermore, the unique behavior of the KM estimate for fan 2 may be representative of the small number of failure occurrences, as represented in Figure 4.25.

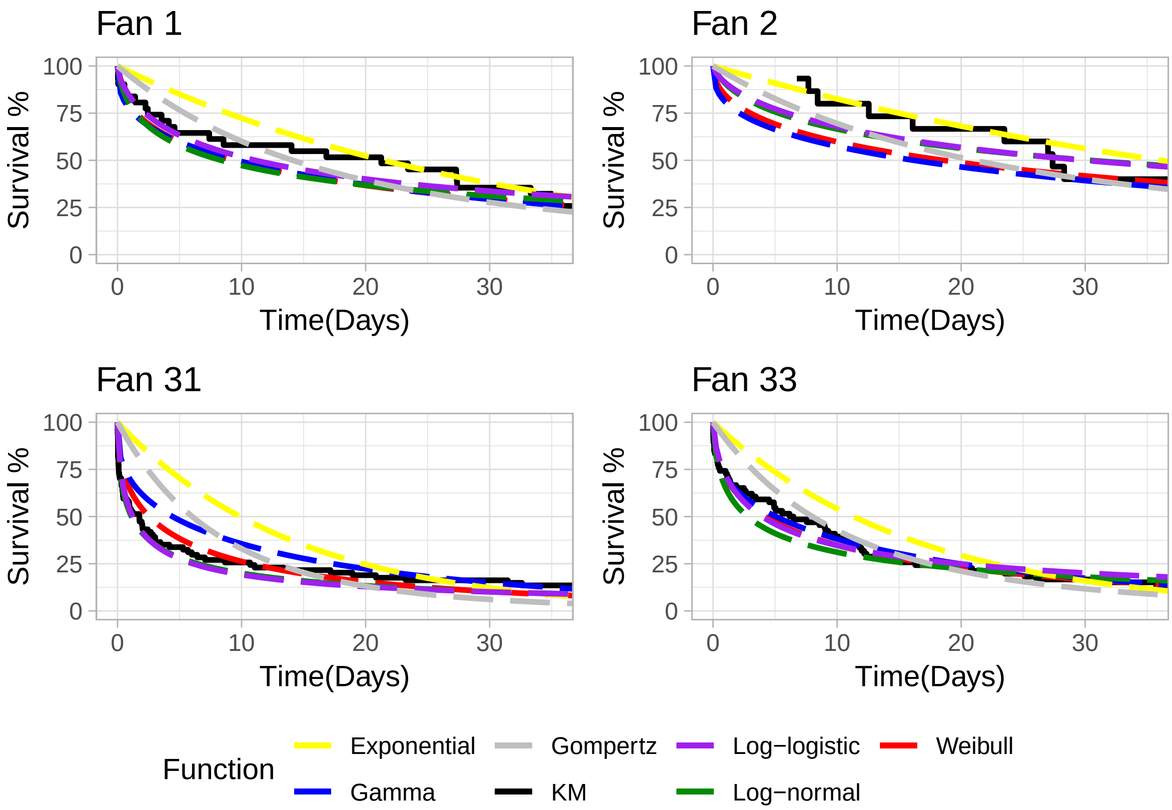

Figure 4.26: Comparison of Kaplan-Meier survival estimate with other model distributions

Table 4.20 contains a comparison of the Akike's Information Criterion(AIC) from each of the respective fitted models. The smallest AIC value of 1256.74 supports the notion that a Weibull model provides the closest approximation of the survival function for each fan.

Table 4.20: AIC comparison of model distribution options

Weibull	Log-normal	Gamma	Exponential	Log-logistic	Gompertz
1256.738	1278.83	1258.585	1473.096	1280.537	1425.85

Following from the conclusions that the reliability of the four fans in section 5 is not equivalent but is well approximated by a Weibull distribution, parametric models can be estimated to model the behavior of respective fan.

As performed in the analysis of the cement mill data, the assumptions of a well fitting Weibull distribution, as well as proportional hazard will be assessed via the complementary log-log transformation. Figure 4.27 contains a plot of the relationship between $log(-log[\hat{S}_c(t_i)])$ and $log(t_i)$, along wit ha least squares line, for each of the fans in section 5.

As evident by the plot, linearity of the points suggests that the Weibull distribution is a sufficient candidate for modelling the reliability of each fan. Furthermore, despite the respective lines for each fan being nearly parallel, fan 2 intersects fans 1 and 33, indicating that the proportional hazards assumption is not valid. This plot further illustrates the unique behavior of fan 2, which may be a consequence of the fewer number of observed failures.

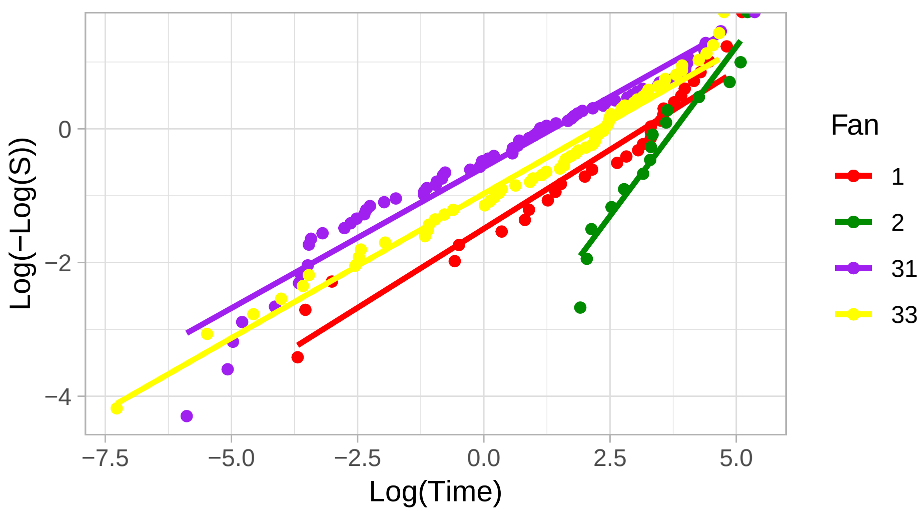

Figure 4.27: Assessment of Weibull fit for each of the fans

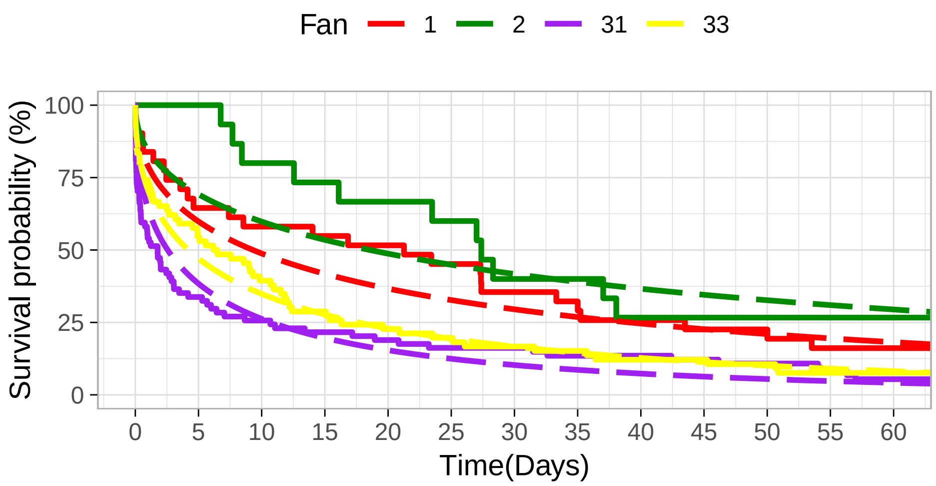

Figure 4.28: Comparison of Kaplan-Meier survival estimate with Weibull model fit

Assuming that the reliability of each fan is not equivalent, but each can be adequately represented by a Weibull function, we can estimate respective model parameters using maximum likelihood estimation. Figure 4.28 shows a comparison of the output of an estimated parametric regression model against the KM survival estimate for each fan. As seen in this plot, the discrepancies between the Weibull model and the KM estimate for fans 1 and 2 are quite pronounced for interfailure durations of less than 30 days. As the Weibull appears to closely approximate the KM estimate for the remaining two fans, this may indicate that fans 1 and 2 have experienced relatively few failures after short

Table 4.22: Parameter estimates for each fan

	Fans			
Parameter	1	2	31	33
Shape: α	0.4831104	0.4831104	0.4831104	0.4831104
Scale: μ	5.4685521	19.8709716	39.6146630	8.9544166

durations. Given the close approximation for fans 31 and 33, it supports the continued use of the Weibull distribution.

Similar to the cement mill analysis, the fitted Weibull model uses the parameterization of α and μ as the shape and scale terms respectively. The shape parameter is held constant, while the fan identifier is used as a categorical covariate with a multiplicative effect on the scale parameter. In this parameterization, fan 31 is used as the baseline, with covariates x_1, x_2, and x_3 representing dummy variables for fans 1, 2, and 33. Fan 31 is chosen as the baseline as it has experienced the highest number of failure events, which is more likely to provide an example of realistic fan performance.

$$S(t) = exp(-(\frac{t}{\mu})^{\alpha}), \qquad log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{4.4}$$

Table 4.21 contains the estimates of the intercept β_0 and β_1, β_2, and β_3 coefficients. Individually, each coefficient represents the logarithm of the ratio of survival times of each fan with the baseline fan(fan 1). As such, positive coefficients such as β_1 and β_2 represent longer survival times for fans 1 and 2, compared to fan 31. Conversely, the negative coefficient for β_3 suggests shorter survival times for fan 22.

Table 4.21: Beta coefficient estimates

Coefficients			
β_0	β_1	β_2	β_3
1.699014	1.290246	1.980185	0.493133

Table 4.18 contains the respective shape and scale parameters for each of the fans derived from the estimated coefficients. As the fan ID is included in the model as a categorical covariate, only one model is being estimated, and thus the shape parameter α remains constant. The dummy variables for each fan have a linear effect on the the log of the scale parameter μ, with each respective value reported in the table.

4.5 Integrated predictive reliability models

In addition to basic models for estimating the reliability of each cement mill without covariates, there is an array of additional methods capable of accommodating all of the previously extracted information.

4.5.1 Integrated models for cement mills

Table 4.23 contains the frequency of relative production output according to quantile for each cement mill. As shown in the table, it appears that all cement mills do not experience the same relative production load. For example, cement mill 1 experiences only 1st(0-36%) and 2nd(36-46%) quantile production loading, with cement mill 3 experiencing only 1st quantile production loading. Conversely, cement mill 4 experiences loading across the entire range, while mill 5 experiences only high loading(Q3:46-53%, Q4:53-67%). Based on this information, it may be valuable to account for production rate when modelling reliability.

