Quantitative Risk Analysis using Real-time Data and Change-point Analysis for Data-informed Risk Prediction

George Jordan

A thesis submitted in partial fulfilment of the requirements of the University of East London for Degree of Professional Doctorate in Data Science

July 2019

Abstract

Incidents in highly hazardous process industries (HHPI) are a major concern for various stakeholders due to the impact on human lives, environment, and potentially huge financial losses. Because process activities, location and products are unique, risk analysis techniques applied in the HHPI has evolved over the years. Unfortunately, some limitations of the various quantitative risk analysis (QRA) method currently employed means alternative or more improved methods are required. This research has obtained one such method called Big Data QRA Method.

This method relies entirely on big data techniques and real-time process data to identify the point at which process risk is imminent and provide the extent of contribution of other components interacting up to the time index of the risk. Unlike the existing QRA methods which are static and based on unvalidated assumptions and data from single case studies, the big data method is dynamic and can be applied to most process systems. This alternative method is my original contribution to science and the practice of risk analysis

The detailed procedure which has been provided in Chapter 9 of this thesis applies multiple change-point analysis and other big data techniques like, (a) time series analysis, (b) data exploration and compression techniques, (c) decision tree modelling, (d) linear regression modelling. Since the distributional properties of process data can change over time, the big data approach was found to be more appropriate. Considering the unique conditions, activities and the process systems use within the HHPI, the dust fire and explosion incidents at the Imperial Sugar Factory and the New England Wood Pellet LLC both of which occurred in the USA were found to be suitable case histories to use as a guide for evaluation of data in this research.

Data analysis was performed using open source software packages in R Studio. Based on the investigation, multiple-change-point analysis packages *strucchange* and *changepoint* were found to be successful at detecting early signs of deteriorating conditions of component in process equipment and the main process risk. One such process component is a bearing which was suspected as the source of ignition which led to the dust fire and explosion at the Imperial Sugar Factory.

As a result, this this research applies the big data QRA method procedure to bearing vibration data to predict early deterioration of bearings and final period when the bearing's performance begins the final phase of deterioration to failure. Model-based identification of these periods provides an indication of whether the conditions of a mechanical part in process equipment at a particular moment represent an unacceptable risk.

The procedure starts with selection of process operation data based on the findings of an incident investigation report on the case history of a known process incident. As the defining components of risk, both the frequency and consequences associated with the risk were obtained from the incident investigation reports. Acceptance criteria for the risk can be applied to the periods between the risks detected by the two change-point packages. The method was validated with two case study datasets to demonstrate its applicability as procedure for QRA. The procedure was then tested with two other case study datasets as examples of its application as a QRA method. The insight obtained from the validation and the applied examples led to the conclusion that big data techniques can be applied to real-time process data for risk assessment in the HHPI.

Acknowledgement

I blame all of you because the last five years of my life as a doctoral student has been a period of sustained challenges and suffering. One day, readers of this thesis may, perhaps, not consider how much some of you played a bigger role in steering me through the storms with your encouragement and support. I believe you know who you are, but I owe you for this work, which is my original contribution to science and the practice of risk analysis because as Harvey Mackay explains, “None of us got to where we are alone. Whether the assistance we received was obvious or subtle, acknowledging someone's help is a big part of understanding the importance of saying thank you.”

I thank my supervisor, director of studies, and mentor, Professor Allan Brimicombe, who was always available when I need support. His mentorship and continuous support make me feel indebted to the world and hope to match someday. I thank Dr. Yang Li, who was my co-supervisor and help steer me through the challenges when required.

My sincere appreciation goes to my young but supportive family Cynthia, Kelton and Jheylane Jordan, who went through the sustained suffering as a result of my limited availability because of my studies. The odd hours of working on my studies, coupled with the challenges that the family went through supporting another member who was terminally ill, was a testimony that as a species we are coded with the DNA to help one another. I dedicate this thesis to you and the memories of little Phelan Jordan who gave us the lesson to never give up until our missions are accomplished.

My sincere thanks go to the leadership and professionals at Chilworth Technology who provided some of the funding towards my studies and gave my family the necessary support during the most difficult years when little Phelan was going through treatments in hospital. I also extend my acknowledgments my bereavement counsellor Dawn Robinson whose support over the last two years has help steady my ship and always make made me aware that there is hope after my loss.

Special thanks to my auntie Ms. Thelma Lomley, brother Otis Okine, friends Leslie Mensah and Peter Bosompem whose precious support has been with me throughout this studies, and Miss Rita Afiademnyo whose occasional criticisms, particularly when I switch into self-distractions, gave me the realisation that the mission can be accomplished. Well, … I want you all to know that you owe me for the pains I had to endure over the five years of my doctoral studies.

Please note that no editor has been mentioned in this acknowledgement and therefore I am responsible entirely for the writing, drafting and editing of my thesis. Thus, an editor has not been used in the construction of any part of this thesis.

To the loving memories of Master Phelan Louis Nii Omanye Jordan (our little Blue Ninja)

Contents

List of Figures

List of Tables

Chapter 1 – Introduction to Research

1.0. Introduction

Accidents in many process industries are rare but generally results in high consequence such as explosions, releases of highly toxic materials and major fires. One such incident is the catastrophic explosion of a pressure vessel at the Loy Lange Box Company (USA), which killed four people, left another in critical condition and destroyed the properties of two nearby companies (CSB, 2017). Due to the variability of the processes undertaken by these industries and the hazardous nature of materials the use, operations within these industries must be monitored and managed to ensure that occurrences of such incidents are minimized to help protect lives and properties.

The cause of these accidents also varies from type of equipment's, operational failures or external factors. Various compliance requirements have been put in place to monitor the activities of these industries. One such legislation is the Control of Major Accident Hazards Regulations which focus on the risk of incidents within process industries in the UK (Anderson, 2005). This regulation ensures that process industries take all measures to manage and minimize major accident hazards from their activities.

Some available incident data suggests that the incidents in these industries occur either by direct causes which happen immediately prior to an undesirable event or further away in time or space with most of the incidents suspected to be the fault of frontline staff/operators (CSB, 2006). The potential for process activities to create a major risk which disruption the operation of the process, cause fatalities, and disrupts the socio-economic activities in areas affected by the incident necessitates the need to have a system that could detect and/or predict the causality of the various risks to help reduce the devastating effect of hazards within the process. As a result, risk analysis is required.

Risk analyses has been widely applied in many industries to help improve safety as part of the process design and operation. Quantitative risk analysis (QRA) is one of the risk analysis tools applied to manage safety as part of the licensing or operation activities. Due to the importance of QRA in many application areas (Marhavilas et al. 2011) to management and process operators, engineering and safety professionals (Saleh & Pendley 2012; Wybo & Van Wassenhove 2016), QRA is included as a tool for risk analysis in many reference publications, textbooks (Goerlandt, Khakzad & Reniers, 2017) and considered an important topic to teach in many discipline areas.

Some of the areas where QRA is applied as a risk analysis tool includes, nuclear installations, offshore oil and gas installations, transportation platforms (road, air and maritime), manufacturing and chemical installations, various security establishments (including cyber security), and

environmental related areas such as construction industry. The application of QRA in these areas has been extensively reviewed by previous researchers (Vinnem 1998; Garrick & Christie 2002; Li et al. 2012; Pasman & Reniers 2014; Taroun 2014; Khan et al. 2015, Cherdantseva et al. 2016; Goerlandt, Khakzad & Reniers 2017).

On a more fundamental level, some authors have raised issues about some of the challenges of application of QRA to detect risk in process industries. Thus, although a powerful tool, some of the QRA methods used in industries has numerous pitfalls which must be carefully considered before applying for risk assessment. Example of these pitfalls includes data quality and data size (Kerin 2017). These pitfalls mean that application of the QRA methods have the potential of providing misleading outcomes especially in situations where there are insufficient samples and series of assumptions are made to establish the events-time relationships.

Despite some of the success of the existing QRA methods, there is evidence of major accidents within the industry which means that some of the assumptions made by the peers of experts for QRA in situations of limited data may be unreliable (Apostolakis 2004). It could therefore be inferred that some of the reported case histories of process safety incidents may be due to some of the pitfalls of QRA or incorrect risk assessments (Russ 2010).

To help overcome some of the pitfalls of insufficient data, some facilities within the industry applies industry-standard software packages which relies on data from incident databases to perform QRA's (Pariyani et al. 2010). However, some research have also suggested that even the data on most industrial incidents on records (e.g., government agency investigations, technical literature) have limitations due to lack of data on some less severe and near-miss incidents which are usually missing from the records (CSB, 2003). As a result, most incident reports could be underestimated due to the insufficient or missing data. Even in facilities where major incidents have occurred, there is evidence of substantial variability in the data of the events that led to the incident. Thus, applying industry-standard packages for QRA also have pitfalls.

Several publications by expert groups (e.g. CPS 2000, CSChE 2004) have also discuss the effect of limited data on QRA. For instance, whereas enough data could be obtained for a QRA from matured process facilities, this cannot be the case for newly operated or less matured facilities (CPS 2000). As a result, the CSChE Guide 2004 incorporated other relevant guidelines, including the Dutch Guide 1988 (CSChE 2004), which suggests that where possible the QRA process must include,

- the use of hazardous substances from the process facilities for hazard identification, consequence, and frequency analysis,

- individual risk profile (i.e. risk versus distance from risk source) must have two-point approximation for risk the estimation, and
- risk evaluation component based on individual risks at levels "allowable for different levels" of population density.

Considering advances in the technology applied within the process facilities, it is obvious that each process generates abundance of data. Thus, the lark of sufficient data for QRA may be due to the industry not being data ready. Research in QRA is therefore faced with the challenges of conducting risk analysis using the existing QRA methods to catch-up with the complex but expanding process systems. Due to dynamic conditions of these facilities and the limited data use for QRA, it is becoming extremely difficult to directly detect process risk. As a result, most of the existing QRA methods depend on assumptions made by peers of experts in safety and are usually validated by single case study research (Yildiza, Dikmen & Birgonul 2014).

Even where enough data is available, the in-depth understanding of trends in the data which is of immense importance for detecting the risk in the process for proactive strategies to be adapted to mitigate the risks is larking. Thus, a good insight into the trend and distribution of the data is of vital importance for any meaningful development of risk and risk mitigation within the industry. This may involve investigating and applying QRA method which incorporates sufficient data. This calls for a new approach to QRA which involves the use of large set of data like data from the process operation itself

In light of the above, this study provides an alternative QRA method called big data QRA method, which relies entirely on big data techniques and real-time as a major contribution to science and practice, provides a detail procedure for its application. The study selects change-point analysis as the preferred big data technique to detect risk in process systems, after other big data methods like exploratory data analysis, feature extraction, has been applied to the process monitoring data. This is because changepoint analysis has been applied for time series datasets for research over the years.

The concept of change-point analysis has its origin in the article first published by E.S. Page in 1954 with a focus on sequential detection of a change in the trend in process quality control data in manufacturing process (Page 1963). The methods for the detection of change-points can be sequential (online), where the analysis may be performed on as data becomes available to identify abrupt changes near the most resent observation, or (b) retrospective (offline), where one off analysis may be performed on historical time series dataset (Page 1963).

Also, for the QRA method to be deem successful it must be able to (a) handle data of different dimensions, (b) detect single and multiple risk events and (c) have the ability to handle historical

and real-time datasets (Aminikhanghahi & Cook 2017; Fisher & Jensen 2018). Detail discussion of the method procedure which is a major contribution to science and practice are provided in Part 3-Methodology, where the study investigates big data techniques for the big data QRA method and using one of the available datasets and provide the detail procedure.

1.1. Theories

The study will consider various complementary theories as part of selecting appropriate theory for the thesis. The theories considered includes (a) Risk theories; (b) Behavioural theories; (c) System and operation theories; and (d) Organisation theory. The appropriate theory will be selected after careful evaluation of these theories and based on consideration of the process of managing risk within process systems, with emphasise that each process is unique with specifications based on factors including the target market, geographical locations of the process/users of the products, and therefore could have its own unique characteristics and management. Detail investigation of the theories will be provided in Part 1 -Risk.

1.2. Evolution of Ideas

The study provides some clarification on this research to help provide clarity to the readers of this thesis. This thesis concerns quantitative risk analysis, focusing on the use of big data techniques and real-time process monitoring data. It explores this topic within the context of the highly hazardous process industries (HHPI), this being industries who use chemicals which may be toxic, flammable, combustible, reactive or a combination of these properties. Ideas for the study was initiated by acquisition and merger of two companies which specialises safety in 2013 companies. One of these companies, which is UK based, specialises in process safety. The second company is a USA based company which specialises in behavioural safety. The event provoked the thoughts about whether the focus of risk should be on human behavioural elements at the process site or on the process itself.

The thesis therefore adopts interdisciplinary approach to present the varied means by which process risk is managed by starting with the concept of risk and its theories. This is followed by sourcing peer reviewed research publication on QRA method as applied within the target industry regarding the focus topic and analysing them to highlight gaps. This then proceeds to real-life case histories of process safety (PS) incidents and critiquing the corresponding incident investigation reports to ascertain whether available dataset could be used for the research.

This then goes on methodology where the research explores how to fill the gaps by considering various theoretical and conceptual framework to select appropriate theory for the research. This is followed by investigating various big data techniques and software packages to obtaining the Big

Data QRA method and provide a step-by-step procedure to make it easier to apply for risk analysis.

Following this is the fourth part of the thesis where the procedure was will is validated and tested to check its validity and reliability. It then concludes that QRA can be performed using a procedure which rely entirely on big data techniques and real-time process monitoring data. The overall process is represented by the flowchart of Figure 1.