Table 4.23: Count of interfailure durations by production rate

Equipment	Production Rate	Percentage
06CM41	Q1	8
06CM41	Q2	37
06CM41	Q3	31
06CM41	Q4	24
06CM11	Q1	75
06CM11	Q2	25
06CM31	Q1	100
06CM51	Q2	23
06CM51	Q3	31
06CM51	Q4	46

[a] Q1:0-36%, Q2:36-46%, Q3:46-53%, Q4:53-67%

4.5.1.1 Stratified extended Cox models for cement mills

As explained in Therneau (2000) an extension of the Cox model allows for stratification, such that the observations are divided into disjoint strata. Each strata has its own baseline hazard function, but common coefficient values for the coefficient vector β. Thus, the hazard function for interfailure duration i in stratum k has the form $h_k(t)e^{X_i\beta}$. Stratification is useful because it allows for adjustment of confounding covariates, or covariates which do not satisfy the proportional hazards assumptions, such as time-dependent covariates. An unfortunate aspect of stratification in the extended Cox model is that as the baseline hazard function is not estimated, the effect or importance of the strata is not estimated(Therneau 2000).

Furthermore, fitted Cox models may include the interaction between strata and covariates, which identifies whether the effect of covariates differs by strata. Including each covariate by strata interaction is equivalent to modeling each strata separately(Therneau 2000).

An extended Cox model including both time-independent and time-dependent covariates takes the form:

$$h_k(t, \mathbf{X}(t)) = h_k(t) \exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right] \tag{4.5}$$

With $X_1, X_2, \ldots X_{p_1}$ the time-independent covariates and $X_1(t), X_2(t), \ldots X_{p_2}(t)$ the time-dependent covariates of interest(Kleinbaum, Klein, and Samet 2006). However, when the extended Cox model includes time-dependent covariates, it may no longer satisfy the proportional hazards assumption, as both the baseline hazard and the covariates depend upon time. The proportional hazards assumption will be verified for all covariates.

Table 4.24 contains a comparison of extended Cox model fits for the available cement mill data using the "coxph" function in the *Survival* package in R. Model 1 is the full model, containing all covariates, as well as all covariate by strata interactions. Model 2 contains all main effects, while Model 3 contains only the significant main effects of *Production Replace*.

As the *Replace* covariate indicates whether the previous maintenance action involved replacement or not(0:no, 1:yes), the baseline value is set to 0. Thus, the model is estimating the effect of replacement when compared to non-replacement. Additionally, as Model 1 includes interaction terms with each cement mill, cement mill 4 was chosen as the baseline, as it is more operational, and at the widest range of production capacity.

In the case of this analysis the model is *stratified* meaning that the estimated coefficients for the main effects represents an effect that is present in all cement mills. Including an *interaction* between a maint effect and a strata, seeks to identify whether the the effect of a variable is different for a particular cement mill.

Despite Model 1 indicating a significant interaction between *Production* and *CM1*, Model 3 results in a lower AIC without including any interactions. Furthermore, as cement mill 4 is the baseline, this interaction implies that the effect of production is not the same for cement mill 1. This notion is expected, because as shown in Table 4.23, cement mill 1 spends 75% of operating time at Q1 production levels and 25% at Q2 production levels, having never operated at higher capacity rates. As such, the interaction term will be excluded from consideration in favor of Model 3.

The use of the proportional hazards framework assumes that the effect of covariates is constant over time. PH models that include time-dependent covariates imply a restriction on these covariates such that $\delta(t) = \delta$(Therneau 2000). This means that although the value of the time-dependent covariates change over time, the estimated coefficient must not. Thus, if the estimated coefficient is constant over time, the time-dependent covariate satisfies the proportional hazards assumption.

In order to evaluate the PH assumption, Therneau (2000) recommend the both a statistical test of the Schoenfield residuals and visual inspection of the residual plots. The *cox.zph* function from the *Survival* package, which tests for an interaction between a covariate and time, was used to formally test the proportional hazards assumption for both models 2 and 3. When testing for a change in the estimates over log(time), the results for model 2 and 3 indicate that all covariates except for *Production* satisfy the proportional hazards assumption. However, when performing the same test using the real time scale, *Production* is now proportional under model 3.

As the proportionality assumption refers to whether the estimated coefficient for *Production* changes over time, this behavior can be assessed and interpreted using plots of the Schoenfield residuals against time. Figure 4.29 contains the Schoenfield residual plot for the estimate of *Production* under model 3 on both real-time(left plot) and log(time)(right plot) scales. The blue lines represent the estimated coefficient for *Production* under model 3, the circles represent the Schoenfield residuals, and the solid line represents a LOWESS(locally weighted scatterplot smoothing) fit of the estimated coefficient. Under proportionality of *Production* the LOWESS fitted line would be horizontal, representing a constant value for the estimated coefficient over time.

Despite a significant test result, neither plot indicates a clear trend in the Schoenfield residuals. The left plot shows that the LOWESS fit is severely effected by the observed failures that occur between the 50-100 day range where observations are relatively sparse. In the right plot, the LOWESS fit demonstrates an upward trend near the larger survival times, where observations are more sparse. In both plots, the departure of the fitted line from the estimated coefficient value(blue line) is not well supported by the observed residuals, which do not exhibit an extreme trend. Based on analysis of the residual plots, the results of the statistical test will be ignored, and the proportional hazards assumptions will be considered satisfied.

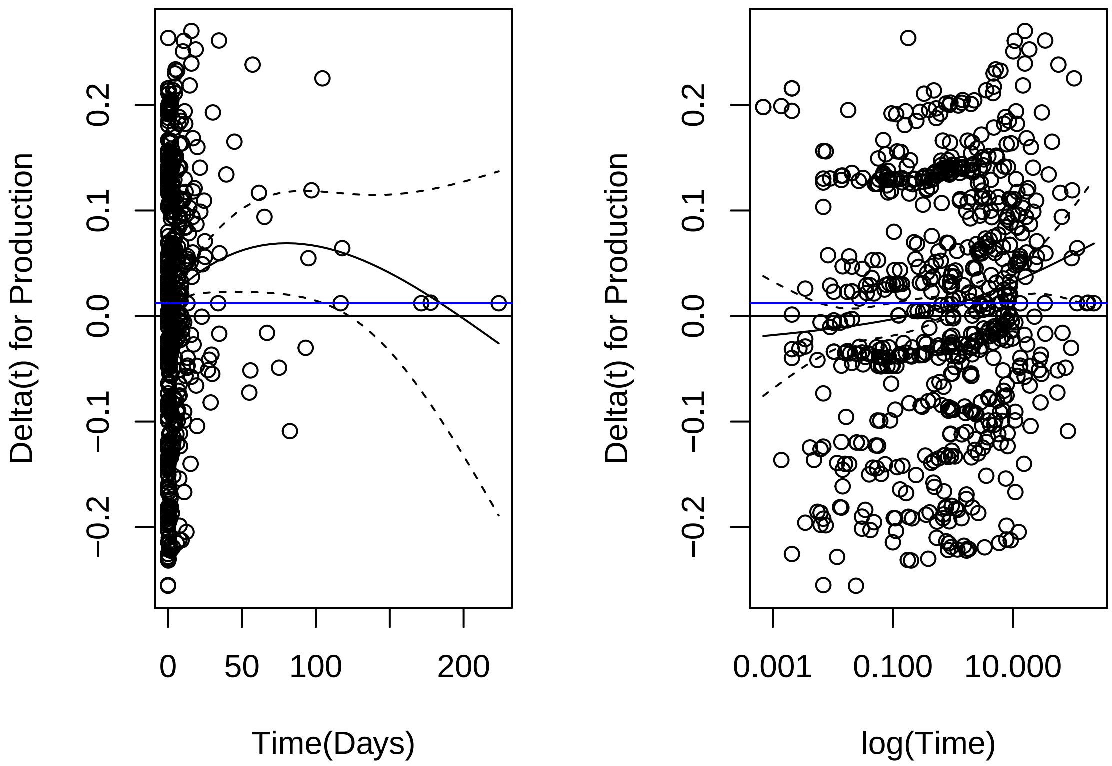

Figure 4.29: Time-dependent coefficient plots of Production for extended Cox model 3. The left plot represents the change in the effect of production over real time, and the right plot represents the change in the effect of production over log(time).

For Model 3, the estimated coefficients for $\beta_{Replace}$ and $\delta_{Production}$ are provided, along with respective standard errors in parenthesis. An estimate of 0.0121 for $\delta_{Production}$, yields a 95% confidence interval of [1.004, 1.021], representing the respective hazard ratio of a 1 unit increase in production rate(measured in percent). This confidence interval indicates that the hazard, or instantaneous risk of failure, increases between 0.4% and 2.5% following a 1% increase in production. This estimate aligns with intuition regarding maintenance and reliability, in that an increase in machine utilization relative to maximum capacity would negatively effect reliability due to increased wear and tear.

Additionally, an estimate of -0.328 for $\beta_{Replace}$ yields a 95% confidence interval of [0.521, 0.996] for the respective hazard ratio comparing interfailure durations following "replacement" maintenance actions versus "non-replacement" maintenance actions. In other words, this confidence interval suggests that replacing a failed component yields between a .004% and 47.9% decrease in the instantaneous risk of failure as compared to a repair. This estimate is also intuitive as a replacement action would ideally have a renewal effect, thereby improving reliability immediately after.

Table 4.24: Comparison of stratified extended Cox models for cement mills

	Models		
	(1)	(2)	(3)
Production	0.008 (0.006)	0.012*** (0.004)	0.012*** (0.004)
Replace	−0.185 (0.292)	−0.376** (0.185)	−0.328** (0.175)
Maint.	−0.143 (0.131)	−0.114 (0.094)	
Numstop	0.011 (0.012)	0.012 (0.009)	
Repairtime	−0.004 (0.378)	0.008 (0.023)	
Production*CM1	0.032** (0.013)		
Production*CM3	−0.010 (0.016)		
Production*CM5	0.003 (0.013)		
Replace*CM1	0.067 (0.445)		
Replace*CM3	0.416 (1.064)		
Replace*CM5	−0.611 (0.508)		
Maint.*CM1	0.222 (0.298)		
Maint.*CM3	−0.025 (0.358)		
Maint.*CM5	−0.109 (0.238)		
Numstop*CM1	−0.018 (0.025)		
Numstop*CM3	0.055 (0.046)		
Numstop*CM5	−0.003 (0.029)		
Rep.time*CM1	0.011 (0.379)		
Rep.time*CM3	0.025 (0.392)		
Rep.time*CM5	−0.742 (1.116)		
AIC	6081.92	6067.33	6064.63
Observations	3,001	3,001	3,001
Log Likelihood	−3,020.961	−3,028.664	−3,030.313
Wald Test	32.310** (df = 20)	15.480*** (df = 5)	13.310*** (df = 2)
LR Test	31.054* (df = 20)	15.648*** (df = 5)	12.350*** (df = 2)
Score (Logrank) Test	29.198* (df = 20)	15.389*** (df = 5)	11.956*** (df = 2)

Note: *p<0.1; **p<0.05; ***p<0.01

4.5.1.2 Parametric model estimation for cement mills

Although the extended Cox model provided insight into the effects of production rate and replacement on the hazard ratio, nothing is known about the actual baseline hazard function of the cement mills. As such, a fully parametric model is used to provide a complete solution for modelling the reliability of each cement mill.