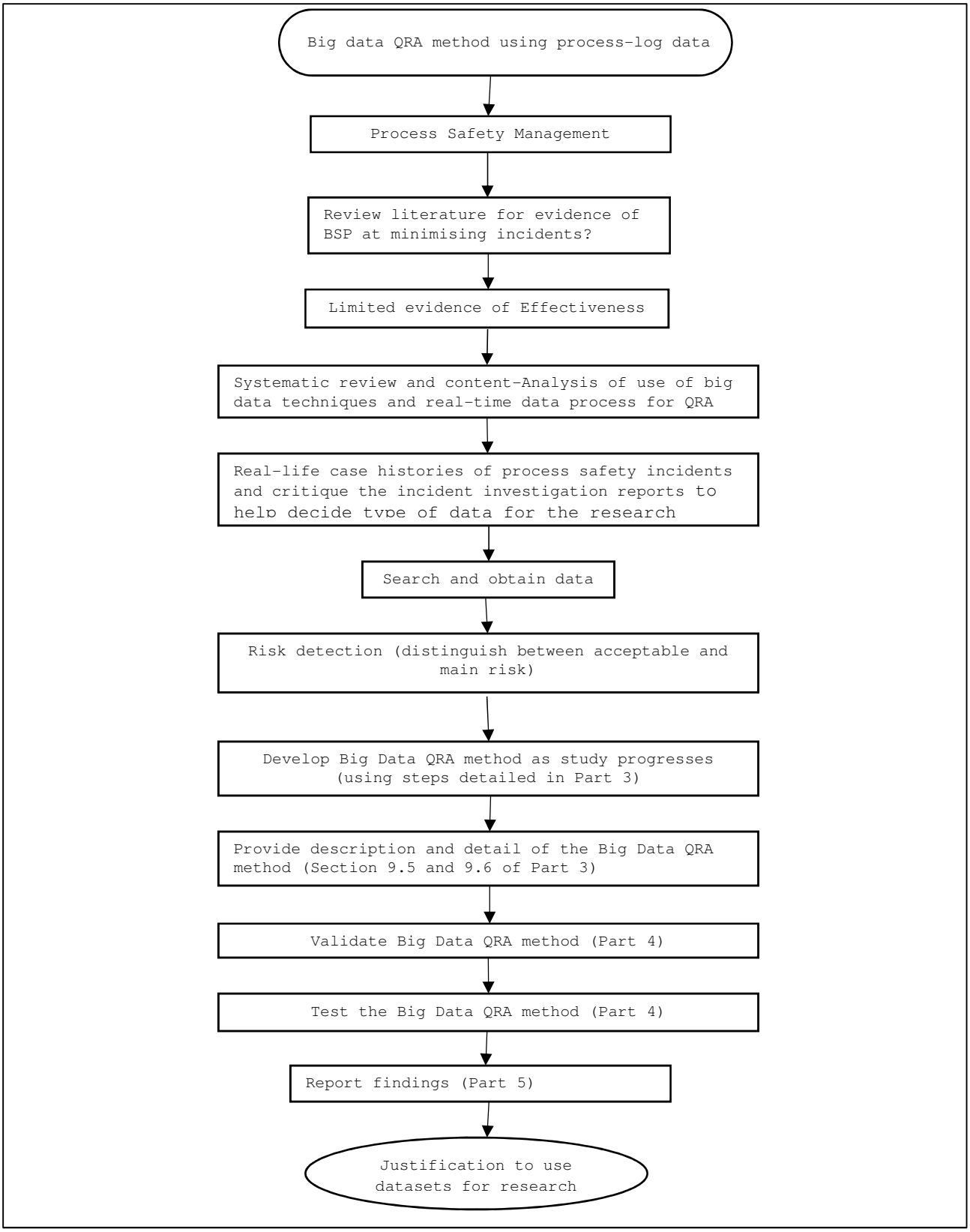


Figure 1: Flowchart illustrating how ideas evolved

1.3. Aims and objectives

This study aims to provide a big data QRA method and detail step-by-step procedure as a major contribution to science and practice. This method will rely entirely on big data techniques and real-time process monitoring data. The objectives of this study include

- Consider risk in process safety management (PSM) as a concept, explain some of the terminologies of risk in the context of their usage in the research, and provide clarification between risk assessment and risk analysis by presenting quantitative risk analysis (QRA) as a method and distinguishing QRA from other forms of risk analysis procedures involving the use of numerical data (e.g. Preventative maintenance).
- Perform a review of published literature on behavioural safety programs (BSP) and QRA methods as applied for managing risks in the HHPIs to ascertain whether the focus of safety should be on the behaviour of frontline personnel of on the process itself.
- Present real-life case histories of process safety incidents, and basic concept of dust fire and explosions with definitions of measurable properties of dusts to provide clarity of their usage in the thesis. The study will also critique the final reports on the incident investigation and its findings to determine whether available data sets could be applied for the research study.
- Consider and select big data techniques, PC software packages to apply, challenges associated with the use of a big data QRA method for risk analysis and how they can be resolved.
- Develop, validate and test the big data QRA method using available datasets, and provide detail procedure of the method.

1.4. Research Structure

The structure for the remainder of the thesis consists of the following:

Part 1- Risk as a Concept, Definition of Key Terms and Theoretical Framework, where the study introduces risk as a concept, provides definition of key generic terms in the context of their usage and select a theoretical framework for the thesis.

Part 2- Literature Review and Systematic Content-analysis, Real-life Case Histories Process Safety Incidents and Data, where the study reviews scientific research publication on QRA methodologies applied in HHPI's with emphasis on application of big data techniques and use of real-time data as a means of exploring any gaps in science and practice before commencing the research. The study also discusses real-time case histories of process safety incidents and data available for the research.

Part 3- Methodology, where the study investigates various big data techniques and software packages leading to obtaining the big data QRA method and presents the detailed step-by-step method procedure.

Part 4- Data Analysis, where the study validates the big data QRA method and test its applicability for risk analysis.

Part 5- Discussion, Conclusion and Recommendation, where general discussion of the research findings in relation to the research objectives and questions, contributions of the research to science and practice, the limitations of the research, suggestions for future research, and conclusion of the research are provided.

1.5. Conclusion

This chapter has introduced the research as well as the aims and objectives. Outline of the research structure has also been provided. Due to the scope that the research may cover, the following sub-questions base on unbiased opinion was applied to help answer the research question.

- How is the knowledge about risk in the HHPIs utilized and which of the risk theories can the study adopt for this research?
- Is behavioural safety programs (BSPs) more effective at preventing major accidents in the HHPIs for the focus of risk to be on monitoring human elements instead of monitoring the process itself?
- What are the existing QRA methods applied in the HHPIs and what are some of the challenges encountered with their use?
- Is there any evidence of existing review and systematic content-analysis of published research on the use of big data techniques and real-time data for QRA in the HHPI?
- How are big data techniques and data from the process operation applied for QRA and what are some of the challenges to overcome and practical solutions?
- Can big data techniques and real-time data be applied to obtain an effective QRA method for use in the HHPI?

These questions may help identify any potential gaps in science and practice and/or close any relevance in the gaps. Answering the questions is therefore not a one-off activity, but part of the process cycle aimed at providing answers as the study progresses. It is also hoped that answering the questions may help adjust the study design were necessary in order to provide more insight into the study.

Next is Part 1- Risk Concept, Definition of Key Terms and Theoretical Framework, where the study introduces risk as a concept, provides definition of key generic terms in the context of their usage and select a theoretical framework for the thesis.

Part 1
Risk as a Concept, Definition of Key Terms and Theoretical Framework

Chapter 2 – Risk as a Concept

2.0. Introduction

In Chapter 1, the thesis presents the aim of the research as providing a QRA method which relies entirely on big data techniques and real-time process monitoring datasets as original contribution to science and the practice of risk analysis. This was followed by a brief presentation of an overview of the research. The study then presents the research objectives, the outline of the research, the scope that the research may cover, then the questions the research seeks to answer in order to arrive at the original contribution to science and practice. The chapter then provide the research structure as five parts (i.e. Parts 1 – 5).

In the light of the above, this chapter will introduce risk as a concept, provide definitions of key generic terms associated with risk in the context of their usage in the thesis and select appropriate theoretical framework for the research. The overall approach adopted commences with a discussion of process safety management, followed by a risk as a concept, and provides the definition of risk terminologies in the context of their use in the thesis. This will be followed by evaluation of conceptual and theoretical framework to aid the selection of appropriate framework for the research, then risk assessment and finally quantitative risk analysis. The general outline of the presentation of Part 1 is provided as Figure 2.

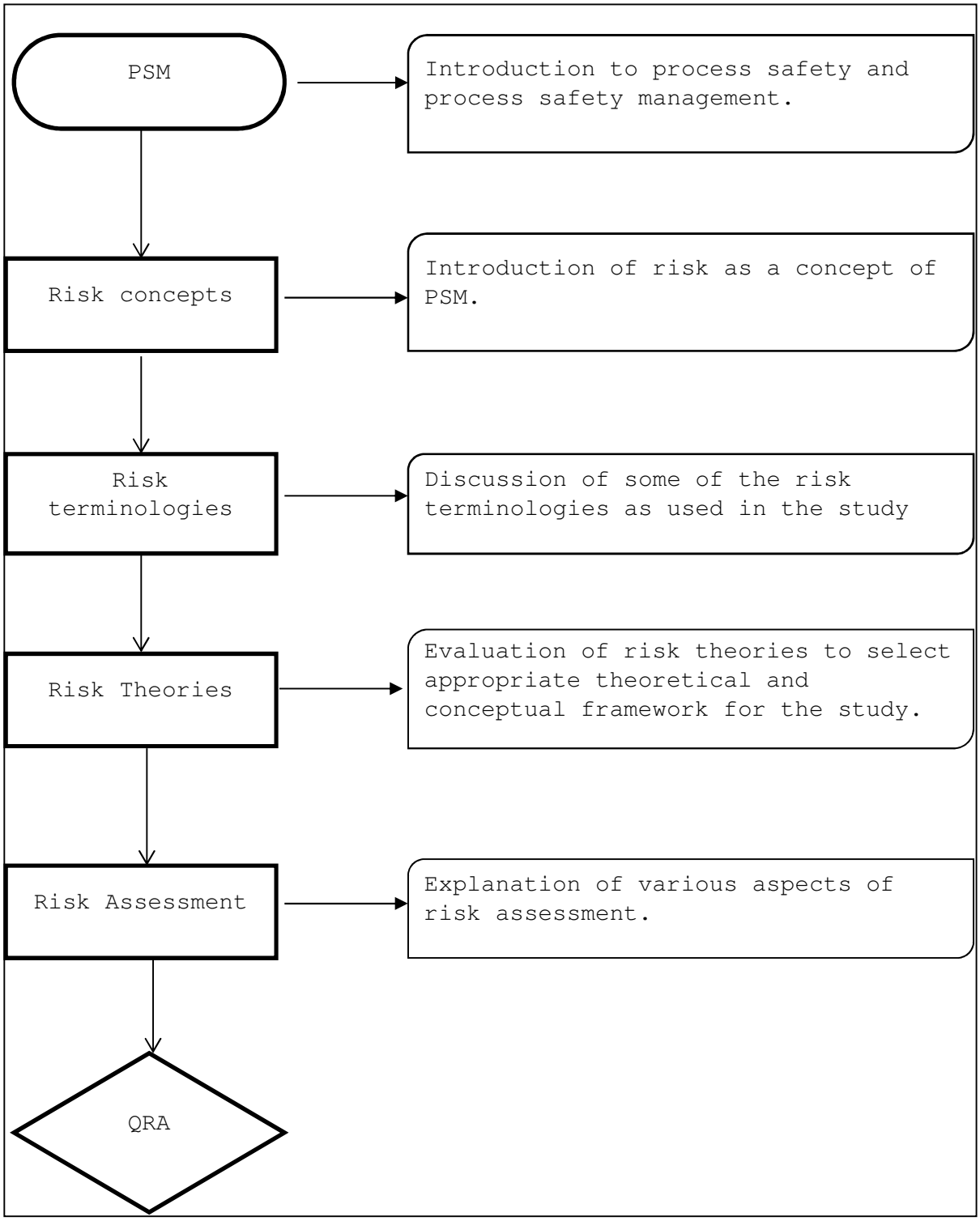


Figure 2. Flowchart illustrating the general outline of Part 1.

2.1. Process and Process Safety Management

The term 'process' in the context of this study can be described as a system of equipment's applied within the HHPIs which have controls and accessories and used as part of the overall processing within the plant. The term 'process system' in the context of this study, includes the component equipment's of the HHPI whose operation may involve the transport of raw materials,

intermediate or final products within the equipment. For instance, in a HHPI, a process system could be equipment's such as silos, bucket elevators, dryers, conveyor belts. As with every environment, these process equipment's are sometimes at risk with devastating consequences which may affect the facility and the surrounding population and properties. The risks must be managed to ensure safety of humans, properties, and the environment within and around the facility, hence the term process safety management (PSM).

PSM could be could be explain by an adapting a definition by the American Institute of Chemical Engineers' (AIChE) Centre for Chemical Process Safety (CPS), as a procedures for managing the reliability of operating systems and handling hazardous substances by applying good design principles, engineering, and operating practices which deals with the prevention and control of incidents that have the potential of a devastating consequences such as the release hazardous materials or high energy output such as toxic effects, fire, or explosion and could ultimately result in serious injuries, property damage, lost production, and environmental impact (CCPS 2010, p. xvii.). For the purpose of illustration, the underlying concept of risk management within the HHPI, this thesis has adapted a figure from ISO and Ghahramanzadeh (ISO 2009, p.14; Ghahramanzadeh 2013, p. 38) and presented as Figure 2.1.