Table 4.25 contains a comparison of fully parametric AFT models estimated using the “aftreg” function from the *eha* package in R, using a Weibull baseline function. As the results of the extended Cox procedure ruled out interaction effects, Model 1 includes only main effects, of which *Production*, *Replace*, and *Maint.* are significant. Model 2 is a reduction of Model 1, eliminating *Repairtime* and *Numstop* as insignificant terms. In order to fully parametrize the model, scale and shape parameters μ_k α_k are estimated for each of the strata(k), with the estimated coefficients being common between strata.

In contrast to the extended Cox model, the estimated coefficients of an AFT model represent logarithms of ratios of survival times. In other words, positive coefficients indicate longer survival times, and negative coefficients indicate decreased survival times.

In terms of the estimates for Model 2, $exp(\delta_{Production}) = 0.977$ indicates that a one unit increase in production rate corresponds to a .977(95%CI[.962, .993]) ratio of relative survival times, or a 2.3% reduction in survival time. Additionally, $exp(\beta_{Replace}) = 1.848$ which suggests that survival durations immediately following replacement actions are a 1.848(95%CI[1.73, 1.97]) times longer than those following non-replacement actions. Finally, $exp(\beta_{Maint.}) = 1.291$, which suggests that a single additional planned maintenance event results in survival times that are 1.291(95%CI[1.107, 1.506]) times longer.

All three of these main effects estimates are intuitive, and align with the inference drawn from the extended Cox model estimation. In summary, an increase in production rate corresponds to decreased equipment reliability, while performing replacement actions, and increasing the number of planned maintenance interventions both significantly increase the reliability of an equipment.

Table 4.25: Comparison of parametric accelerated failure time models for cement mills

	Models	
	(1)	(2)
Production	-0.023^{***} (0.008)	-0.023^{***} (0.008)
Replace	0.601^{*} (0.331)	0.614^{*} (0.318)
Repairtime	-0.009 (0.048)	
Maint.	0.219^{**} (0.089)	0.256^{***} (0.077)
Numstop	0.010 (0.013)	
log(scale):CM4	1.400^{***} (0.388)	1.420^{***} (0.387)
log(shape):CM4	-0.635^{***} (0.040)	-0.639^{***} (0.040)
log(scale):CM1	1.390^{***} (0.290)	1.405^{***} (0.286)
log(shape):CM1	-0.745^{***} (0.078)	-0.756^{***} (0.076)
log(scale):CM3	2.178^{***} (0.356)	2.196^{***} (0.355)
log(shape):CM3	-0.860^{***} (0.109)	-0.861^{***} (0.109)
log(scale):CM5	2.587^{***} (0.455)	2.604^{***} (0.455)
log(shape):CM5	-0.441^{***} (0.074)	-0.439^{***} (0.074)
Observations	3,001	3,001
Log Likelihood	$-1,314.651$	$-1,315.007$
Note:		*p<0.1; **p<0.05; ***p<0.01

As a fully parametric model with scale and shape μ_k and α_k estimates for each strata k, these estimates can be used to model the effect of covariates on the individual reliability of each cement mill. Using the reported estimates from Model 2, where $\hat{S}_k(t)$ represents the reliability function for each cement mill k, and *log(scale)* in Table 4.25 corresponds to β_k, the AFT model has the following parameterization:

$$\hat{S}_k(t) = exp(-(\frac{t}{\mu_k})^{\alpha_k}),$$
$$log(\mu_k) = \beta_k + \beta_{Replace}Replace + \beta_{Maint.}Maint. + \delta_{Production}Production \tag{4.6}$$

Figure 4.30 contains a comparison of the parametric AFT fit from Model 2 for each of the cement mills at differing production levels(20%, 40%, and 60%). As a follow-up from the earlier interpretation of the effect of production rate on interfailure time, these plots visually express how an increase in production rate results in a decrease in the reliability, for each cement mill. Such a fully parametric solution may be especially useful for capacity planning or when planning future maintenance interventions, as it allows for simulation of the reliability an single cement mill. Moreover, parametric models could facilitate the use of real-time reliability monitoring as production demands and environmental factors change.

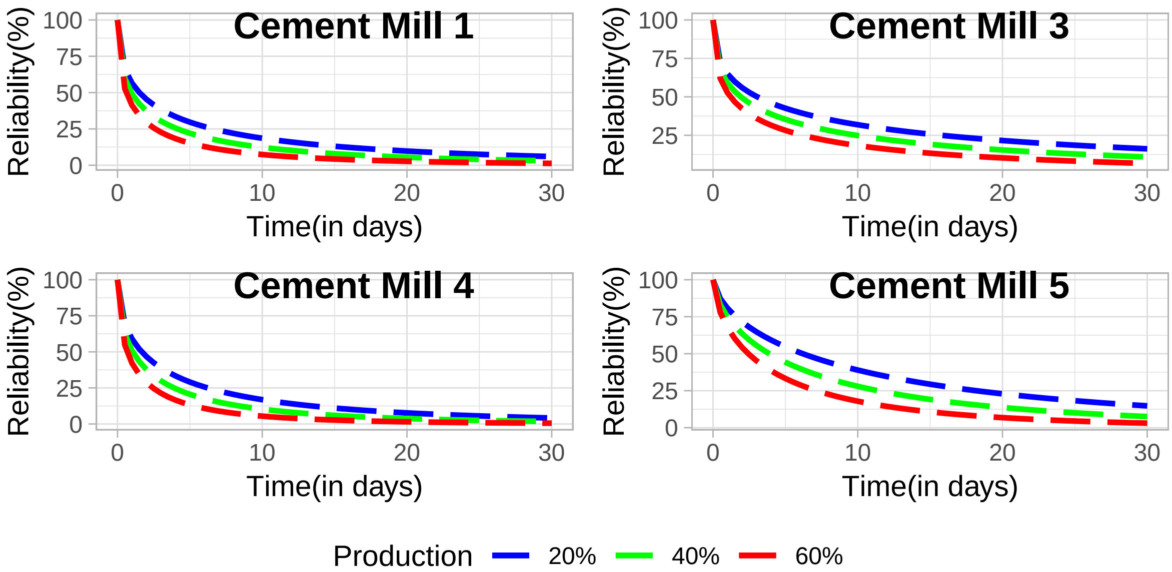

Figure 4.30: Comparison parametric AFT survival curves at differing production levels

In addition to calculating the reliability for each cement fill, conditioned on certain covariate values, a parametric model allows for easy calculation of the median time until failure as well as the **mean time before failure(MTBF)**, conditional on covariates. Given the same parameterization as used for Model 2, where the time until failure $T \sim Weibull(exp(X'\hat{\beta}), \hat{\alpha})$, the MTBF is the expected value of T such that:

$$MTBF = \hat{E}(t_i) = \int_0^{\inf} \hat{S}(t)dt = \exp(X'\hat{\beta})\Gamma(1+\hat{\sigma}) \tag{4.7}$$

where $\hat{\sigma} = \frac{1}{\hat{\alpha}}$,and Γ represents the Gamma function(Liu and Lim 2018).

Furthermore, the median time until failure conditional on covariates X is:

$$\text{median}(t_i) = \exp(X'\hat{\beta})(\log(2))^{\hat{\sigma}} \tag{4.8}$$

Following from the procedures demonstrated by Liu and Lim (2018), it is possible to estimate confidence intervals for the MTBF using the Delta method to derive the standard error of the MTTF. The standard error can be estimated as follows:

$$SE = \left\{ \begin{pmatrix} \frac{\partial E(\hat{t}_i)}{\partial \hat{\beta}} \\ \frac{\partial E(t_i)}{\partial \hat{\sigma}} \end{pmatrix}^t \Sigma_{\hat{\sigma}\hat{\beta}} \begin{pmatrix} \frac{\partial E(\hat{t}_i)}{\partial \hat{\beta}} \\ \frac{\partial E(t_i)}{\partial \hat{\sigma}} \end{pmatrix} \right\}^{\frac{1}{2}} \tag{4.9}$$

with $\sum_{\hat{\sigma}} \hat{\boldsymbol{\beta}}$, the covariance matrix of both $\hat{\beta}$ and $\hat{\sigma}$.

For example, the MTBF of cement mill 4, conditional on 50% production, no replacement, with 2 maintenance interventions can be derived as follows:

$$\begin{aligned} \exp(X'\hat{\beta})\Gamma(1+\hat{\sigma}) &= exp(\beta_k + \beta_{Replace}Replace + \beta_{Maint.}Maint. + \delta_{Production}Production)(1 + \frac{1}{\hat{\alpha}}) \\ &= exp(\beta_k + \beta_{Replace}(0) + \beta_{Maint.}(2) + \delta_{Production}(50))(1 + \frac{1}{\hat{\alpha}}) \\ &= exp(1.42 + (0.614 \cdot 0) + (0.256 \cdot 2) + (-0.023 \cdot 50))(1 + \frac{1}{0.528}) \\ &\approx 3.97 days \end{aligned} \tag{4.10}$$

Functions from the *ciTools* package in R were adapted in order to estimate the standard error of the MTBF using the AFT parameter estimates from Model 2. Figure 4.31 contains a comparison of the MTBF for each cement mill, conditional on a range of covariate values, against the observed failures. Specifically, the points on each plot indicate observed cement mill failures according to respective relative production rate and failure time. The shape of the point indicates whether the previous maintenance action involved a replacement or not. The size of the point indicates the cumulative number of maintenance interventions(preventative, planned, etc.,) that were carried out during the gap time. The solid line and red ribbon represent the MTBF and respective 95% CI conditional on *no replacement*, *two maintenance interventions*, and indicated production level. The dashed line and blue ribbon represent the respective MTBF and 95% CI conditional on *replacement* with remaining covariates identical.

As evident in the plots, most of the failures above the confidence intervals represent gap times immediately following replacement(triangles), gap times in which multiple maintenance interventions occurred(large circles), or both(large triangles or large circles). This visual representation supports the earlier conclusions regarding the affect of each covariate on reliability, as well as the insights into the difference in behavior between cement mills.