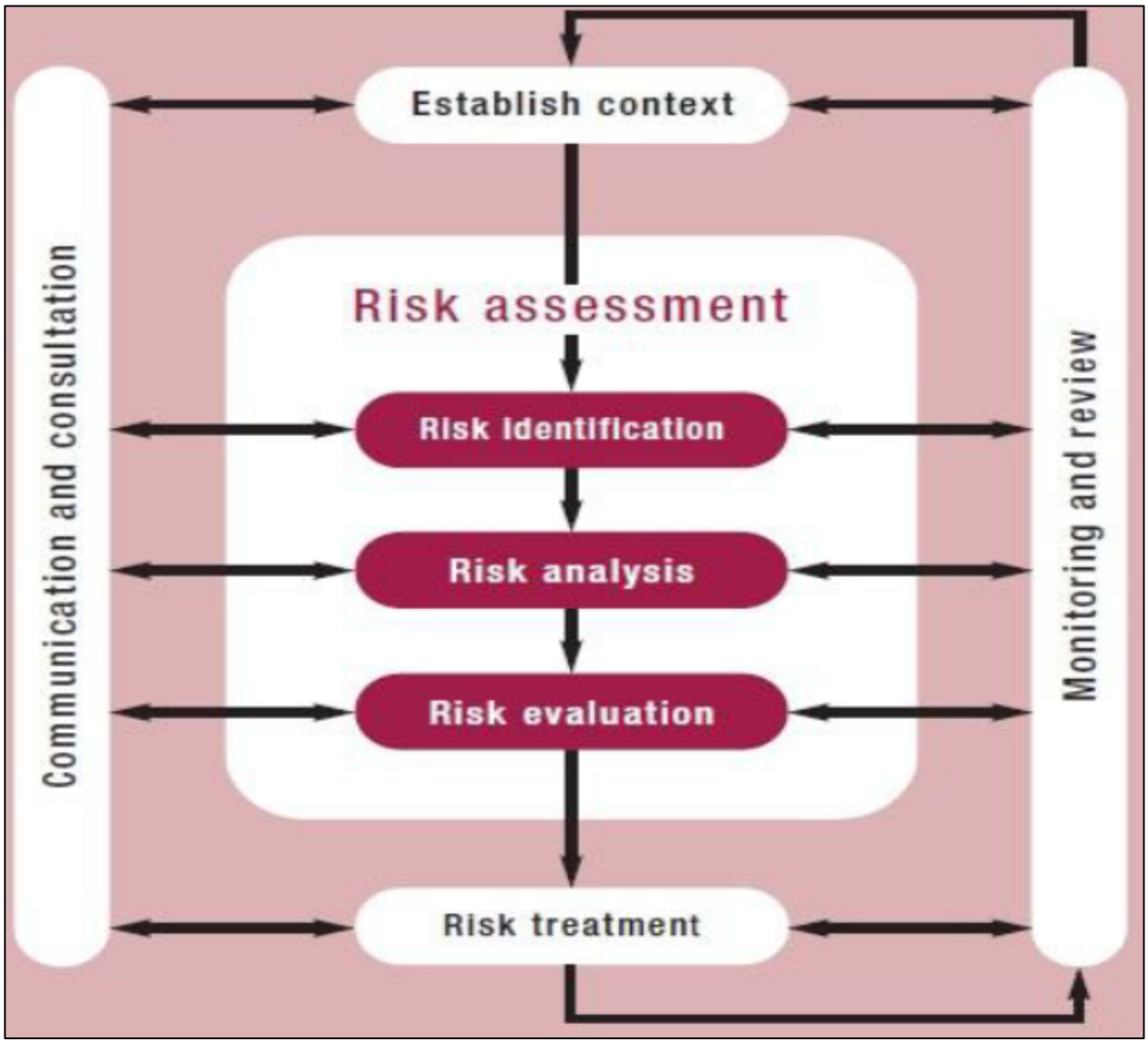


Figure 2.1: Risk management process (Source: Ghahramanzadeh 2013, p.38).

Process safety (PS) is sometimes used interchangeably with chemical safety or chemical process safety when dealing with protection against the toxic effects of chemicals. A typical process system is unique, in that its operations and the product or service could be different from the operations and products or services from other similar process systems. Examples of process systems include process such as sugar processing plant, oil refinery desulfurization plant and industrial chemical plant.

Typically, the process system applies financial, human, material and environmental resources for delivery of a specific product and services. As a result, different aspects of the process undertaken within the system must be managed properly in order to deliver the product and services. Due to its uniqueness, all information relating to a specific process system including safety need to be obtained for its safety management. Thus, every aspect of the process e.g. cost of running the process, the risks associated with the process operations must be determined and managed.

Because the process and its operations exist in an environment, other factors influenced by the environmental setting must be considered in PSM.

Over the years, research in PSM have shown that some process safety events can be catastrophic leading to the potential for loss of life, property, and environmental damage (Kerin 2017, p. 9). This study produces some PS incidents and incidents which had severe impact on the environment as Tables 2.1a and Table 2.1b respectively.

Thus, despite the advanced in technology and information, PS incidents are still with us. Due to the risk associated with their process operations, organisations have adopted interventions linking to safety and risks.

Table 2.1a: Some process safety incidents with related fatalities adapted from Kerin (2017)

Year	Location	Installation	Incident	Fatalities
1974	Flixborough, England	Chemical plant	Explosion	28
1977	Westwego, Louisiana, USA	Grain handling plant	Dust explosion	36
1984	San Juanico, Mexico City, Mexico	LPG terminal	Explosion, fire	>600
1984	Bhopal, India	Chemical plant	Toxic release	>3000
1988	Norco, Louisiana, USA	Refinery	Explosion	7
1988	Piper Alpha oilfield, North Sea	Upstream oil	Explosion, fire	167
1989	Pasadena, Texas, USA	Petrochemical	Explosion, fire	23
1992	LaMede, France	Refinery	Explosion	6
1992	Guadalaja, Mexico	Gas pipeline	Gas leak, sewer explosion	252
1998	Longford, Victoria, Australia	Gas processing	Explosion	2
2000	Mina Al-Ahmadi, Kuwait	Refinery	Explosion, fire	5
2001	Campos Basin, Brazil	Upstream oil	Explosions	11
2001	Toulouse, France	Chemical plant	Explosions	31
2003	Chongqing, China	Natural gas filed	Explosion, toxic release	243
2004	Skikda, Algeria	Gas processing	Explosion	27
2005	Texas City, Texas, USA	Refinery	Explosion	15
2005	Mumbai High North Field, India	Upstream oil and gas	Fire	22
2010	Macondo, Gulf of Mexico	Upstream oil	Explosion	11

Year	Location	Installation	Incident	Fatalities
2011	Laverton, Victoria, Australia	Chemical factory	Toxic release	1
2012	Paraguana Peninsula, Venezuela	Refinery	Explosion, fire	48
2014	Kunshun, Jiangsu, China	Metal products factory	Metal dust explosion	146
2015	Bay of Campeche, Gulf of Mexico	Upstream oil	Fire	4
2015	Tianjin, China	Chemical Storage	Explosion	173
2016	Gazipour, Bangladesh	Plastic packaging factory	Explosion	33
2017	Cambria, Wisconsin USA	Dry corn milling	Explosion and fire	5
2017	St. Louis, MO USA	Box Company	Explosion	1
2018	Pasadena, Texas USA	Chemical plant	Explosion and fire	-
2019	Waukegan, Illinois USA	Chemical plant	Explosion and fire	4
2019	Crosby, Texas USA	Chemical plant	Explosion	1

Table 2.1b: Some process safety incidents and their environmental impact (Source: Kerin 2017)

Year	Location	Installation	Incident	Environment Impact
1976	Seveso, Italy	Chemical plant	Chemical runaway reaction released 2,3,7,8-tetrachlorodibenzop- dioxin (TCDD)	Contamination of locally grown food, widespread death and emergency slaughtering of animals to prevent chemical entering the food chain.
1984	Bhopal, India	Chemical plant	Uncontrolled chemical reaction released methyl isocyanate gas and other chemicals	Broad-scale death of plants and animals created food shortages in the short term; long-term effects still impact plants, animals and people 30 years later.
1986	Chernobyl, Ukraine	Nuclear power	Overpressure led to steam explosion, fragmentation of fuel core and release of radiation	Contamination of the food chain resulted in a higher risk of cancer, death and reproductive loss in plant and animal populations up to 30 km from the site; strategies such as soil removal and exclusion zones were employed to mitigate the impact with the long-term effect determined by the half-life of the radionuclides; broader land contamination occurred with weather conditions and radioactive rainfall determining the level and range of contamination.
2009	Montara, Timor Sea	Upstream oil	Blowout and fire led to an oil spill that continued for 74 days, contaminating an estimated 90,000 km^2 of the Timor Sea	Oil and dispersants damaged coral and seaweed beds, impacting on fishing grounds with damage to mangroves putting villages at risk of flooding.
2010	Macondo, Gulf of Mexico	Upstream oil	Blowout of wellhead and release of an estimated 650 million litres of oil into Gulf of	Described as the “worst environmental disaster in American history” by the US Natural Resources Defence Council (NRDC), the oil and dispersants had a devastating impact on marine plants (including death of seaweed beds) animals

Year	Location	Installation	Incident	Environment Impact
			Mexico.	and birds, and severely impact fishing and tourism.
2011	Fukushima, Japan	Nuclear power	A tsunami resulting from an earthquake struck the coast, impacting the power plant resulting in a meltdown, and release of radiation across a large area.	Surrounding area remains highly radioactive, with some 160,000 evacuees still living in temporary housing; clean up estimated to take 40 years with some land unfarmable for centuries.
2014	Houston TX, USA		Release of highly toxic methyl mercaptan	Toxic chemicals released into the atmosphere.
2016	Mississippi, USA		Gas Plant Explosion and Fire	Release of toxic fumes into the atmosphere
2018	Oklahoma USA		Gas Well Blowout and Fire	Toxic gases and chemical decomposition products released into the atmosphere
2018	Superior, WI USA		Oil refinery fire and explosion	Release of highly toxic fumes into the atmosphere
2019	Philadelphia, PA, USA		Refinery Fire and Explosions	Release of toxic vapour on combustible products into the atmosphere
2019	Cosby, TX USA		Fatal fire and explosion at chemical plant	Release of highly toxic chemicals and combustion products into the atmosphere.
2019	Waukegan, IL. USA		Chemical release from manufacturing plant	Release of highly toxic chemicals into the atmosphere

There are various professional organisations and international associations that identifies and addresses issues relating to process safety around the world. They include American Institute of Chemical Engineers (AIChE), Institution of Chemical Engineers (IChemE), US Occupational Safety and Health Administration (OSHA), UK Health and Safety Executive (HSE), who aims at developing process safety professionals at identifying and addressing process safety requirements and provide guidelines for process safety management. These professional institutions provide their own standards and guidelines for managing different aspects of process safety. One such standard is the Guidelines for Process Safety Documentation (AIChE 1995) which details the elements of process safety.

The relevant literature on PSM has been reviewed (Amyotte & Lupien 2017). The 14 elements of process safety which are aimed at identifying best practices for process safety practitioners include

- Process Safety Information (PSI)
- Process Hazard Analysis
- Operating Procedures
- Training
- Contractors
- Mechanical Integrity
- Hot Work
- Management of Change
- Incident Investigation
- Compliance Audits
- Trade Secrets
- Employee Participation
- Pre-start up Safety Review, and
- Emergency Planning and Response.

Thus, keeping compliance in PSM requires documentation and other process safety information (PSI) which is also useful for relevant authorities including organisations which deals with emergencies such as the National Fire Protection, medical institutions and insurance providers (Kingsley & Kaelin 2012). Because each process is unique, the requirements of documentation may differ slightly.

The key documents (Kingsley & Kaelin 2012) include information on

- Process description

- Process flow diagram
- Piping and instrumentation drawing (P&ID)
- Electrical area classification drawing
- Process hazard analysis (PHA)
- Safety data sheets (SDS)
- Design basis for emergency systems and devices
- Start-up or shutdown operating procedures
- Normal operating procedures
- Emergency procedures
- Management-of-change procedure
- Maintenance records
- Other supporting documents (e.g. material and energy balance; process chemistry; materials of construction; equipment arrangement; plot plant; ventilation design; emergency planning; upper and lower control limits; consequence of process deviation; and accident/incident investigation reports).

Although PS incorporates some elements of occupational health and safety (OHS) in managing safety within an organisation with functional elements (IChemE 2015) like organisational culture, system procedures, knowledge and competency, as well as engineering design, they differ in areas like mechanisms of causation (Kerin 2017, p 4). This is because PS is not just focus on managing potential losses but losses which are usually associated with higher levels of energy and release of potentially toxic materials with devastating effect. PS incidents are less frequent as OHS incidents and focuses on engineering designs and the consequences failure in these designs are likely to be severe.

Due to the uniqueness of process activities, each process has its own unique characteristics which determine the PSM approach adopted and the way it is manage. Despite this, there are interdependent factors that leads to the production of the final products including process reactions and risks. Out of these factors, the study selects risk as a concept for discussion.

2.2. Risk

The term risk can be traced to two probable origins (a) the Italian word “risco” and (b) the Spanish word “riesco”, both derived from the Latin word “resecum” which refers to any dangers that threatens ships (Liuzzo et al, 2014). Risk may therefore be described as the chance that an unfavourable event could be caused by the presence of a hazard, leading to causing harm to humans or damage to property and the environment (CCPS, 2000 p. 6).

This description of risk has led to the general use of risk and hazard interchangeably. To provide useful distinction between risk and hazard, the HSE has provided a clearer description as follows (HSE 2001, p.6):

- Hazard is the potential for harm arising from an intrinsic property or disposition of something to cause detriment.
- Risk is the chance that something of value or someone will be adversely affected in a stipulated way by the hazard.

The CCPS on the other hand gave the following definitions (CCPS 2000, p.6):

- Hazard is a chemical or physical condition that has the potential for causing damage detriment to people, property, or the environment.
- Risk is the chance of a human injury, environmental damage or economic loss in terms of both the incident likelihood and the magnitude of the loss or injury.

Aside from these definitions, the hazard can be described as a condition caused by an intrinsic property because what causes that hazard is sometimes more remote than the true hazard that represents them. Hence, it makes more sense to recognise a physical hazard that has the potential to cause the risk so that any control measures to mitigate the risk can be provided.