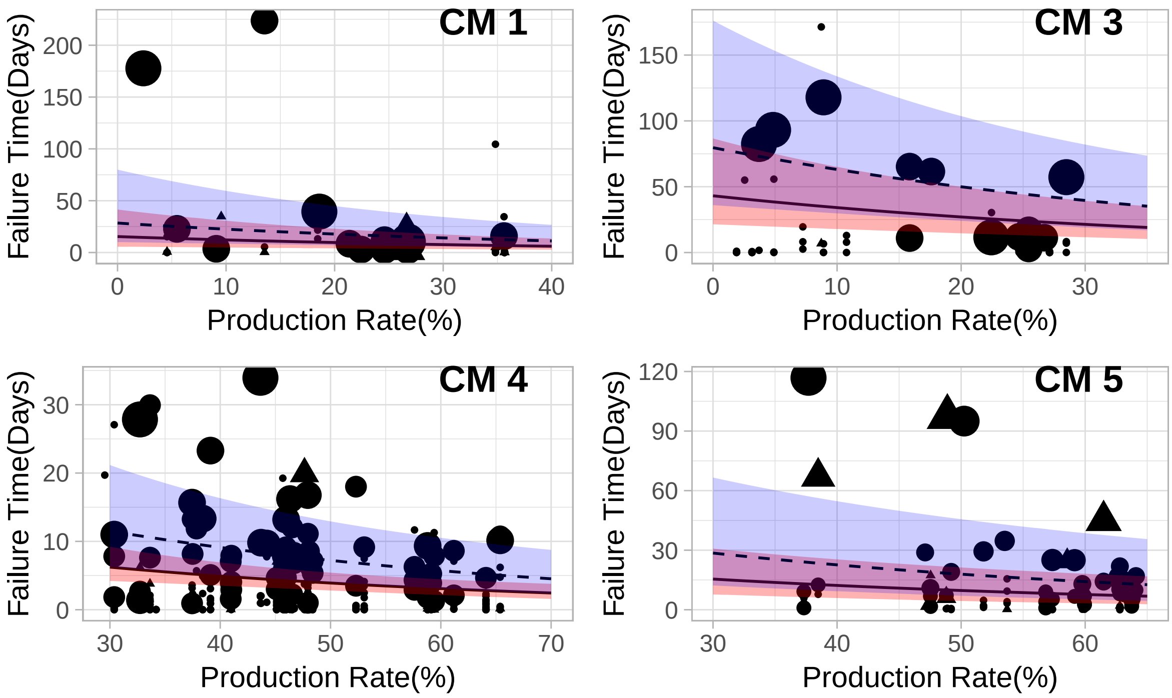

Figure 4.31: Comparison of estimated MTBF against observed failure time for each cement mill

4.5.2 Integrated models for fans

In addition to performing a basic parametric model estimation, the large quantity of available covariates for the interfailure duration of each fan make estimation of more advanced models possible. Similar to the cement mill analysis, an extended Cox model will be used to estimate a semi-parametric reliability model, given the large number of time-dependent covariates. Additionally, in order to facilitate improved estimation, the full set of fan data will be used to fit the extended Cox models.

As there is a large number of covariates, an extended Cox model will be estimated using a subset of the covariates, namely *Production*, *Replace*, *Repairtime*, and *Numstop*, in addition to interactions between the covariates and strata. The aim of the first estimation will be to identify which, if any, of these covariates are significant, along with possible interactions. The results of this estimation will be carried over into an estimation using the condition monitoring covariates, namely the vibration measurements, The results of the second estimation will provide a basis for later fully-parametric modelling.

4.5.2.1 Stratified extended Cox models for fans

Table 4.26 contains a comparison of extended Cox model fits for the available fan data using the “coxph” function in the *Survival* package in R. Model 1 is the full model, containing all covariates, including all covariate by strata interactions. Model 2 is a reduction of Model 1, containing the same covariates with the exception of the insignificant *Replace by strata* interaction term. Model 3 contains only the main effects of *Production*, *Replace*, *Repairtime*, and *Numstop*. Although *Repairtime* is no longer significant in Model 3, it was retained as it was significant in the additional models. Statistical tests of the Schoenfield residuals for each model indicate that all covariates satisfy the proportional hazards assumption.

As the *Replace* covariate indicates whether the previous maintenance action involved replacement or not(0:no, 1:yes), the baseline value is set to 0. Thus, the model is estimating the effect of replacement when compared to non-replacement. Additionally, as Models 1 and 2 include interaction terms with each strata, section 5 fan 31 was chosen as the baseline, as it is had experienced a large number of failure events.

In the case of this analysis the model is *stratified* meaning that the estimated coefficients for the main effects represents an effect that is present in all fans. Including an *interaction* between a maint effect and a strata, seeks to identify whether the the effect of a variable is different for a particular fan.

As evident in the table, Models 2 and 3 have similar AIC values, despite Model 3 containing far fewer covariate values. Given that the model parameters are being estimated using data from 10 fans operating in different areas of the plant, it is not surprising that some of the interaction terms appear to be significant.

For Model 3, the estimated coefficients for $\delta_{Production}$, $\beta_{Replace}$, $\beta_{Repairtime}$, and $\beta_{Numstop}$,are provided, along with respective standard errors in parenthesis. An estimate of -0.0183 for $\delta_{Production}$, yields a 95% confidence interval of [0.967, 0.997], representing the respective hazard ratio of a 1 unit increase in production rate(measured in percent). This confidence interval indicates that a 1% increase in production rate corresponds to between a 0.3% and 3.3% reduction in the instantaneous risk of failure. Unlike the cement mill analysis, an increase in the the production rate is estimated to reduce the risk of failure for fans. Since the fans operate at high speeds, there will be more variability in fan speed during light production, when cooling requirements fluctuate. As the production and need for consistent cooling increases, the fans operate with less variability, which may reduce the propensity for wear resulting from speed variations.

Additionally, an estimate of -0.586 for $\beta_{Replace}$ yields a 95% confidence interval of [0.355, 0.873] for the respective hazard ratio comparing interfailure durations following "replacement" maintenance actions versus "non-replacement" maintenance actions. Based on this confidence interval, performing a replacement action results in between a 12.5% and 64.5% reduction in hazard as compared to a repair. This result corroborates the result obtained from the cement mill analysis.

Furthermore, an estimate of 0.017 for $\beta_{Repairtime}$ yields a 95% confidence interval of [0.996, 1.038] for the respective hazard ratio resulting from a 1 unit(hour) increase in repair time(downtime). As this CI includes 1, Model 3 suggests that repairtime may not have an effect on the hazard function immediately after. Finally, an estimate of 0.174 for $\delta_{Numstop}$ yields a 95% confidence interval of [1.069, 1.324] for the respective hazard ratio following each additional non-failure stoppage event. This estimate suggests that each time the equipment is stopped for neither failure nor maintenance, the instantaneous risk of failure increases between 6.9% and 32.4%. This may be attributed to factors such as corrosion, start-stop wear, or load variations of the respective equipment during non-failure stoppages.

Table 4.26: Comparison of stratified extended Cox models for fans

	Models		
	(1)	(2)	(3)
prod	−0.007 (0.015)	−0.007 (0.015)	−0.018** (0.007)
rep1	−0.482** (1.033)	−0.384 (0.316)	−0.586** (0.280)
reptime	−0.244* (0.181)	−0.247* (0.179)	0.017 (0.013)
numstop	0.671** (0.371)	0.671** (0.371)	0.174*** (0.084)
prod:equipment05FN02	0.012 (0.028)	0.012 (0.028)	
prod:equipment03FN08	0.086 (0.129)	0.086 (0.129)	
prod:equipment05FN01	−0.011 (0.023)	−0.011 (0.023)	
prod:equipment05FN33	−0.017 (0.023)	−0.017 (0.023)	
prod:equipment06FN11	−0.544*** (76.809)	−0.547*** (76.809)	
prod:equipment06FN41	0.001 (0.066)	−0.0005 (0.065)	
prod:equipment06FN51	−0.304*** (0.164)	−0.305*** (0.164)	
prod:equipment06FN52	−0.693 (0.683)	−0.693 (0.683)	
prod:equipment06FN53	−0.030 (0.021)	−0.030 (0.021)	
rep1:equipment05FN02	(0.000)		
rep1:equipment03FN08	(0.000)		
rep1:equipment05FN01	0.133 (1.119)		
rep1:equipment05FN33	(0.000)		
rep1:equipment06FN11	(0.000)		
rep1:equipment06FN41	−0.064 (1.295)		
rep1:equipment06FN51	(0.000)		
rep1:equipment06FN52	(0.000)		
rep1:equipment06FN53	0.180 (1.236)		
reptime:equipment05FN02	0.285** (0.182)	0.288** (0.180)	
reptime:equipment03FN08	−0.217 (0.454)	−0.214 (0.454)	
reptime:equipment05FN01	0.259* (0.182)	0.262* (0.180)	
reptime:equipment05FN33	0.845* (0.432)	0.848* (0.431)	
reptime:equipment06FN11	(0.000)	(0.000)	
reptime:equipment06FN41	0.326** (0.190)	0.330** (0.189)	
reptime:equipment06FN51	−5.245*** (3.219)	−5.242*** (3.219)	
reptime:equipment06FN52	−2.361 (1.941)	−2.358 (1.941)	
reptime:equipment06FN53	0.166 (0.196)	0.173 (0.191)	
numstop:equipment05FN02	−2.013*** (1.132)	−2.014*** (1.132)	
numstop:equipment03FN08	(0.000)	(0.000)	
numstop:equipment05FN01	−0.172 (0.470)	−0.164 (0.465)	
numstop:equipment05FN33	−0.500 (0.385)	−0.501 (0.385)	
numstop:equipment06FN11	(0.000)	(0.000)	
numstop:equipment06FN41	−1.023*** (0.717)	−1.058*** (0.698)	
numstop:equipment06FN51	(0.000)	(0.000)	
numstop:equipment06FN52	(0.000)	(0.000)	
numstop:equipment06FN53	−1.090* (0.789)	−1.090* (0.789)	
AIC	1826.49	1820.57	1819.58
Observations	805	805	805
Log Likelihood	−884.247	−884.284	−905.791
Wald Test	354.770*** (df = 29)	318.710*** (df = 26)	18.110*** (df = 4)
LR Test	57.205*** (df = 29)	57.132*** (df = 26)	14.116*** (df = 4)
Score (Logrank) Test	51.159*** (df = 29)	50.828*** (df = 26)	14.422*** (df = 4)
Note:			*p<0.1; **p<0.05; ***p<0.01

A final estimation using extended Cox models was performed to evaluate the usage of the different condition monitoring covariates, namely the original vibration measurements and the results from the two PCA procedures. Building upon the results from Model 3, Table 4.27 contains a comparison of an additional extended Cox model estimation with the inclusion of the vibration measurements. Model 4 includes the main effects from Model 3, in addition to the 9 available vibration measures. Models 5 and 6 both contain terms for *Production*, *Repairtime*, *Numstop*, and *Replace*, in addition to the principal component scores from their respective procedures, denoted as A and B.

As evident in the table, all but 3 of the terms included in Model 4 are significant, all are significant in Model 5, and in Model 6 all but the last three principal component scores are significant. Additionally, despite containing more covariates, Model 4 results in the lowest respective AIC value. This suggests that retaining the original vibration measurements instead of a dimension reduction may result in an improved model fit, in addition to having an easier interpretation. Furthermore, regardless of the condition monitoring covariates included, each model results in similar estimates for the first 4 coefficients.