Some researchers have classified risk into the following based on precautionary issues (Luizz et al. 2014):

- Residual or acceptable risk - a situation where the risk is not supported by science-based evidence.
- Certain or unacceptable risk – a risk whose cause-effect link between the event and the damage caused can be scientifically proven
- Uncertain risk - a risk which is not yet scientifically proven, but whose existence cannot be ruled out.

However, risk cannot be discussed without defining other terms which are also used interchangeably including uncertainty and loss.

2.2.1. Uncertainty and Loss

Uncertainty has been explained as a situation where there is no previous historic data or event relating to the situation under consideration (Ashan & Sakale 2014). Thus, uncertainty may refer to ‘spread of an outcome' or the likelihood that the proposed outcome may not be

as predicted. Hence when a risk is estimated, there must be a clear documentary evidence to show the procedure, data and any assumptions made, so that the outcome may not be variable or incomplete (White 2008). Accordingly, a risk may relate to the same underlying concept (unknown future) but the probability of the event occurring can be assessed due to the existence of information and historical data. This suggests that the difference between risk and uncertainty is in the ability for probability estimation of the event occurring which is found in risk but not uncertainty. The consequence of risk can therefore be negative as well as positive.

Generally, estimating risk involves combining consequence data and the incident frequency data. Thus, the process for a risk estimate is always accompanied by the possibility of some uncertainties. Where there is established knowledge about issues with the data used in the risk estimation, the uncertainty associated with the risk estimate is 'knowledge uncertainty'.

Knowledge uncertainty arise when there are gaps in management of knowledge process and in the data with which the process operates (Berztiss 2004). As a result, there is the need qualify the evidence to ensure that the conclusions of the process do not go beyond what is known. The uncertainty associated with a risk estimate must therefore be modelled to alleviate any concerns relating to the validity of the risk estimation method. The HSE (2001) use Figure 2.2a to illustrate one of the processes by which risk uncertainty can be modelled.

The figure shows a horizontal axis along which uncertainty in consequence of the risk increases and a vertical axis along which uncertainty in the likelihood increases. Thus, moving along the directions of the horizontal axis means robustness of the risk analysis decreases because assumptions made in the analysis are not be validated. As a result, uncertainty in the consequence increases on the assumption that the risk is solely based on the consequence. Alternatively, moving towards the bottom of the vertical axis the degree of uncertainty about likelihood increases so the consequence is allocated to the hazard.

At the onset of both axes, the method for the risk analysis is more robust because the assumptions are verified as part of the process. Moving along the diagonal, reliance on past experience of generic hazard increases. As a result, robustness of the method decreases due to lack of information about the risk. So, whereas knowledge uncertainty suggests that more information could generally decrease in uncertainty, this does not necessarily apply to the probability of a risk.

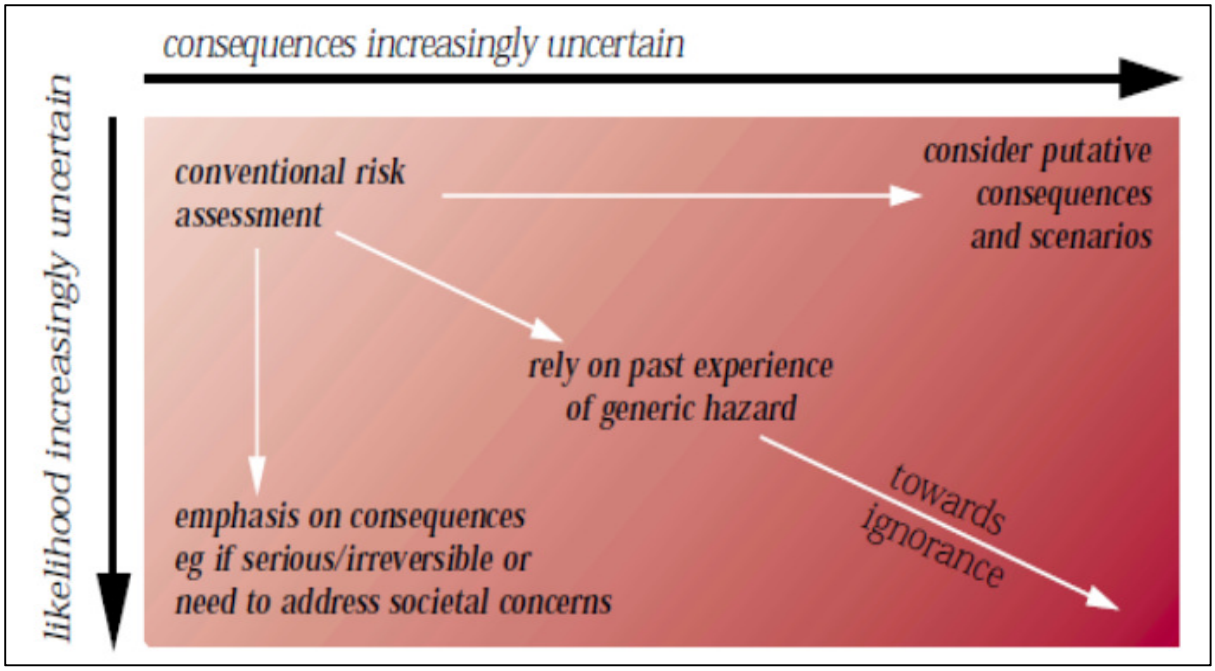


Figure 2.2a: Procedures for tackling uncertainty when assessing risks (Source: HSE 2001)

Loss may be due to the negative outcome of risk (Ingram 2014). Thus, although risk and loss are sometimes used interchangeably, loss categorises the degree of severity and frequency of risk within operation (Pezier 2002). In terms of data, loss and risk expresses the extent by which a model performs against the data, but this depends on the difference in the type of data (Li 2018). Li also explains that loss expresses how well a model performs against a training data while risk measures the loss across the entire data.

Pezier (2002) use log-frequency/log-severity (Figure 2.2b) to categorise operational risk into four based on the losses expected from operations as:

- Normal operational risk, in which expected losses are more significant than the expected risk.
- Ordinary operational risk, in which both expected risks and expected losses are significant.
- Exceptional operational risk, in which the expected risks are much more significant than expected losses.
- Immaterial operational losses, for which the expected risks and expected losses are significant.

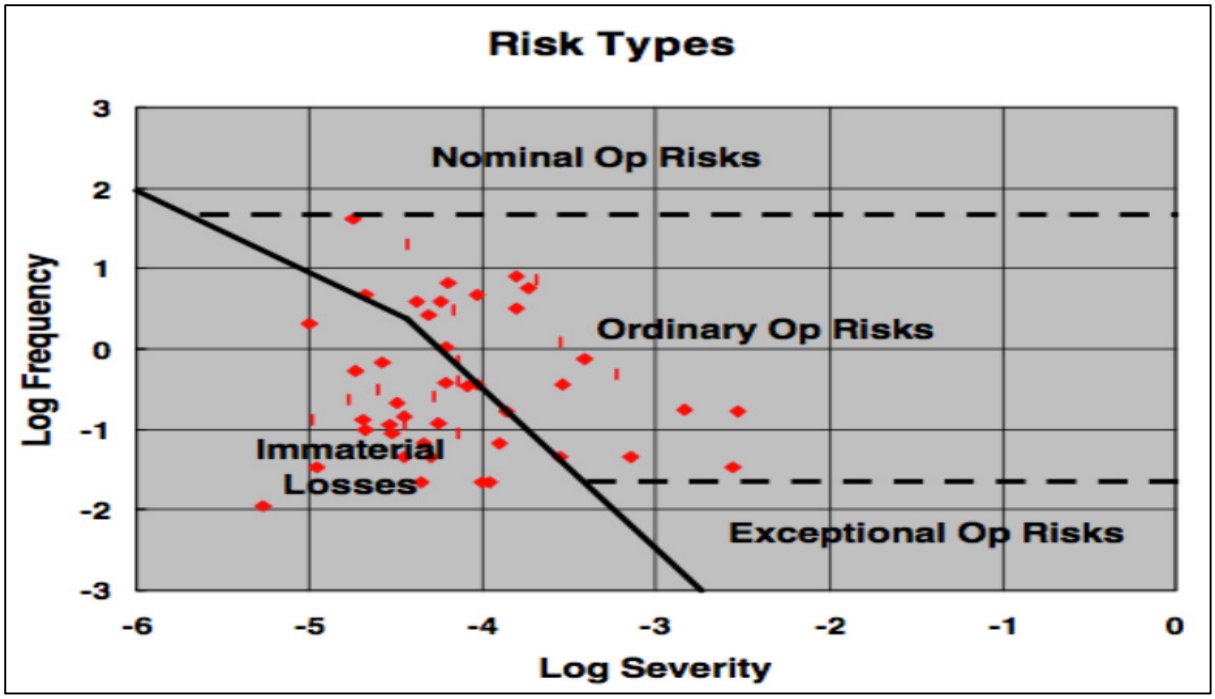


Figure 2.2b: Categories of operational risk (Source: Pezier 2002).

Another component of risk is 'uncertainty avoidance' which is one of the five cultural dimensions identified by Hofstede and Bond (Hofstede & Bond, 1984). They define culture as a collective program of the mind which distinguishes members of one group or a category of people from another (Hofstede 2001). The five cultural aspects were described as masculinity-femininity, collectivism-individualism, uncertainty avoidance, long term-short term orientation, and power distance (Hancıoğlu. et al. 2014).

Uncertainty avoidance was explained by Hofstede and Bond as a degree by which people feel threatened by ambiguous situations from which they create beliefs and institutions to avoid (Hofstede & Bond, 1984). Institution and organisations have adopted uncertainty avoidance to shape the pattern of behaviour, and expectations of individuals and groups within the organisation (Maxell 2013). It is imperative to note that the terms, hazard, uncertainty, loss, and uncertainty avoidance are all considered as components of the risk. However, they must not be considered as "risk" which is the concept under investigation within this study. The models and methods use for risk investigations are expected to be 'robust'.

The term robust refers to a measure of the capacity of a model or method to remain unaffected by small but deliberate variations in its parameters and provides an indication of its reliability during normal usage (ICH 2005). This depends on the relationship between available input data and the outcome. Most risk investigation, analysis models and methods are built on complex assumptions which means that their performance is valid if the

underlying assumptions are true, otherwise the outcome of using the model or method can be misleading. As a result, the method for a risk analysis must not just be robust but must also be 'effective'.

Effectiveness of a model or method refers to its ability to identify or detect the risk, determine the probability of the risk occurring, its consequence as well as the uncertainty around its impact, and the significance of the risk in relation to the process objectives (Hongping, et al. 2018; Zhang, et al. 2018). Hence, the model or method must be able to determine the known and the unknown. The model and method must also include 'sensitivity analyses' so that the outcome can be evaluated and be devoid of limitations.

Sensitivity analysis helps determine the impact or ability of the models or methods to determine the association of the parameters of the model or method and the outcome. It involves the determination of the parameters which aids the prediction of the outcome and allows identification and ranking of the most important factors which leads to great improvements in the output factors (Khoshroo, et al. 2018). Thus, in performing a QRA, the analysis must investigate how uncertainty outputs can be allocated to different input parameters and usually focuses on (a) identifying the most dominant parameters; (b) highlighting factors which may require additional research for strengthening the knowledge base; and (c) determine insufficient parameters which can be eliminated to avoid over parameterization (Hong & Purucker 2018). Figure 2.2c represent the list of some of the categories of the most commonly used sensitivity analysis methods.

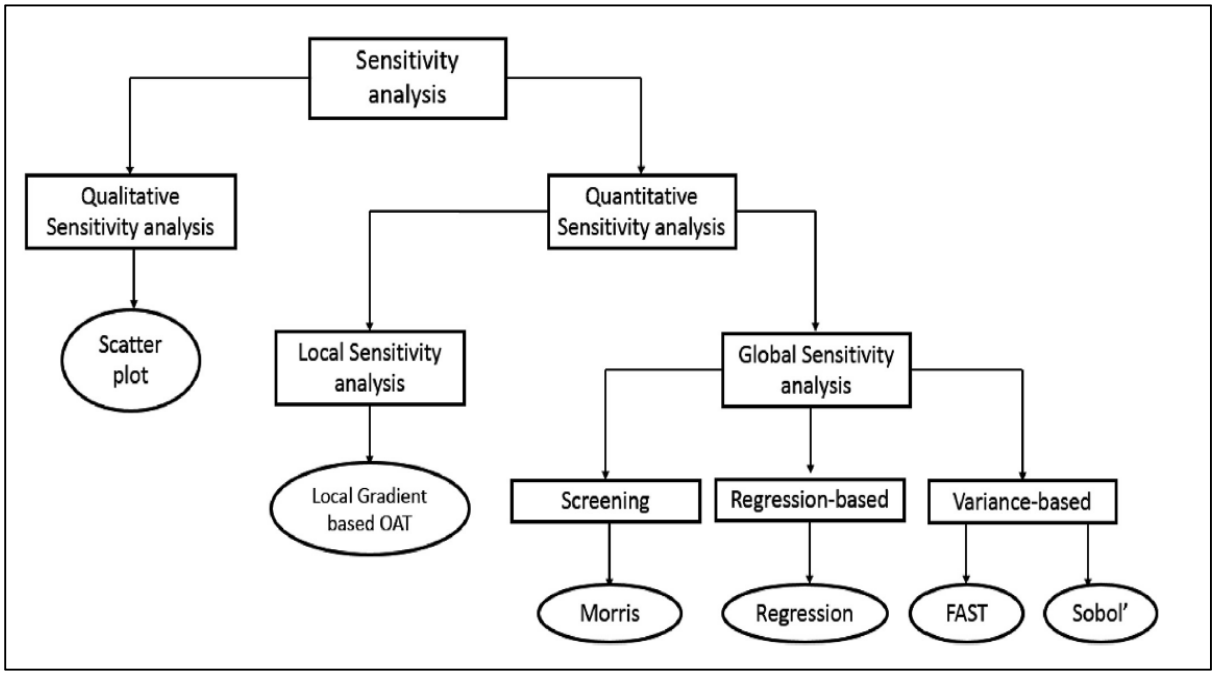


Figure 2.2c: Category of sensitivity analysis methods (Source: Hong & Purucker 2018)

2.3. Risk Theories

A theoretical framework acts as a blueprint for an entire research study, serves as a guide on which to build and support the study, and provides the structure which defines how the study is conducted (Grant & Osanloo 2014). Grant and Osanloo further explain that the theory for a research offers a conceptual basis for understanding, analysing, and designing ways to investigate a problem. Thus, the researcher must define the approach to the research problem and provide the rationale for how and why the study is being conducted so a reader could understand the researcher's inclination with regards to issues being address by the study. Kitchel and Ball (2014) explains that a theory incorporates a set of inter-related concepts, definitions, and propositions that present the phenomenon which specify relation among variables by explaining and predicting the phenomena.

Therefore, the theoretical and conceptual framework must "explain the path of a research and grounds it firmly in theoretical constructs" (Adom, Hussein, & Agyem 2018, pp. 438), and aim to make research findings more meaningful and acceptable. Accordingly, the theoretical framework also explains the conceptual framework and the two ideas appear to be similar in nature. However, they differ in their approach, style, and utilization within a study.

Thus, every study requires a theoretical and conceptual framework together with a literature review because all three functions (Rocco & Plankhotnik 2009) builds the foundation for the study, demonstrate how the study advances knowledge, conceptualize the study, assess design and instrumentation, and provide a reference point for interpretation of findings (Merriam & Simpson 2000). These functions are not necessarily fulfilled by the review or framework, but functions in a comparable manner and therefore use interchangeably. This study therefore evaluates various theories to help choose appropriate theoretical assumption for the QRA method.

2.3.1. Evaluation of Risk Theories for Selecting Theoretical and Conceptual Frameworks

In the UK, the risk regulator (the HSE) require that any risk created by an entity must manage to ‘as low as reasonably practicable’ (ALARP) or ‘so far as is reasonably practicable' (SFAIRP) levels by the risk creator (Russ 2010; HSE 2014). The two terms essentially mean the same and at their core is the concept of ‘reasonably practicable’ which involve weighing a risk against the effort, time and money needed to control it (HSE 2014). Thus, the HSE only inspects the quality of risk management but does not instruct individuals or organisations on how to manage risks. Over the years, several theories have been

integrated into the concept of risk. They include human behavioural theories, organisational theory, system behavioural theory, and contingency theory.

2.3.1a. Human Behavioural Theory

Behavioural theories have been integrated into the concept of risk to investigate human behaviours within organisations for managing process risks. There has been some attempt to investigate the relationship between human behaviour and the level of risk, which establish that some of the accidents are humanly motivated (Wilde 1982). However, within the concept of risk, one could split behavioural theories into (a) human behaviour theory and (b) system behavioural theory.

Over the years, various literature on human behaviour theory has been reviewed. For instance, Guldenmund has reviewed various literature on safety culture and safety climate within organisational safety practices over a 20-year period (Guldenmund 2000). Guldenmund found that though organisations practice safety culture and climate, there is lark of agreement on the cause, content and consequence of the models to specify the relationship of the concepts with risk management or safety performance. This study will review behavioural safety programs (BSP) to access if the focus of safety should be on human behavioural elements or the process itself as part of the steps to obtain the suitable QRA method for the HHPIs in Part 2.

2.3.1b. Organisational Theory

Organisational theory involves an effort by organisations to achieve with acceptable risk which is usually describe as safe operations. As explained by Grote (2012), more high-risk industries are adopting safety management with an emphasis on learning from the different risks and the corresponding limits to generalise safety management methods. Organisational safety management is expected to be designed, run, and assessed on three crucial parameters namely "the kinds of safety to be managed, the general approach to managing uncertainty as a hallmark of organizations that manage safety, and the regulatory regime within which safety is managed" (Grote 2012, pp. 1983). However managerial direct involvements are generally influenced by organisational issues and therefore there is the need to keep the balance for managing risk based on cost and effect.

2.3.1c. System Behavioural Theory

Within typical process system are various components which operate independently or in a system of a loop. These individual components perform special functions of their own which

contributes to the total operation of the processing system and hence their individual behaviour can be studied separately. Assumptions underlying the study of system behaviours includes separation of the process into subsystems which operate independently with results which could be analysed separately from that of the entire system.

Systems theory therefore explains a system as "interdependent components working together in a cooperative manner to accomplish a purpose" (Smit 2010, p.7). This theory sounds good to be considered in process terms due to the complexity of process systems and interrelations between different sub-parts.

Leveson and Stephanopoulos, on the other hand, explain system theories as an approach which focuses on the entire process system without decomposing its behaviour into individual events over time (Leveson & Stephanopoulos 2014). Thus, the study of system behaviour theory includes studies into properties of the entire process system, as well as human, social, legislative and regulatory guidelines surrounding activities within the process. Leveson and Stephanopoulos also suggest that applying system theory could lead to new types of risk and accident analysis. Some researchers have also proposed that system domain theory involves the ‘man-machine systems' and ‘system ergonomics' which originated from the United States in the 1960s (Swuste et al. 2014).

2.3.1d. Contingency theory

As discussed above, each risk is unique and therefore must be managed in accordance with its specific characteristics and location within a specific period. Since the focus of this study involves the detection of risk which falls under managing risk, the uniqueness of the nature of process risk means they cannot be managed in ‘one best way’. Therefore, in choosing the appropriate theoretical framework for the thesis, contingency theory was selected since the concept of contingency theory and the focus of the thesis have a correlation.

Although the contingency approach suggests that “there is no one best approach" for managing risk, it does not oppose the existence of alternative pathways which might be more appropriate for each specific contingency. As with other theories, there could be some objections to contingency theory but its suitability for this study is based on its risk-based concept which is the focus of the study. Because contingency theory is assumed to be a risk management concept, the study selects contingency analysis, which is an aspect of contingency theory for this thesis.

Contingency analysis (CA) has been applied to minimise risk in the HHPI's, nuclear, oil and gas, aviation and healthcare industries, and in event of emergencies (Everdij & Blom, 2016).

They also explain that CA involves identifying potential accidents and elevating adequacies of emergency measures. Because unpredictable events may have a devastating impact on resources and operations, the CA must include a list of all potential contingency occurrences and the post-contingency process which could be applied to minimise any contingency violations in the HHPI (Everdij & Blom 2016).

2.4. Risk Assessment and Risk Analysis

Risk assessment and risk analysis has also been used interchangeably over the years. However, some of the regulatory organisations involved in risk management e.g. Factor Analysis of Information Risk (FAIR), OSHA and HSE, explains that risk analysis involves identifying potential threats to the organization for which the related vulnerabilities must be analysed, whereas risk assessment involves evaluating existing controls and their effectiveness to the potential threats (HSE 2014; Copland 2017; Ventiv-Aon 2017). The process to identify, analyse, evaluate, handle exposure to losses, monitor risk control measures and minimise adverse effects are described as risk management. The relationship between risk management, risk assessment and risk analysis may represent by Figure 2.4a.

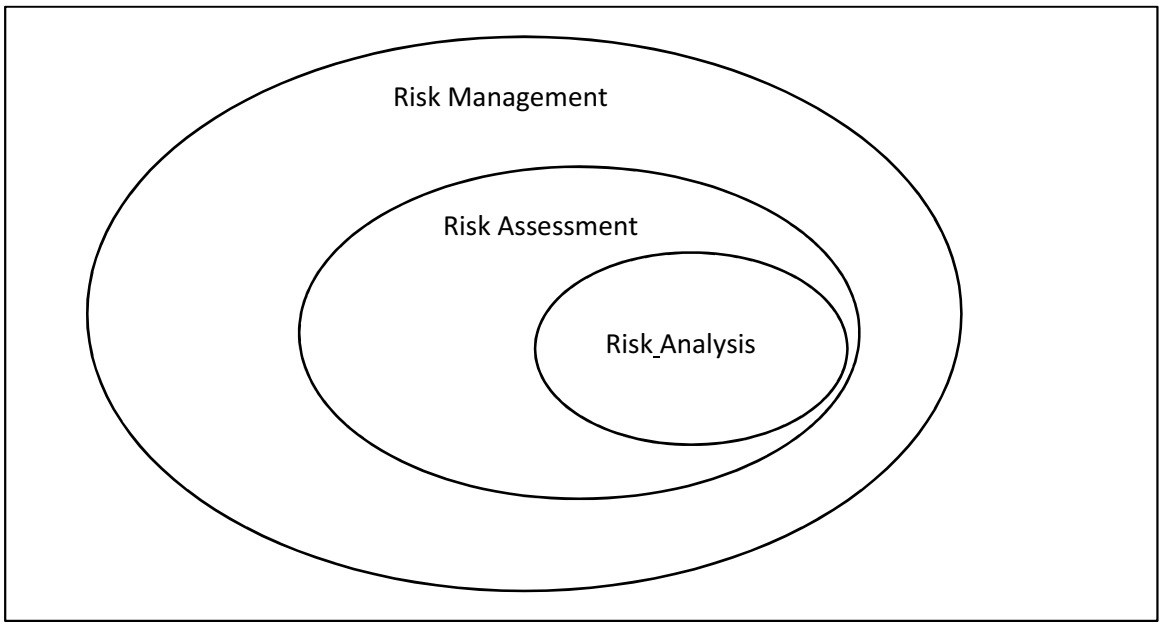


Figure 2.4a: Relationship between risk management, risk assessment and risk analysis (Adapted from: www.healthguardsecurity.com)

Unfortunately, some risk management involves assessments which are less accurate and therefore produce results which poorly inform organisations about prioritization and cost-ineffective decisions and this violates the purpose risk analysis (Copland 2017). As discussed in 2.3.1d, each risk is unique and therefore must be managed in accordance with its specific characteristics and location within a specific period. After identifying the risks through some collaborative effort from all safety experts and process personnel, a root

cause analysis is performed to find the root cause of a hazard and to determine how to mitigate the risk. Thus, a risk analysis method could be either qualitative or quantitative depending on the type of data available for the analysis.

2.4.1. Qualitative Risk Analysis

Qualitative risk analysis usually applies the descriptive methods to determine the impact and probability of risk. Qualitative risk analysis techniques are therefore modelled with qualitative knowledge (e.g. describe situations or scenarios). They include failure mode, effects, and criticality analysis (FMECA), hazard operability (HAZOP), layer of protection analysis (LOPA), safety integrity levels (SILs), probabilistic risk assessment (Webber et al., 2012). As an example, the study will use LOPA and SILs to explain a qualitative risk analysis.

2.4.1a. Layer of Protection Analysis

Layer of protection analysis (LOPA) refers to a calculation of residual risk "used to assess the requirements for safety-critical instrument loops" (de Salis 2012; p. 183). Outcome of a LOPA is used to reduce risks within a typical process system and is recommended in the standard for level of system safety performance requirements including the BS EN 61511 which is recommended by the HSE (HSE 2019 p.7). The HSE guide explains that safety-instrumented systems and other risk reduction measures which form part of the overall safety of the process plant must conform to the requirements as set in the standard.

One key approach to safety includes the use of hazard and risk assessment to identify how risk can be located in process equipment's to ensure their safe operations. A typical safety system for instruments also known as safety instrument system (SIS) includes

- Alarms on the processing system which help operators to suspect issues in their operations.
- Controls on basic process equipment's which are linked to the basic process control systems (BPCS) on the plant.

2.4.1b. Safety Integrity Levels

Safety integrity levels (SIL) refer to a measure of the safety system performance. A typical SIL assessment is based on the idea that each risk event may have several properties within the process design and operation that reduces the unwanted event's likelihood (de Salis 2012). The HSE electrical and control and instrumentation (EC&I) guide (HSE 2019) suggests that SIL assessments should be applied to all lifecycle phases of the instrument

loop and requires safety-instrumented system (SIS) to achieve compliance. As part of SIL assessment, safety instrumented function (SIF) is first determined from hazard identification assessment before designing of the SIS. To meet compliance information applied to develop the SIS must be meet compliance to BS EN 61511 required for a given SIL. There are four discrete SIL integrity levels based on the probability of failure which are assigned based on controllability categories as exemplified in Table 2.4 (Charlwood, Turner, & Worsell 2004).

Table 2.4: SIL assignments categories (Charlwood, Turner & Worsell 2004)

Controllability Category	Acceptable Failure Rate	Integrity Level
Uncontrollable	Extremely improbable	4
Difficult to control	Very remote	3
Debilitating	Remote	2
Distracting	Unlikely	1
Nuisance only	Reasonably possible	0

Thus, LOPA techniques for a proposed instrument performed by a team of assessors can also be used for SIL analysis. Like other qualitative risk assessment methods including FMEA and Checklist analysis (Giannini et al. 2006), LOPA is a hazard identification approach with Checklist analysis being the simplest tool (Giannini et al. 2006). All these methods rely on the hard-won experience of operators and specialist which makes the methods extremely difficult to use without some level of competency.

2.4.1c. Issues with Qualitative Risk Analysis Methods

Some research has highlight issues with the qualitative risk analysis methods applied within the industry. For instance, a review of LOPA on the data use reveal that sometimes the data sources, data quality and data type applied are inappropriate and the level of uncertainty varies and do not incorporate sensitivity study (Chambers, Wilday & Turner 2009). With more emphasis on published data, one would expect a LOPA to rely on data from published sources instead of knowledge from experienced process operator's maintenance staff and expertise of safety professionals.

Applying qualitative evaluation to all identified scenarios help organisations to prioritise the most important problems and propose measures to help mitigate any risk associated with the problem identified (Kotek & Tabas 2012). As a result, qualitative methods like HAZOP, SIL and LOPA can be performed at the beginning of the quantitative risk analysis process. Because qualitative risk analysis is not the motivation for this research, the study will focus on quantitative risk analysis (QRA).