Table 4.27: Comparison of extended Cox models for fans using original values and PCA values

	Models		
	(4)	(5)	(6)
Production	−0.017** (0.007)	−0.018** (0.007)	−0.019** (0.007)
Repairtime	0.019** (0.013)	0.019* (0.013)	0.017* (0.013)
Numstop	0.185*** (0.082)	0.177*** (0.084)	0.180*** (0.083)
Replace	−0.576** (0.299)	−0.575** (0.284)	−0.561** (0.282)
ADE	−1.291*** (0.399)		
ANDE	0.504*** (0.331)		
HDE	−0.367 (0.903)		
HDE.CAV	1.171 (1.080)		
HDE.ENV	−1.569 (4.230)		
HNDE	0.546** (0.415)		
HNDE.ENV	4.294** (2.803)		
VDE	0.232*** (0.114)		
VNDE	−1.065*** (0.511)		
PCA1		1.459* (0.877)	
PCA2		−3.052*** (1.402)	
PCA3		3.822* (2.289)	
PCA4		2.498** (1.461)	
PCB1			−0.331* (0.191)
PCB2			0.205 (0.246)
PCB3			−0.334 (0.308)
PCB4			0.058 (0.215)
AIC	1807.39	1822	1823.62
Observations	805	805	805
Log Likelihood	−890.695	−903.001	−903.810
Wald Test	95.500*** (df = 13)	26.830*** (df = 8)	23.830*** (df = 8)
LR Test	44.308*** (df = 13)	19.698** (df = 8)	18.079** (df = 8)
Score (Logrank) Test	47.668*** (df = 13)	21.086*** (df = 8)	18.874** (df = 8)
Note:			*p<0.1; **p<0.05; ***p<0.01

4.5.2.2 Parametric model estimation for fans

Although the extended Cox model provided insight into the effects of maintenance related condition monitoring covariates on the hazard ratio, nothing is known about the actual baseline hazard function of the fans. As such, we advance the analysis with a fully parametric model to provide a complete solution for modelling the reliability of each fan. Although the extended Cox models were estimated using data from all fans, the fully parametric model will be estimated using only fans from section 5.

Table 4.28 contains a comparison of fully parametric AFT models estimated using the "aftreg" function from the *eha* package in R, using a Weibull baseline function. Model 1 includes the 4 maintenance related covariates in addition to the original vibration measurements. Model 2 is a subset of Model 1, including the only significant covariates from Model 1. After estimating the reduced model, some of the significant terms from Model 1 are no longer significant in Model 2. Model 3 contains a further reduction, retaining only significant covariates from Model 2.

The estimated coefficients of an AFT model represent logarithms of ratios of survival times; positive coefficients indicate longer survival times, and negative coefficients indicate decreased survival times. In terms of the estimates for Model 3, $exp(\beta_{Repairtime}) = 0.968$ indicates that a one unit(hour) increase in repair time corresponds to a .968(95%CI[.950, .987]) ratio of relative survival times. In other words, each hour of repair time corresponds to between a 1.3% and 5% reduction in reliability immediately following the maintenance action. Given this result, a longer repair time may be indicative of a more severe failure, or a failure for which an appropriate maintenance action proves difficult. Moreover,longer repair time implies intensive maintenance actions, and the possibility of human error, which may reduce equipment reliability.

Additionally, $exp(\delta_{ANDE}) = 0.644$ which suggests that a one unit(mm/s) increase in ANDE(axial non-drive end) vibration corresponds to a 0.644(95%CI[.459, .904]) ratio of relative survival times. This confidence interval indicates that the survival decreases between 9.6% and 54.1% with a 1mm/s increase in ANDE vibration. Finally, $exp(\delta_{HNDE.ENV}) = 0.001$ which suggests that a one unit(gE) increase in HNDE.ENV(horizontal non-drive end enveloping) vibration corresponds to a 0.001(95%CI[.000001, .191]) ratio of relative survival times. Similarly, a 1gE increase in HNDE.ENV vibration corresponds to between a 99.99% and 80.9% decrease in survival time. Each of the estimated effects for fan vibrations are intuitive, given their purpose of monitoring the health condition of the equipment.

Table 4.28: Comparison of parametric accelerated failure time models for fans

	Models		
	(1)	(2)	(3)
Production	0.011 (0.011)		
Repairtime	−0.027*** (0.009)	−0.032*** (0.011)	−0.032*** (0.011)
Numstop	−0.086 (0.159)		
Replace	0.459 (0.552)		
ADE	0.573** (0.283)	0.327 (0.249)	
ANDE	−0.509** (0.259)	−0.481*** (0.160)	−0.440** (0.173)
HDE	0.323 (0.617)		
HDE.CAV	−0.884 (0.662)		
HDE.ENV	7.266* (4.181)	3.705 (4.589)	
HNDE	−0.012 (0.361)		
HNDE.ENV	−8.212*** (3.061)	−8.849*** (2.985)	−7.493** (2.977)
VDE	−0.112 (0.235)		
VNDE	−0.106 (0.435)		
log(scale):Fan31	3.441*** (1.221)	2.637*** (0.991)	3.864*** (0.864)
log(shape):Fan31	−0.912*** (0.090)	−0.910*** (0.090)	−0.911*** (0.090)
log(scale):Fan2	5.886*** (1.349)	5.573*** (0.978)	6.690*** (0.910)
log(shape):Fan2	0.563** (0.230)	0.319 (0.200)	0.355* (0.204)
log(scale):Fan1	4.903*** (1.392)	4.672*** (1.037)	5.920*** (0.917)
log(shape):Fan1	−0.381** (0.155)	−0.404*** (0.150)	−0.409*** (0.152)
log(scale):Fan33	4.206*** (1.207)	3.367*** (0.977)	4.595*** (0.848)
log(shape):Fan33	−0.678*** (0.102)	−0.680*** (0.101)	−0.689*** (0.101)
Observations	526	526	526
Log Likelihood	−603.363	−607.256	−608.826
Note:			*p<0.1; **p<0.05; ***p<0.01

As a fully parametric model with scale and shape μ_k and α_k estimates for each strata k, these estimates can be used to model the effect of covariates on the individual reliability of each cement mill. Using the reported estimates from Model 2, where $\hat{S}_k(t)$ represents the reliability function for each cement mill k, and *log(scale)* in Table 4.28 corresponds to β_k, the AFT model is represented in Equation 4.11.

$$\begin{aligned} \hat{S}_k(t) &= exp(-(\frac{t}{\mu_k})^{\alpha_k}), \\ log(\mu_k) &= \beta_k + \beta_{Repairtime} Repairtime + \delta_{ANDE} ANDE + \delta_{HNDE.ENV} HNDE.ENV \end{aligned} \tag{4.11}$$

Figure 4.32 contains a comparison of the parametric AFT fit from Model 3 for each of the fans at differing *ANDE* vibration levels(3mm/s, 4mm/s, and 5mm/s), with *Repairtime* and *HNDE.ENV* held constant at their mean values(3.96hrs and .102gE respectively).

From the earlier interpretation of the effect of vibration readings on interfailure time, these plots visually express how an increase in vibrations results in a decrease in the reliability, for each fan. The plot representing fan 2 further illustrates behavior that is different from the other fans. This is supported by the difference in the *log(scale)* and *log(shape)* parameter estimates for fan 2 as compared to the estimates for the remaining fans, as shown in Table 4.28.

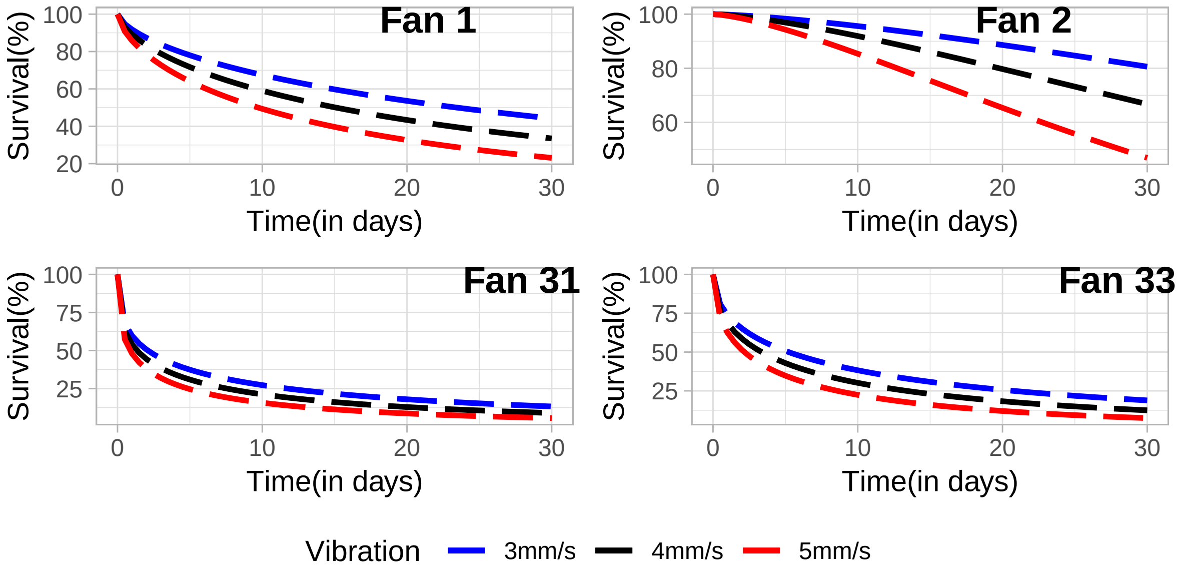

Figure 4.32: Comparison parametric AFT survival curves at differing production levels

As demonstrated in the cement mill analysis, it is possible to use the parameter estimates from Model 3 to derive the MTBF, conditional on covariates, for each of the respective fans. For example, the MTBF of fan 31, conditional on 4mm/s of ANDE vibration, a 4hr previous repairtime, and an HNDE.ENV measurement of .1gE can be derived using Equation 4.12.

$$\begin{aligned}\exp(X'\hat{\beta})\Gamma(1+\hat{\sigma}) &= exp(\beta_k + \beta_{Repairtime}Repairtime \\ &+ \delta_{ANDE}ANDE + \delta_{HNDE.ENV}HNDE.ENV)(1+\frac{1}{\hat{\alpha}}) \\ &= exp(\beta_k + \beta_{Repairtime}(4) + \delta_{ANDE}(4) + \delta_{HNDE.ENV}(.1))(1+\frac{1}{\hat{\alpha}}) \\ &= exp(3.86 + (-.032 \cdot 4) + (-.440 \cdot 4) + (-7.493 \cdot .1))(1+\frac{1}{0.402}) \\ &\approx 11.17 days \end{aligned} \quad (4.12)$$

Functions from the *ciTools* package in R were adapted in order to estimate the standard error of the MTBF using the AFT parameter estimates from Model 3. Figure 4.33 contains a comparison of the MTBF for each fan, conditional on a range of covariate values, against the observed failures.

Specifically, the points on each plot indicate observed fan failures according to respective ANDE vibration value and failure time. The size of the point indicates the duration of the last repair(in hours) prior to the start of the gap time. The solid line and red ribbon represent the MTBF and respective 95% CI conditional on a 2hr repair time and indicated production level. The dashed line and blue ribbon represent the respective MTBF and

95% CI conditional on 10hr repair time with remaining covariates identical. For both confidence intervals, the HNDE.ENV value is held constant at its mean of .102gE.