2.4.2. Quantitative Risk Analysis

A quantitative risk analysis (QRA) method applies various mathematical and statistical techniques as well as quality numerical data. It is widely applied as a tool to improve safety from the process design stage to its operation and deemed very important because of its practical application for decision-making on safety (Goerlandt, Khakzad & Reniers 2017). Several reviews of risk analysis methods suggest that QRA is applied from Oil and gas to chemical installations (Khan et al. 2015). This study provides a more thorough systematic review and content analysis of QRA in Part 2.

Other concepts such as preventative maintenance (PM) and remaining useful life (RUL) also uses numerical data and are sometimes assumed to be QRA methods. To help distinguish QRA from these methods, the study provides a brief description of these methods in the next two sections. The study begins by presenting a classical overview a typical PM based on the bathtub curve.

2.4.2a. Preventative Maintenance

Manufacturing processes are prone to increasing wear over time due to usage or age and are therefore affected by occasional failures resulting from deterioration. As a result, they must be repaired and maintained to avoid downtime and revenue losses. Industry, therefore, requires a procedure for reducing the occurrence of the failure by conducting a periodic maintenance procedure. PM has been widely applied in industry. For instance, it has been applied to continuously deteriorating system which are subject to stress such as internal vibration signal of ball-bearings (Deloux, Castanier & Berenguer 2009), optimizing uptime and performance in vehicle fleets (Chaudhuri, 2018) and reducing breakdowns and maintenance costs of other manufacturing machines (Bastos, Lopes & Pires 2014).

A PM is sometimes based on statistical analysis and historical life data of the specific equipment (Rezvanizaniani, Dempsey & Lee 2014). One example of the statistical life of a manufacturing process equipment is presented by the bathtub curve which represents a hypothetical failure rate with time (Figure 2.4b).

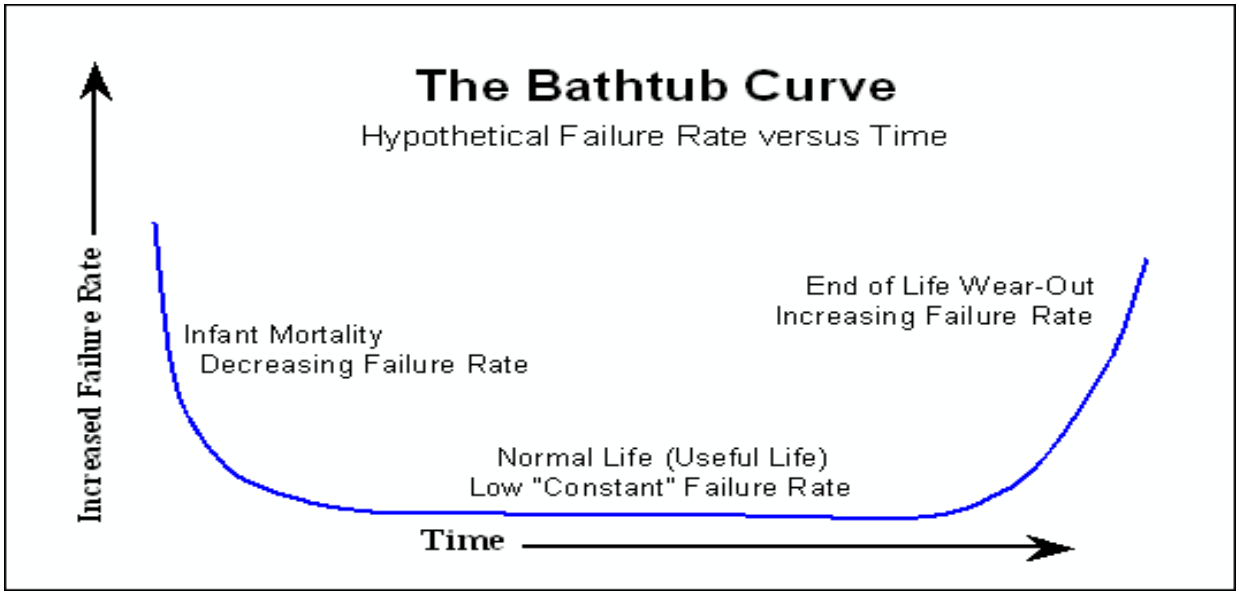


Figure 2.4b: The Reliability Bathtub Curve (Source: Wilkins 2002)

The bathtub curve shows three periodic events (Wilkins, 2002)

- A decreasing failure rate.
- A period of low or relatively constant failure rate, and
- A wear-out period shows increasing failure rate.

This curve suggests that the failure process of a new equipment during the first few usages may be due to manufacturing or installation problems. This is followed by an extended period where the probability of failure is low. Then a period where the probability of failure increases. This requires a Risk-based maintenance (RBM) which is a system of predictive analytics which involves maintenance management and statistical tools. A typical RBM "aims to improve maintenance planning and decision making by reducing the probability and consequences of the failure of equipment" (Xu et al. 2013, p. 1).

According to the Norwegian Standard for Oil and Gas Industry, the selection and prioritisation of maintenance activities for RBM concepts are based on the principles of risk analysis (NTC 2001). Hence the requirements used to determine the RBM programs require contingency plans which begins with a risk assessment model. A typical process diagram for RBM is shown in Figure 2.4c.

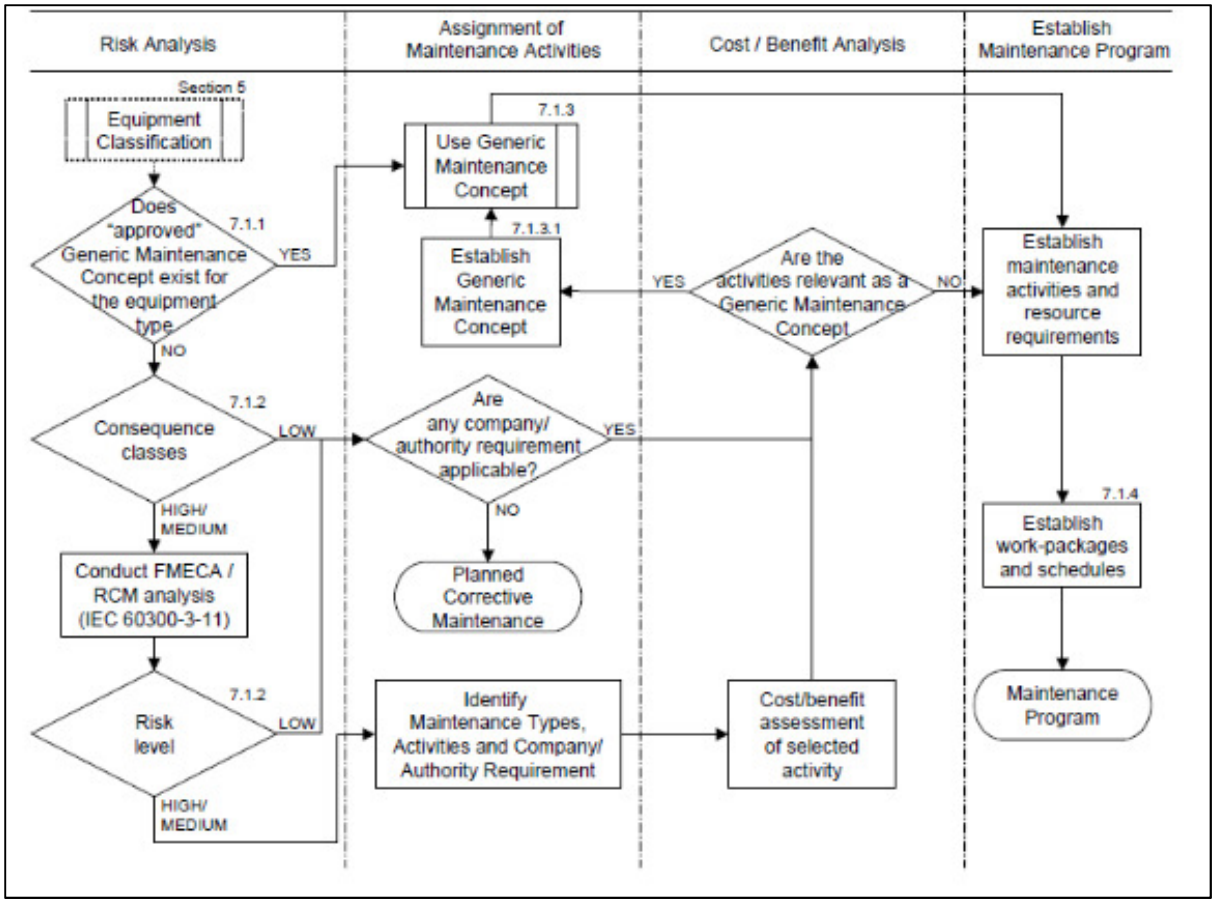


Figure 2.4c: Process diagram for RBM (Source: NTC 2001)

Previous work has described RBM as a proactive approach for safety management using non-destructive techniques which determines the status of equipment before a breakdown occurs (Hashemian & Bean 2011; Perl, Mulyukin & Kossovich 2017). It compliments PM because it is from the condition of equipment and not the statistics of its life expectancy. For instance, whereas a planning proactive maintenance depends on the use of mean-time-between-failure (MTBF) statistics (e.g. machine inspection or repair data) to prevent the occurrence of equipment malfunctioning; the PM applies equipment operation data.

According to the HSE, PM and proactive maintenance are part of maintenance management (MM) system which "should deliver the effective inspection, maintenance and testing activities that assess the condition of plant, detect deterioration and remedy the identified shortcomings" (HID Inspection Guide Offshore: Inspection of Maintenance Management, p. 3) and is a standard for addressing issues relating to risk of subsequent failure.

Although some elements of this study appear to show characteristics of PM, as can be seen in the RBM classification tree of Figure 2.4c, a QRA on its own is not a PM because a PM incorporates many processes including a risk assessment stage.

2.4.2b. Remaining Useful Life

Remaining useful life (RUL) has been applied to predict the lifespan of process systems with the goal of reducing minimising failure in both manufacturing and service sectors as part of the maintenance decision-making process (Okoh et al. 2014). It has been applied as a prognostic technique where condition indicators are applied to estimate remaining time of failure (Ragab et al. 2014). The process involves using statistical and machine learning algorithms to historical data to discover hidden patterns (Witten, Frank & Hall 2011).

Various statistical and data-driven statistical techniques including linear regression, support vector machine learning, Bayesian model, Hidden Markov model have all been applied in RUL estimation (Wang et. Al. 2007; Tian et al. 2012; Kim et al. 2012; Son et al. 2013; Caesarendra et al. 2017;). RUL has been applied extensively in PM to estimate functioning and reliability of operating equipment's. This gives degradation indicators which helps safety and maintenance managers to schedule machine maintenance to help minimise downtimes. Although some attempt has been made to use RUL to predict the risk of failure (Tian et al. 2012; Ragab et al. 2014) it differs from QRA methodology because RUL is mainly applied as a fault diagnostic model for PM.

Considering the above discussions, the study concludes that PM and RUL are proactive measures. These together with other protective measures has been classified into the following (HSE 2000):

- Passive, which minimises the hazard using process and equipment design features by reducing the frequency or consequence of the hazard with minimum functioning of any device.
- Active, which applies engineering controls and any safety devices on the processing system detect anomalies within process operations
- Procedural, which applies operating procedures, and other operation management approaches to prevent incidents (HSE 2000).

2.5. Conclusion

This chapter began with an outline of the framework for the presentation of Part 1, followed by PSM, introduction of risk as a concept of risk and definition of key generic terms in the context of their usage in the research. After that, the study evaluated some theoretical and conceptual framework from which contingency theory was selected as the appropriate theory for the study. A discussion of risk assessment and QRA was also presented.

Next is Part 2, where the study presents a review of research literature publications BSP to help establish whether the focus of safety in the HHPI must be on behavioural elements of PS or on the process itself. Where the review establish that the focus must be on the process itself, the study will perform a systematic review and content-analysis of published research literature on existing QRA methods use in the HHPI.

Part 2
Literature Review and Systematic Content-analysis, Real-life Case Histories Process Safety Incidents and Data

Part 2 - Background

In Part 1, the study introduced process safety management (PSM) and risk as a concept. The study then evaluated conceptual and theoretical frameworks to help select a theory for the research. This was followed by risk assessment, after which quantitative risk analysis (QRA) which is the focus of this research was discussed. As part of the objectives to provide a QRA method which relies entirely on big data techniques and real-time datasets as a major contribution to science and practice. The study now presents:

- A systematic review and content-analysis of research publications relating to review questions.
- Basic concepts, particularly of dust fire and explosions.
- Definition of some properties and terms use in dust fire and explosions and in the context of the study.
- Real-life case histories of industrial dust fire and explosion incidents and a critique the final reports on the investigation these incidents by a relevant authority.
- Provide a justification for using real-life incidents as case study for the research.
- Introduction of available datasets for the research.

To make the presentation of the thesis more structured and easier to read, Part 2 presents the chapters for the list above in a way of taking the reader through literature reviews to the type of datasets available for the research. The overall approach is illustrated by the flowchart of Figure 3a.