As evident in the plots, the MTBF decreases substantially as the value of the ANDE vibrations increases. In contrast with the representation from the cement mill analysis, which included production level, number of maintenance interventions, and replacement status, these plots only depend upon values that are recorded as the result of a manufacturing process. However, this plot may also be helpful for guiding production decisions, as it provides insight into the consequences of operating equipment that is experiencing high vibration levels, or has undergone long repairs.

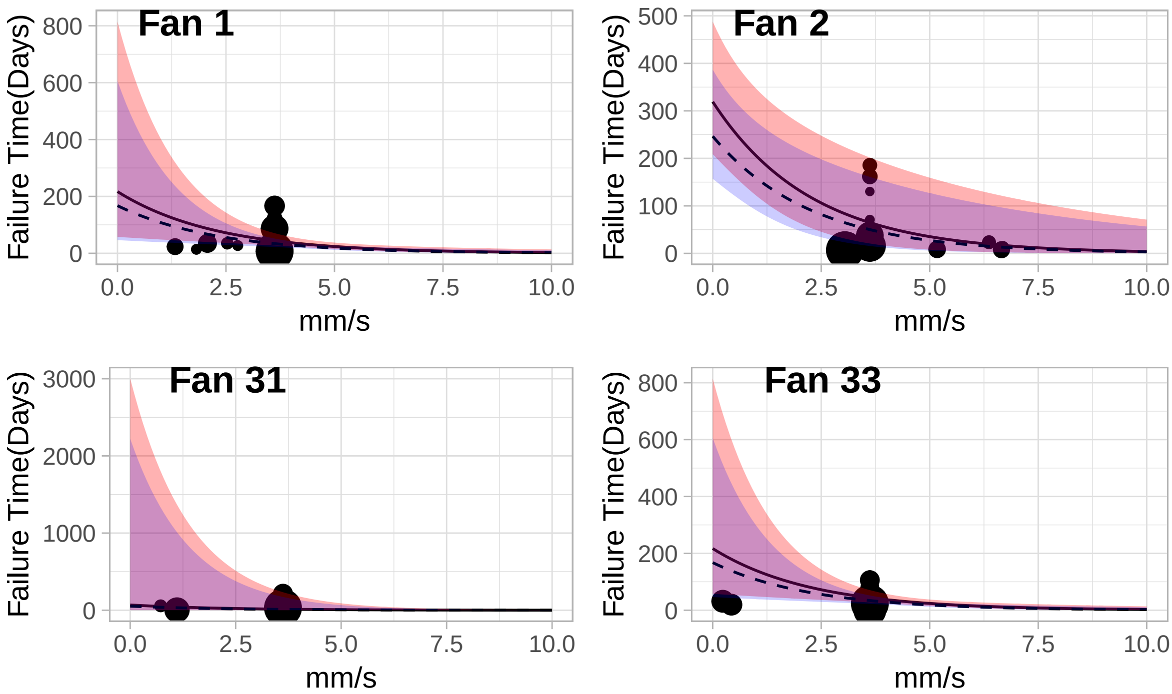

Figure 4.33: Comparison of estimated MTBF against observed failure time for each fan

4.6 Classification models

In addition to modelling reliability using traditional survival analysis techniques, the wide array of covariates that have been extracted from the data sources allow for an equally wide selection of modelling tools. During this research, the role of covariates representing observed vibration measurements have been of key interest, as they, ideally, represent the current health status of the equipment. Such indicators of the condition of the equipment are especially valuable as they are virtually non-invasive, meaning that the equipment can be monitored without the need to stop and re-start the machine. In contrast, the majority of the extracted covariates are byproducts of the frequent stoppage events that the equipment experiences.

In theory, a maintenance engineer that is able to continuously(or intermittently) monitor the condition of equipment is able to use this insight for the purpose of either scheduling a *maintenance intervention* or identifying the optimal *maintenance action* for the next failure. Over time, this procedure of *maintenance decision support* would be reflected in the historical records of both the condition indicators and the respective past maintenance actions. The following section will explore a predictive modelling approach in an attempt to classify maintenance actions using historical covariates.

4.6.1 Predicting maintenance action from covariates

The aim of the first predictive model was to train artificial neural network(ANN) using the integrated fan dataset to predict the respective maintenance action conditional on respective covariates. As the fan dataset contains historical records pertaining to a specific equipment during a unique time between two failures, there are multiple rows representing the change in the time-dependent covariates leading up to each failure. In order to consistently provide a fixed number of inputs to the ANN, only the final observation in each gap time is used to train the network. This implies that the ANN is training using only the values observed immediately prior to failure, and not the *history* of the gap time.

In addition, 70% of the data is selected for use to train the model, with the remaining 30% used to evaluate the model performance. All models were trained using the *neuralnet* package, which allows the user to determine the number of neurons and hidden layers, the desired error function, the activation function, and the training algorithm, among other features.

As the larger fan dataset will be used for the predictive modelling, the 12 of the covariates used in extended Cox Model 4 in the previous section will be retained, in addition to a covariate representing the time of the imminent failure(*Tstop*). Although a model comparison is not provided, several ANNs were trained and compared using resilient backpropagation, sum of squared error(SSE) and cross-entropy(CE) error functions, in addition to logistic and hyperbolic-tangent activation functions. The results of this model were largely equivalent regardless of function choice, and the results of a network with a single hidden layer containing 8 neurons using the SSE and logistic functions is summarized by the confusion matrix in Table 4.30

When evaluated with the test set, the accuracy of the network was 9.6%, as reflected in the confusion matrix. Such a poor classification results is not entirely surprising given the non-uniform distribution of maintenance actions, as illustrated previously in Table 4.6. As a result, there are relatively few samples in some categories for the purpose of training, and even fewer in the smaller testing set. The results of this model are summarized in the confusion matrix presented in Table 4.29, which also reports the sensitivity and specificity for each category of maintenance action.

4.6.2 Predicting replacement probability from covariates

Given the difficulties of the previous model in classifying maintenance actions into sparsely represented categories, a further analysis was performed after dichotomizing the maintenance action into a binary indicator of whether the maintenance action was a *Replace* or

Table 4.29: Confusion matrix ANN predicting category of future maintenance action from covariates

Predicted Class.	Observed Class.							
	1	2	4	5	9	12	Sensitivity	Specificity
1	9	3	77	1	0	1	1	0.03
2	0	0	1	0	0	1	0	0.98
4	0	1	0	0	0	0	0	0.94
5	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	1

Table 4.30: Comparison of models for predicting replacement

Model	Type	MSE	Accuracy	Sensitivity	Specificity
1	6	0.114	0.883	0.222	0.953
2	8	0.113	0.872	0.444	0.918
3	12	0.105	0.894	0.222	0.965
4	10	0.100	0.894	0.333	0.953
5	GLM	0.110	0.883	0.000	0.976
6	GLM	0.088	0.894	0.000	0.988

Note:
Type: GLM for logistic regression, otherwise it indicates the number of neurons in the respective hidden layers

any other category. Using a binary response not only provides a more balanced distribution with which to train and test, it allows for the usage and comparison of traditional statistical classifiers such as a logistic regression.

Table 4.30 provides a comparison of ANNs with several different configurations against two logistic regression models. Models 1, 2, 3, and 4 all include the same 13 covariates, namely *Production*, *Numstop*, *Tstop*, *Repairtime*, *ADE*, *ANDE*, *HDE*, *HDE.CAV*, *HDE.ENV*, *HNDE*, *HNDE.ENV*, *VDE*, and *VNDE*. All four models were trained using resilient backpropagation, a SSE error function, and a logistic activation function with 6, 8, 12, and 10 neurons in a single hidden layer respectively. During exploration, modeling including more than one hidden layer were attempted, however the procedure did not converge. The logistic regression corresponding to Model 5 contained the same 13 covariates and the ANN models. As the significance tests provided by Model 5 indicated that *Production* was the only significant covariate, Model 6 was estimated with only production as a predictor.

As the motivation behind these models is to identify conditions that indicate an equipment is likely to be replaced, a replacement value of 1, indicating that replacement was performed, is classified as a *positive*. In the context of the classification results in Table 4.30, *sensitivity* corresponds to the the percentage of *positives* correctly identified. In other words, this represents when the model correctly identified that a replacement occurred. Conversely, *specificity* corresponds to when the model correctly identified that the equipment did not need replacing. As evident in the table, Model 1 performs the best classification in terms of overall accuracy, sensitivity, and specificity.

The confusion matrix for Model 1 is provided in Table 4.31, in which the columns in-

Table 4.31: Confusion matrix for Model 1

Predicted	Observed	
	0	1
0	81	7
1	4	2

dicate the observed value of replacement, and the rows indicate the predicted value of replacement. Although the accuracy and specificity of model 1 are quite high, the sensitivity is still quite low. The confusion matrix in Table 4.31 provides additionalcontext to these values, as it identifies that there are only 9 *replacement* actions compared to 85 *non-replacement* actions in the testing set. The fact that the model almost always classifys observations as *non-replace* indicates that it is unable to identify a consistent pattern between the covariates and the resulting maintenance actions. Given our knowledge that the plant maintenance engineers are currently unable to use all available data to decide on the appropriate maintenance action, it is not surprising that the model suggests these actions are inconsistent. If the plant can use the framework of integration and analysis demonstrated throughout this research to identify a set of maintenance rules based on equipment conditions(e.g., if vibration exceeds threshold, perform inspection), then classification models would theoretically be able to identify these rules via patterns in future data.

Chapter 5

Conclusion

This chapter will provide a conclusion to the research and summarize the extend to which the research objectives were achieved. In addition, the results from the statistical analysis of cement mills and fans will be summarized, detailing the implications that the results have on equipment maintenance. Finally, a post-analysis discussion will provide identify suggestions for improvement of the methodology and recommendations for future work.

5.1 Research objectives

In summary, the main objective of this research was to develop an integrated predictive model, incorporating failure event records, production output records, and vibration observations, that can be used to predict the reliability and behavior of the mechanical equipment of the plant. In order to achieve this objective, several specific objectives were first completed.

First, each data source was aggregated, pre-processed, and structured in order to make them suitable for statistical analysis. The second specific objective was to perform a descriptive analysis on the prepared data sources in order to identify important characteristics regarding the scope of the data collected, and to extract preliminary insights. Additionally, the data was used to perform a criticality analysis which identified the cement mills and fans as critical equipment within the plant. The third objective was to build and demonstrate the use of reliability models for the critical equipment. The finally objective involved an integration of the failure event records, the production output records, and the condition monitoring(vibration measurements) records for the purpose of building an integrated predictive model.

As the research successfully accomplished all specific objectives, the primary objective was also accomplished, and several integrated predictive models were generated. The stoppage event records and the monthly production records were integrated and used to build extended Cox and accelerated failure time reliability models for 4 cement mills in plant section 6. A complete integration of all three data sources was performed and used to estimate an extended Cox model for 10 fans and an accelerated failure time model for 4 fans. Furthermore, the integrated fan data was used to build artificial neural network and logistic regression models for the purpose of predicting future maintenance actions from the observed event, production, and vibration measurements.

5.2 Analysis results

As the integrated predictive models estimated the effect of equipment-specific variables on the reliability of the equipment, the results of these models, and respective inference, can be used to derive future maintenance decision support. Based on the integrated data for cement mills, the extended Cox model(semi-parametric) and accelerated failure time model(fully parametric) both indicate that an increase in production rate corresponds to an increase in the risk of failure, or decrease in reliability. Additionally, both models indicate that replacing a component of the cement mill in response to a failure significantly increases reliability immediately after, as compared to simply repairing the cement mill. Finally, the accelerated failure time model also indicated that performing additional maintenance interventions, before failure occurs, significantly increases the reliability of the cement mill.

Based on the integrated data sources for fans throughout the plant, the extended Cox model provided the same conclusion regarding production rate(an increase in production yields a decrease in reliability) and equipment replacement(replacing a component increases reliability more than repair). However, this model also indicated that each time a fan must be stopped, for reasons other than failure or maintenance, the reliability significantly decreases once it resumes operation. As the accelerated failure time model was estimated using only 4 of the fans from the same section, it identifies several different significant effects. This model suggests that as the duration of the repair time increases, the reliability of the fan immediately following the repair is decreased. Furthermore, it indicates that as the observed vibrations(ANDE and HNDE.ENV) increase, the reliability of the fan decreases.

Lastly, the ANN and logistic regression classification models managed to provide more accurate predictions of future action as the complexity of the models decreased. When predicting the future maintenance action from covariates, the model behaved poorly. However, when reducing the output to a classification *replace* or *non-replace* the model prediction greatly improved.

In conclusion, these predictive models provide a clear indication of the value of using data integration to identify changes in equipment reliability. In addition to tracking and accounting for historical behavior and performance, monitoring the condition of an equipment through vibrations can provide a key indicator of the health of equipment, which can be used to motivate maintenance interventions.

5.3 Post-analysis discussion

This section will serve to briefly summarize some points of discussion regarding the research, including suggestions for future work, improved methodology, and alternative analyses.

As the source data for this research is extensive, yet largely unstructured, this research demonstrates the need for structure and standardization when making and recording observations. Additionally, it demonstrates the benefit of having compatibility between data sources based upon a standardized reference. Although text mining was used to classify a failure mechanism and maintenance action for all failure events, there is tremendous value in documenting this information in real-time according to engineering standards. Additionally, tracking and classifying the degree of repair(e.g., complete, partial, etc.) or specific component replaced will allow for improved estimation of reliability, and more specific maintenance actions.

The use of vibration measurements to predict equipment reliability demonstrates the necessity for more frequent, or continuous, monitoring of equipment. Given the irregular observation intervals and need for imputation, it is likely that the model estimation would greatly improve with more frequent observations. Continuous vibration monitoring would allow maintenance engineering to precisely identify the health status of an equipment and allow for instantaneous maintenance intervention if deemed necessary.

Additionally, when integrating the data sources and accounting for both the occurrence of non-failure stoppages(*Numstop*) in addition to maintenance interventions(*Maint.*), the use of a cumulative sum makes an assumption that the effect of both types of events has an additive relationship. In terms of non-failure stoppages, this assumption may be correct, as repeated stoppages may cause an *accumulation* of wear, which might be well represented by a cumulative sum. In contrast, the effect of planned maintenance, in theory, should promote an immediate increase in reliability, which would eventually diminish with time. In this regard, the effect of repeated planned maintenance interventions may be better represented by a different function. An alternative to a cumulative sum could be a measure of the time since last maintenance, which could receive a slight reduction following each maintenance intervention, while still accounting for the effect of time. This reduction in time could also vary depending upon the type or severity of the maintenance intervention.

Although the plant has experienced a large number of mechanical and electrical failures, the majority of the stoppage events records have not been thoroughly studied. Although the cumulative number of non-failure stoppages was accounted for in the data integration, the number of different stoppage categories could be particularly useful in distinguishing between *intermittent* and *extended* failures, as discussed in the literature review. It could be useful to account for these different non-failure stoppages in reliability models, or even to study the occurrence of these events themselves using intensity or count models.

Appendix

Table 5.1: ISO 14224:2016 standardized failure mechanism classification

Category	Code	Classification
Mechanical failure	1.0	General
	1.1	Leakage
	1.2	Vibration
	1.3	Clearance/alignment failure
	1.4	Deformation
	1.5	Looseness
	1.6	Sticking
Material failure	2.0	General
	2.1	Cavitation
	2.2	Corrosion
	2.3	Erosion
	2.4	Wear
	2.5	Breakage
	2.6	Fatigue
	2.7	Overheating
	2.8	Burst
Instrument failure	3.0	General
	3.1	Control failure
	3.2	Nosignal/indication/alarm
	3.3	Faultysignal/indication/alarm
	3.4	Out of adjustment
	3.5	Software error
	3.6	Common cause/Common mode failure
Electrical failure	4.0	General
	4.1	Short circuiting
	4.2	Open circuit
	4.3	No power/voltage
	4.4	Faulty power/voltage
	4.5	Earth/isolation fault
External influence	5.0	General
	5.1	Blockage/plugged
	5.2	Contamination
	5.3	Miscellaneous external influences
Miscellaneous	6.0	General
	6.1	No cause found
	6.2	Combined causes
	6.3	Other
	6.4	Unknown

Table 5.2: ISO 14224:2016 standardized maintenance activity classification

Code	Action
1	Replace
2	Repair
3	Modify
4	Adjust
5	Refit
6	Check
7	Service
8	Test
9	Inspection
10	Overhaul
11	Combination
12	Other

Table 5.3: Quantiles of vibration measurements by status

Measurement	Status	Min.	Q1	Median	Q3	Max.
ADE	Normal	0.005	1.3635	2.4820	3.8610	360.576
ADE	Alert	3.520	4.0115	4.4390	5.0170	5.981
ADE	Danger	6.109	7.1960	8.9500	13.1860	387.585
ANDE	Normal	0.124	1.6470	2.6470	4.2175	130.573
ANDE	Alert	3.502	3.9920	4.3800	5.1820	5.945
ANDE	Danger	6.170	7.2780	7.8460	9.2160	10.590
HDE	Normal	0.114	1.4990	2.6040	4.6190	500.000
HDE	Alert	3.593	4.1020	4.7750	5.3330	5.904
HDE	Danger	6.034	7.3260	9.7030	12.2650	500.000
HDE.CAV	Normal	0.306	1.6265	2.5590	4.9135	168.149
HDE.CAV	Alert	3.558	3.9045	4.6820	5.0195	5.612
HDE.CAV	Danger	6.367	7.6860	8.7600	11.7530	22.279
HDE.ENV	Normal	0.000	0.0360	0.0730	0.1555	6.916
HNDE	Normal	0.124	1.6580	2.9770	5.4670	500.000
HNDE	Alert	3.518	4.0455	4.6795	5.2590	5.979
HNDE	Danger	6.096	8.0330	9.4090	13.7420	271.508
HNDE.ENV	Normal	0.001	0.0340	0.0660	0.1515	33.502
HNDE.ENV	Alert	4.991	4.9910	4.9910	4.9910	4.991
HNDE.FL	Normal	0.004	1.2870	2.4050	4.2690	52.019
HNDE.FL	Alert	3.557	4.3220	4.6890	5.3610	5.995
HNDE.FL	Danger	6.014	6.8990	10.0060	13.5650	62.193
VDE	Normal	0.131	1.3200	2.2180	3.9440	500.000
VDE	Alert	3.514	3.9970	4.4900	5.3740	5.996
VDE	Danger	6.014	7.3660	10.9775	16.0080	500.000
VNDE	Normal	0.061	1.3150	2.1445	3.6235	500.000
VNDE	Alert	3.531	3.9395	4.4160	4.9560	5.797
VNDE	Danger	6.140	7.3270	9.2060	14.0870	35.022

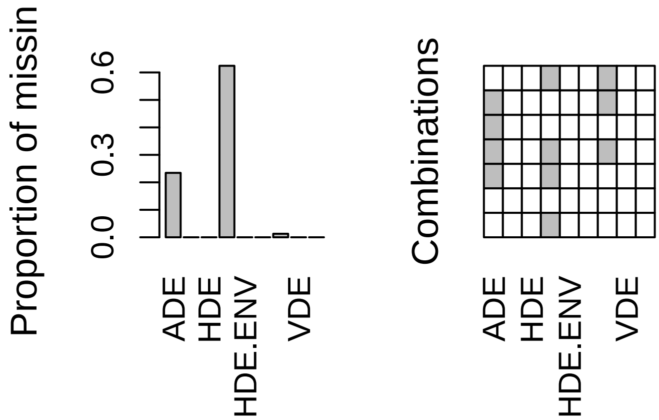

Figure 5.1: Summary of the missing values for vibration variables for fan data. The left plot indicates the proportion of missing values by variable. The right plot indicates the combinations of missing(grey) and non-missing values along with their relative proportions.

Table 5.4: Component loadings for the first 4 principal components using vibration measurements

Variable	PC1	PC2	PC3	PC4
ADE	-0.492	0.144	0.256	0.213
ANDE	-0.455	0.076	0.448	0.199
HDE.CAV	-0.231	-0.175	0.08	-0.703
HDE.ENV	-0.203	-0.53	-0.217	0.185
HDE	-0.184	-0.389	-0.199	-0.379
HNDE.ENV	-0.252	0.396	-0.624	-0.097
HNDE	-0.148	-0.436	-0.339	0.468
VDE	-0.441	0.351	-0.31	0
VNDE	-0.378	-0.201	0.202	-0.12
Cum. Var.	37%	58%	71%	80%

Note:
Cum. Var. represents the percentage of cumulative variance explained by the principal components. HNDE.FL is not included as it is not measured on fans

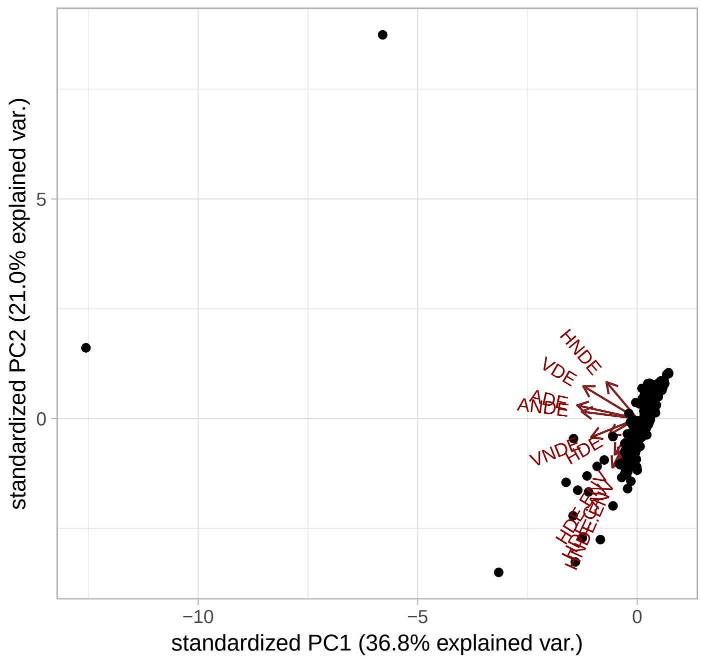

Figure 5.2: Biplot of the fan vibration PCA

Table 5.5: Component loadings for the first 4 principal components using vibration measurements

Variable	PC1	PC2	PC3	PC4
ADE	-0.41	-0.166	0.326	-0.149
ANDE	-0.149	0.596	0.412	-0.085
HDE.CAV	-0.231	0.071	-0.43	0.656
HDE.ENV	-0.463	-0.187	0.097	-0.193
HDE	-0.305	-0.071	-0.323	0.14
HNDE.ENV	0.004	0.36	-0.626	-0.632
HNDE	-0.315	-0.506	-0.152	-0.267
VDE	-0.441	0.17	0.073	0.017
VNDE	-0.395	0.398	-0.075	0.109
Cum. Var.	34%	50%	62%	72%

Note:
Cum. Var. represents the percentage of cumulative variance explained by the principal components.
HNDE.FL is not included as it is not measured on fans

Table 5.6: Universal part-of-speech abbreviations and their meanings

Abbreviation	Meaning
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

References

Aggarwal, Charu C. 2015. *Data Mining: The Textbook.* Cham: Springer International Publishing : Imprint: Springer.