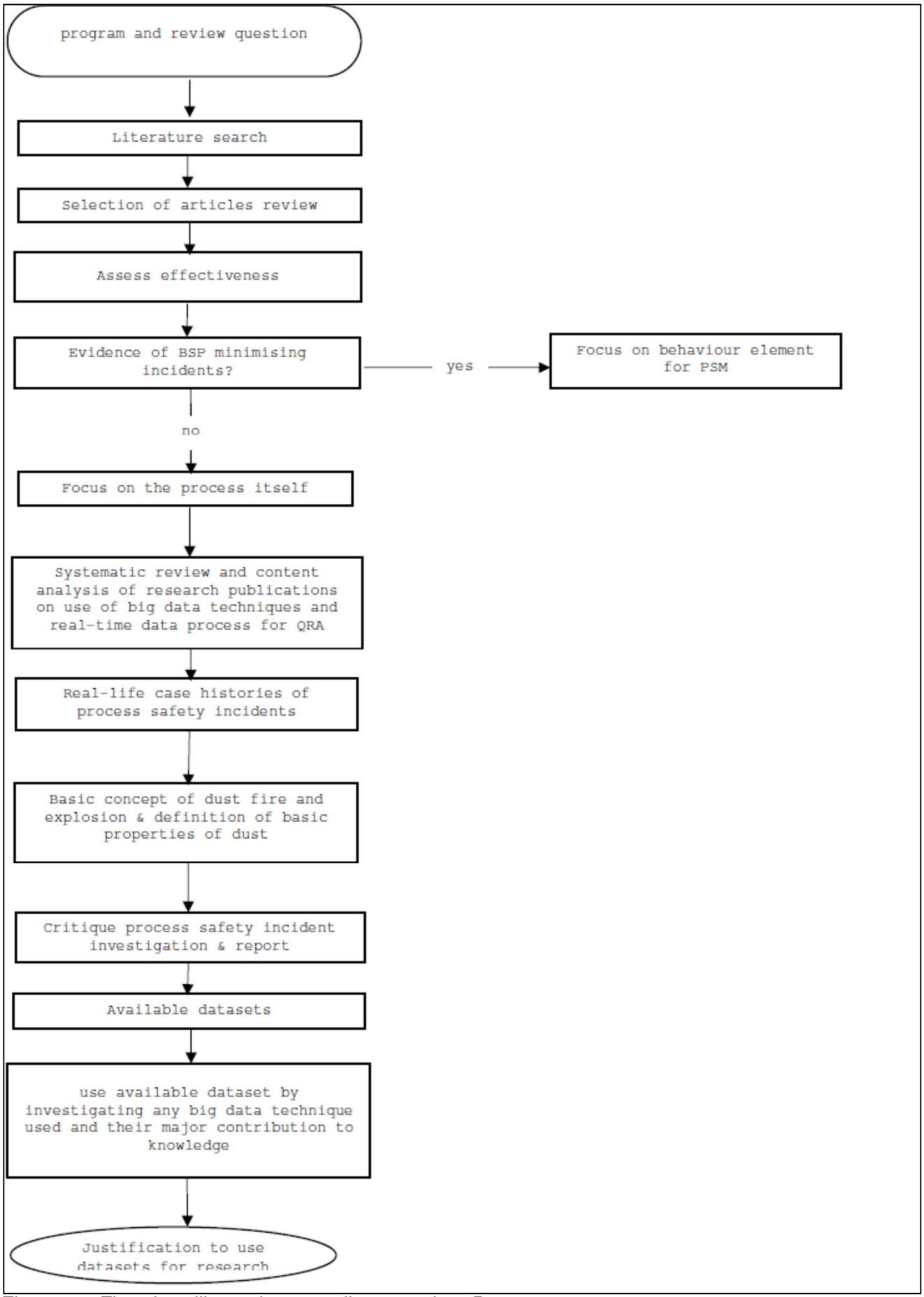


Figure 3a: Flowchart illustrating overall approach to Part 2.

Chapter 3 – Literature Review and Systematic Content-analysis

3.0. Introduction

In Chapter 2, the study discusses BSP as a concept of process safety management (PSM). The concept of risk and was also discuss after which contingency theory was selected as the appropriate theoretical framework for this study. The study then concludes that QRA is the concept for which the study is being conducted.

In this chapter, the study reviews research publication on the concept of behavioural safety programs (BSP) as applied for managing risk within the HHPIs. The aim of this literature review of application of BSP in the HHPI is expected to help establish whether the emphasis of PSM should be on behavioural elements (of personnel) or on the process itself. Where the review reveals that the application of BSP had minimal impact at reducing risks at the HHPI, the study will conclude that the focus of safety must be on the process itself. The study will then proceed with a review of publications of how big data techniques and real-time process monitoring data have been applied in the methods currently use for QRA in the HHPIs. The overall framework of the approach to the literature reviews are presented as the flowchart of Figure 3b.

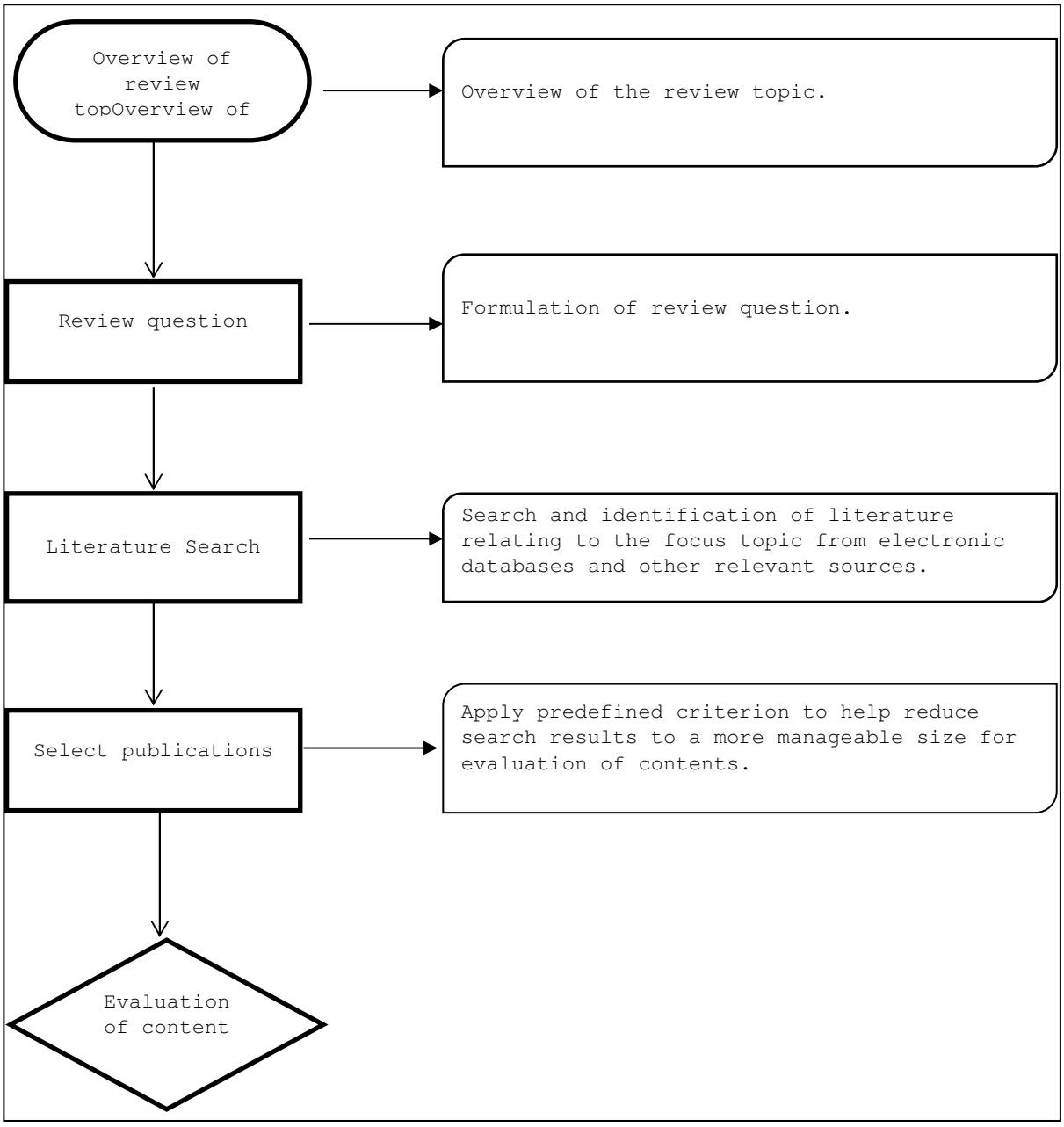


Figure 3b: Flowchart for the framework of the literature review

3.1. Electronic and Manual Literature Search

Electronic databases will be search for peer reviewed research publications on the focus topic to identify relevant articles for the review. Various databases, references of literature and other citations will be search. Predefined filters will be applied to help reduce the search result to levels which could be handle without affecting the time allocated for the research.

3.1.1 Method for the Literature Reviews and Systematic Content Analysis

To identify relevant articles for the review, thirteen electronic chemical engineering databases will be searched using search strings and terms constructed with the aid of the Cochrane Handbook

for Systematic Reviews of Interventions (TCC 2011). They include database for psychology (PsychINFO), UK HSE, EbscoHost, Institute of Chemical Engineering (IChemE), Science Direct, US Environmental Protection (US EPA), European Process Safety Centre (EPSC), European Union Labour Force Survey (EU-LFS), European Federation of Chemical Engineering (EFCE), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Process Safety Incidents database (PSID), International Powered Access Federation (IPAF), Web of Knowledge, and bibliographies of other related articles were also retrieved. Wildcards will be applied to accommodate international spelling variations. The systematic appraisal of the citations will be performed by applying the criteria for inclusion and exclusion. Filters like language, subject domain, and article will be applied to help eliminate any unwanted articles and ensure that only relevant articles are considered.

3.1.2. Filters for Inclusion and Exclusion

Prior to the research, it was deemed that another filter for time of the publication must be considered. Because the major legislation on control of major accident hazards (COMAH) initially came into force in 1999, it was initially decided that publications from 1999 must be considered. However, it was realised that there have been several reviews of both BSP and the existing QRA methods since COMAH 1999. These include

- the review of QRA method by Tixier et al. in 2002.
- Review of BSP at prevention of major hazard incidents by Bell and Healey published in 2006.
- Review of existing QRA methods by Patel & Sohani published in 2013.
- Review of process safety regulations and its enforcement globally by Besserman & Mentzer published in 2017.

COMAH 99 has also been replaced by CPMAH 2015. It was therefore decided that the study sets a filter to exclude publications prior to the year 2007 to help eliminate a repetition of work done by previous research and ascertain whether there could be any gabs which has not been covered in the previous reviews.

Other filters will also be applied to help eliminate any unwanted articles and ensure that only relevant articles are considered. These include

- Filters like language to ensure that only literature published in English language are selected because of the domain language use for this study.
- Filter for industry to eliminate industries like Nuclear, Construction, Health and other industries whose operations do not fall under the HHPI.

- Filters for hazards to eliminate publications which cover hazards for which the controls required are outside those of the HHPI (e.g. outside exposure to solar radiation and flooding).
- Filters will be applied for type of publication to ensure that only peer reviewed publications are considered for this review.
- Filters will be applied to eliminate sources such as books, blogs and patients from the list of literature.
- Because the research involves multidisciplinary approach, no filters will be applied for subject domain but certain domain areas like finance, law, music, politics, natural disaster management, archaeology, to mention a few will be excluded.

3.2. Behavioural Safety Programs *and Major Incidents in the HHPI's: A Systematic Review and Systematic Content Analysis*

As mention in Chapter 2, PSM is covered under various legislations worldwide. One of these legislations which is applied in the UK is the Control of Major Accident Hazards Regulations (HSE 1991). This legislation requires the HHPI's to take all necessary measures to manage their major accident hazards. The legislation has been superseded by COMAH Regulations 2015 (HSE 2015) to include:

- definition of dangerous substances using the harmonised system of classification using the EU's Classification, Labelling and Packaging (CLP) Regulation 2008,
- a transition arrangement for safety reports,
- stronger requirements for public information,
- emergency planning,
- competent authority on inspection and
- broader domino effects duty.

Accidents in the HHPI's occur either by direct causes (i.e. occur immediately prior to the undesirable event) or further away in time or space to the underlying causes which contributed to the event (Anderson 2005). Most of these incidents are deemed to have been due to errors on the part of frontline staff and operators. Owing to this, resources which may help to prevent these incidents are directed towards the frontline employees.

Previous research into major incidents and their causes within the HHPI's around the globe reveals that although the most common cause of these disasters is human error, they are not errors solely caused by frontline staff but by the designers of the processes and managers as well (Kletz 1999, p.48). For instance, Kletz explains that "many accidents have occurred because changes were made in plants or processes and these changes had unforeseen side effects" (Kletz 1999, p48).

Thus, the error of a frontline personnel may be a combinatorial effect involving every level in the organisation. As a result, BSPs which incorporate models based on the premise that behavioural factors make a significant proportion of accidents (Anderson 2005) have been incorporated into managing risk within the HHPIs.

The application of BSP suggests that more consideration is being given to human factors which are relevant to the control of hazards, than the general focus on occupational and/or personal safety. However, there are suggestions that it is extremely difficult to identify appropriate behavioural program models because BSPs are hampered by gaps in the evidence, and in the areas of effectiveness and the behaviour change processes (Lunte et al. 2011).

The elements of the BSPs has been explained and reviewed in various publication (HSE 2002, p.15; Flaming & Lardner 2002, p.3; Sulzer-Azaroff & Austin 2000; Skowron-Grabowska & Sobociński 2018), which suggests that the effectiveness of the program at reducing accidents varies widely. For instance, one review suggests that the effectiveness of programs varied from 2% to 85% improvement (Sulzer-Azaroff & Austin 2000). However, the size of the sample applied in these researches are relatively small. An overview of the BSP elements is represented by Figure 3.2a.