Barszcz, Tomasz. 2019. *Vibration-Based Condition Monitoring of Wind Turbines.* Vol. 14. Applied Condition Monitoring. Cham: Springer International Publishing.

Bishop, Christopher M. 1996. *Neural Networks for Pattern Recognition.* Repr. Oxford: Clarendon.

Blockeel, Hendrik. 2016. *Machine Learning and Inductive Inference.* 5de, herz. uitg., herdr. Leuven: Acco.

Choudhary, Devendra, Mayank Tripathi, and Ravi Shankar. 2019. "Reliability, Availability and Maintainability Analysis of a Cement Plant: A Case Study." *The International Journal of Quality & Reliability Management* 36 (3): 298–313. http://search.proquest.com/docview/2193101527/.

Cleves, William A, William W Gould, Roberto G Gutierrez, and Yulia Marchenko. 2010. *An Introduction to Survival Analysis Using Stata.* 3rd edition. Texas: Stata Corporation.

Coetzee, Jasper L. 1997. "The Role of Nhpp Models in the Practical Analysis of Maintenance Failure Data." *Reliability Engineering & System Safety* 56 (2): 161–68. https://doi.org/https://doi.org/10.1016/S0951-8320(97)00010-0.

Collett, D. 2003. *Modelling Survival Data in Medical Research, Second Edition.* Chapman & Hall/Crc Texts in Statistical Science. Taylor & Francis. https://books.google.be/books?id=Hm/_FfYRHCuAC.

Cox, D. R, and D Oakes. 1998. *Analysis of Survival Data.* Repr. Monographs on Statistics and Applied Probability 21. Boca Raton: Chapman; Hall/CRC.

Elvebakk, G, Bo Lindqvist, and K Heggl. 1999. "The Trend-Renewal Process for Statistical Analysis of Repairable Systems." *Technometrics* 45 (September). https://doi.org/10.1198/004017002188618671.

Everitt, B., and T. Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R.* Use R! Springer New York. https://books.google.be/books?id=BB51TWe94EwC.

Fridolin, Wild. 2019. "CRAN Task View: Natural Language Processing." https://cran.r-project.org/view=NaturalLanguageProcessing.

Gamiz, Maria Luz, and Bo Henry Lindqvist. 2016. "Nonparametric Estimation in Trend-Renewal Processes." *Reliability Engineering & System Safety* 145: 38–46. https://doi.org/https://doi.org/10.1016/j.ress.2015.08.015.

Hartler, G. 1989. "The Nonhomogeneous Poisson Process - a Model for the Reliability of Complex Repairable Systems." *Microelectronics Reliability* 29 (3): 381–86. https://doi.org/https://doi.org/10.1016/0026-2714(89)90624-0.

"ISO 13373-1:2002 (Vibration Condition Monitoring - Part 1:General Procedures)." 2002. Standard. Vol. 2002. Geneva, CH: International Organization for Standardization.

"ISO 13373-2:2016 (Vibration Condition Monitoring - Part 2:Processing, Analysis and Presentation of Vibration Data)." 2016. Standard. Vol. 2016. Geneva, CH: International Organization for Standardization.

"ISO 14224:2016 (Collection and Exchange of Reliability and Maintenance Data for Equipment)." 2016. Standard. Vol. 2016. Geneva, CH: International Organization for Standardization.

Jardine, A. K. S., P. M. Anderson, and D. S. Mann. n.d. "Application of the Weibull Proportional Hazards Model to Aircraft and Marine Engine Failure Data." *Quality and Reliability Engineering International* 3 (2): 77–82. https://doi.org/10.1002/qre.4680030204.

Jardine, A.K.S., and A.H.C. Tsang. 2013. *Maintenance, Replacement, and Reliability: Theory and Applications, Second Edition*. Mechanical Engineering. Taylor & Francis. https://books.google.be/books?id=6sgh7iIQxboC.

Jardine, Andrew K.S., Daming Lin, and Dragan Banjevic. 2006. "A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance." *Mechanical Systems and Signal Processing* 20 (7): 1483–1510. https://doi.org/https://doi.org/10.1016/j.ymssp.2005.09.012.

Josse, Julie, Nicholas Tierney, and Nathalie Vialaneix. n.d. "CRAN Task View: Missing Data." https://cran.r-project.org/view=MissingData.

Kijima, Masaaki. 1989. "Some Results for Repairable Systems with General Repair." *Journal of Applied Probability* 26 (1): 89–102. http://www.jstor.org/stable/3214319.

Kleinbaum, David G., Mitchel Klein, and J Samet. 2006. *Survival Analysis : A Self-Learning Text*. New York: Springer.

Kutner, Michael H, William Li, Christopher J Nachtsheim, and John Neter. 2005. *Applied Linear Statistical Models*. 5th edition. Boston: McGraw-Hill/Irwin.

Lawless, Jerald F. 2007. *The Statistical Analysis of Recurrent Events*. Statistics for Biology and Health. New York, NY: Springer New York.

Lengerich, Eugene. n.d. "STAT 507: Epidemiological Research Methods." https://newonlinecourses.science.psu.edu/stat507/node/81/.

Lindqvist, Bo Henry. 2006. "On the Statistical Modeling and Analysis of Repairable Systems." *Statist. Sci.* 21 (4): 532–51. https://doi.org/10.1214/088342306000000448.

Liu, Enwu, and Karen Lim. 2018. "Using the Weibull Accelerated Failure Time Regression Model to Predict Time to Health Events." *bioRxiv*. https://doi.org/10.1101/362186.

Moore, Dirk. 2016. *Applied Survival Analysis Using R*. https://doi.org/10.1007/978-3-319-31245-3.

Muchiri, Peter, Liliane Pintelon, Harry Martin, and Peter Chemweno. 2013. "Modelling Maintenance Effects on Manufacturing Equipment Performance: Results from Simulation Analysis." *International Journal of Production Research* 52 (November): 3287–3302. https://doi.org/10.1080/00207543.2013.870673.

Nystad, Bent Helge, and Magnus Rasmussen. 2010. "Remaining Useful Life of Natural Gas Export Compressors." *Journal of Quality in Maintenance Engineering* 16 (2): 129–43. https://doi.org/10.1108/13552511011048887.

Pintelon, Liliane, and Alejandro Parodi-Herz. 2008. "Maintenance: An Evolutionary Perspective." In *Complex System Maintenance Handbook*, 21–48. London: Springer London. https://doi.org/10.1007/978-1-84800-011-7_2.

Rausand, M., and A. Høyland. 2004. *System Reliability Theory: Models, Statistical Methods, and Applications.* Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley. https://books.google.be/books?id=gkUWz9AA-QEC.

Rencher, A.C. 2002. *Methods of Multivariate Analysis.* Wiley Series in Probability and Statistics. Wiley. https://books.google.be/books?id=4xGlNAEACAAJ.

Robinson, David, and Julia Silge. 2017. *Text Mining with R: A Tidy Approach.* O'Reilly Media. http://shop.oreilly.com/product/0636920067153.do.

Ruijters, Enno Jozef Johannes, Dennis Guck, Peter Drolenga, Margot Peters, Mariëlle Ida Antoinette Stoelinga, Gul Agha, and Benny Van Houdt. 2016. "Maintenance Analysis and Optimization via Statistical Model Checking: Evaluating a Train Pneumatic Compressor." *Proceedings of the 13th International Conference on Quantitative Evaluation of SysTems, QEST 2016*, 331–47.

Salameh, Jack P., Sebastien Cauet, Erik Etien, Anas Sakout, and Laurent Rambault. 2018. "Gearbox Condition Monitoring in Wind Turbines: A Review." *Mechanical Systems and Signal Processing* 111: 251–64.

Sharma, Subhash. 1996. *Applied Multivariate Techniques.* New York: Wiley.

Swiniarski, Roman W, Witold Pedrycz, and Lukasz A Kurgan. 2007. *Data Mining: A Knowledge Discovery Approach.* Boston, MA: Springer US.

Tanwar, Monika, Rajiv N. Rai, and Nomesh Bolia. 2014. "Imperfect Repair Modeling Using Kijima Type Generalized Renewal Process." *Reliability Engineering & System Safety* 124: 24–31. https://doi.org/https://doi.org/10.1016/j.ress.2013.10.007.

Therneau, Patricia M., Terry M.; Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Wakiru, James M., Liliane Pintelon, Peter N. Muchiri, and Peter K. Chemweno. 2019. "A Review on Lubricant Condition Monitoring Information Analysis for Maintenance Decision Support." *Mechanical Systems and Signal Processing* 118: 108–32. https://doi.org/https://doi.org/10.1016/j.ymssp.2018.08.039.

Wu, Shaomin, and Philip Scarf. 2017. "Two New Stochastic Models of the Failure Process of a Series System." *European Journal of Operational Research* 257 (3): 763–72. https://doi.org/https://doi.org/10.1016/j.ejor.2016.07.052.

Yanez, Medardo, Francisco Joglar, and Mohammad Modarres. 2002. "Generalized Renewal Process for Analysis of Repairable Systems with Limited Failure Experience." *Reli-*

ability Engineering & System Safety 77 (2): 167–80. https://doi.org/https://doi.org/10.1016/S0951-8320(02)00044-3.

Yang, Dong, Hui Li, Yaogang Hu, Jie Zhao, Hongwei Xiao, and Yongsen Lan. 2016. "Vibration Condition Monitoring System for Wind Turbine Bearings Based on Noise Suppression with Multi-Point Data Fusion." *Renewable Energy* 92: 104–16.

Zacks, Shelemyahu. 2012. *Introduction to Reliability Analysis: Probability Models and Statistical Methods*. Springer. https://doi.org/https://doi.org/10.1007/978-1-4612-2854-7.

AFDEL
Straat nr bus (
3000 LEUVEN, BEL
tel. + 32 16 00 0
fax + 32 16 00 0
www.kuleuven.be