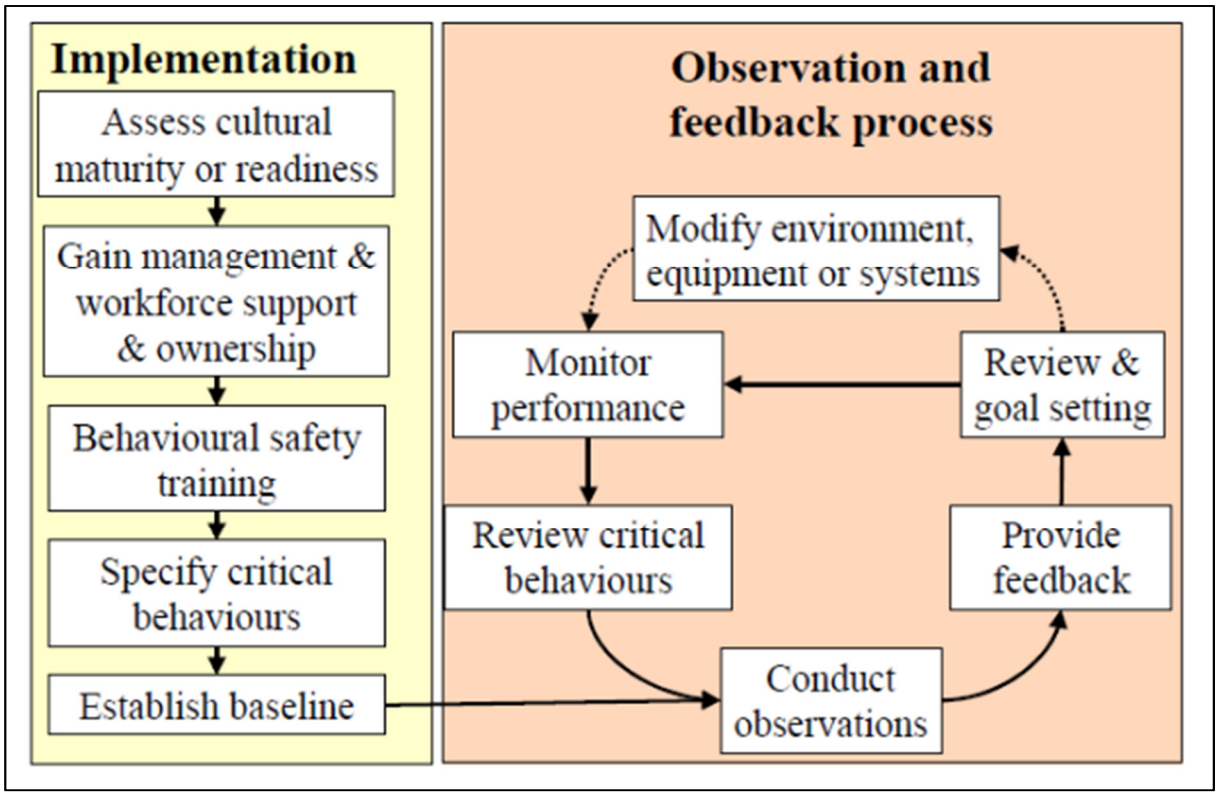


Figure 3.2a: Overview of Behaviour Safety Programs. Source: Fleming & Lardner 2002)

Safety measurements are usually base on lost time indicators (LTIs), process safety performance and organisation failures. Therefore, in addressing safety issues within the organisation, it is

important to consider training and operation procedures, engineering and hardware design, process safety, and human factors such as occupational or personal health and safety. Since QRA forms part of safety measurement, the study performs this systematic review of BSPs to ascertain whether the focus of safety within the HHPI's should be on monitoring behavioural of frontline personnel or on the process itself.

The review aims at examining research publications relating to BSPs in the HHPI's by evaluating the (a) content and procedure by which the programs were intended to exert their effect and (b) effectiveness of the programs at minimising incidents. The review question was formulated from the news about the acquisition and merging of a Behavioural Safety Company in the United States and a Process Safety Company in the United Kingdom.

Search string "Behaviour* OR Performance* OR Conduct* OR activity* OR Action*) AND (Science* OR Model* OR Approach* OR System* OR Assessment* OR Investigate*) AND (Safet* OR Protection* OR Well-being* OR Hazard* OR Danger*) AND (Benefit* OR Impact* OR Health*) AND (Chemical* OR Process* OR Product* OR Manufact*) AND (Communit* OR Volunt* OR Social exclu*," was initially applied by combining the keywords (behavioural, conduct, performance, activity, action), with the assessment terms (science, approach, assessment, investigate, system, model), result related terms (safety, protection, well-being, hazard, danger, benefit, impact), then industry type terms (chemical, hazardous chemicals, product, manufacturing) and other related terms (social, community, voluntary).

The final search string was (Behaviour* Science* model* OR Behavioural* assessment* model* OR Behavioural* Assessment approach* OR behavioural* assessment* investigation*) AND (Chemical* industry* incident* OR chemical* Process* incident*OR Product* Manufact*incident*) AND (safety* OR Hazard*) AND (benefit* OR success*). Details of all the search strings and the corresponding number of citations are provided in Table 3.2a.

Table 3.2a: Search strings and outcome

Search	No. of Citations
(Behaviour* OR Performance* OR Conduct* OR activity* OR Action*) AND (Science* OR Model* OR Approach* OR System* OR Assessment* OR Investigate*) AND (Safet* OR Protection* OR Well-being* OR Hazard* OR Danger*) AND (Benefit* OR Impact* OR Health*) AND (Chemical* OR Process* OR Product* OR Manufact*) AND (Communit* OR Volunt* OR Social exclu*)	0
(Behaviour* safety* OR Behaviour* protection* OR Performance* Safety* OR performance* Protection* OR Activity* OR Conduct* OR activity* OR Action*) AND (Science* OR Model* OR Approach* OR System* OR Assessment* OR Investigate*) AND (Well-being* OR Hazard* OR Danger*) AND (Benefit* OR Impact* OR Health*) AND (Chemical* OR Process* OR Product* OR Manufact*) AND (Communit* OR Volunt* OR Social exclu*)	60958
(Behaviour* safety* OR Behaviour* protection* OR Performance* Safety* OR performance* Protection* OR Activity* OR Conduct* OR activity* OR Action*) AND (Science* OR Model* OR Approach* OR System* OR Assessment* OR Investigate*) AND (Chemical* industry* OR chemical* Process* OR Product* Manufact*) AND (Well-being* OR Hazard* OR Danger*) AND (Benefit* OR Impact* OR Health*) AND (Communit* OR Volunt* OR Social exclu*)	72846
(Behaviour* safety* OR Behaviour* protection* OR Performance* Safety* OR performance* Protection* OR Activity* OR Conduct* OR activity* OR Action*) AND (Science* OR Model* OR Approach* OR System* OR Assessment* OR Investigate*) AND (Chemical* industry* hazard* incident* OR chemical* Process* hazard* incident*OR Product* Manufact*hazard* incident*) AND (Benefit* OR Impact* OR Health* OR Safety)	61320
(Behaviour* OR Behaviour* OR Performance* OR performance* OR Activity* OR Conduct* OR reactivity* OR Action*) AND (Science* OR Model* OR Approach* OR System* OR Assessment* OR Investigate*) AND (Chemical* industry* hazard* incident* OR chemical* Process* hazard* incident*OR Product* manufact*hazard* incident*) AND (Benefit* OR Impact* OR Health* OR Safety* OR Protection)	53625
(Behaviour* Science* model* OR Behavioural* assessment* model* OR Behavioural* Assessment approach* OR behavioural* assessment* investigation*) AND (Chemical* industry* incident* OR chemical* Process* incident*OR Product* Manufact*incident*) AND (safety* OR Hazard*) AND (benefit* OR success*)	48914

3.2.1. Criteria for Including and Excluding Citations

The term 'program' used in this review refers to strategies intended to help reduce major accidents in the HHPIs. As mention in this section, several reviews of BSP has been performed since COMAH 1991 including how behaviour and relevant control measures has help in prevention of major hazard incidents in 2006 (Bell & Healey 2006, p.iii-vi), and the review of literature on process safety regulations and its enforcement globally in 2017 (Besserman & Mentzer 2017). As a result, it a decision was made to review publications from 2007 to ascertain whether there could be any new findings. Any findings which has been covered in previous reviews will be rejected. Other criterion will be applied to help reduce the number of publications to manageable size. The detailed pre-define criterion is provided as Table 3.2b.

Table 3.2b: Criteria for inclusion and exclusion

Inclusion criteria	**Exclusion criteria**
Evaluation of intervention's intended to affect human behaviour compliance (actions that employees take to comply with health and safety measures).	General construction industries (residential, road, offices, etc); health (e.g. hospitals), industries involved in waste disposal of health-related waste materials, etc.
Evaluation of behavioural predecessor such as awareness or attitude towards risk within the controlled industries.	Hazards that have need of controls which are not specific to the Chemical and Major Hazard Industries (e.g. outside exposure to solar radiation, flooding, etc.).
Evaluation of conditions before and after the intervention measures and control group (employers, employees, etc) before and after the incidents and accidents? Evaluate the conditions before and after the intervention process which focused on industries (the process itself e.g. production/manufacturing including use of raw materials, transportation, design, maintenance) before and after the incidents and accidents.	Interventions surrounding alteration to the wider occupational health management structure, or separately targeting management activities. Secondary and tertiary involvement on the premise that worker's motivation levels would be different from primary participation.
Industries to include Chemical or Major hazard industry construction e.g. Nuclear plants, off-shore and on-shore oil and gas installations, chemical and major hazard production plants/factories, hazardous chemical waste and recycling plants/industries, etc.	Publications before 2007
Publications from 2007	

3.2.2. Outcome of Literature Search

The databases search produces 48914 articles. 106 publication were obtained from other relevant sources. With the aid of the selection criteria in Table 3.2b,

- 48488 articles were rejected because they were published before 2007
- 206 were found to be duplications and therefore removed,
- 257 were rejected because they refer to general occupational health and safety issues
- 45 were rejected because of title of the articles,
- 30 were rejected because they did not meet potential research aims
- 33 were rejected because they were not related to HHPIs
- 25 were rejected because they were not related to BSP
- 19 were rejected because they were review papers
- 1 was rejected because it relates to external event (flooding) instead of internal factors
- 10 were rejected because full text article reveal that they did not meet the selection criteria.

Each article reviewed was double-checked to reduce any bias. The 6 articles obtained for full review are detailed in Table 3.2c. Figure 3.2b is a preferred reporting items for systematic reviews and meta-analyses (PRISMA) chat for the selection of the citations.

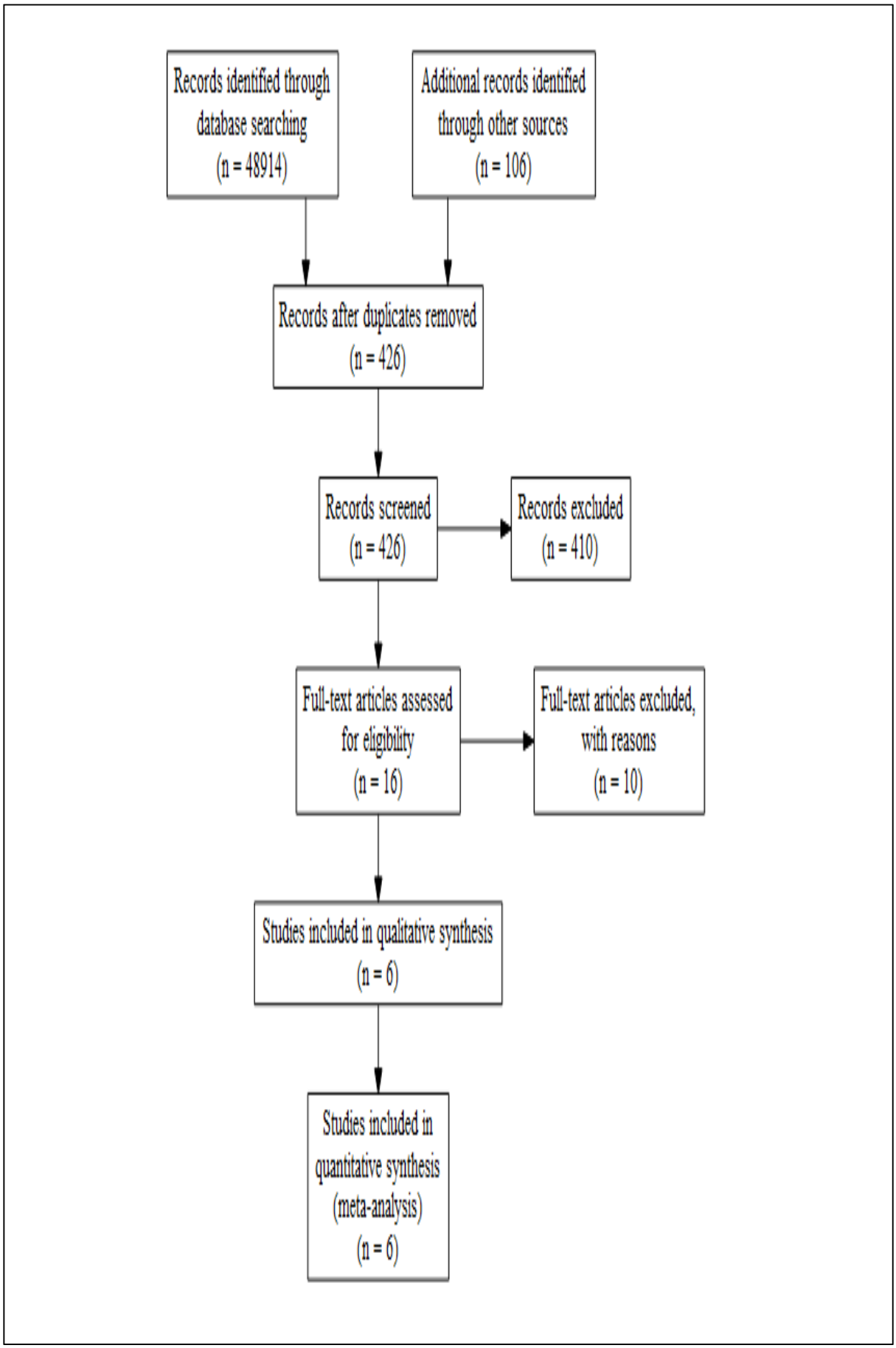


Figure 3.2b: Flow chart of the selection process, (adapted from Lunt et al 2011).

Table 3.2c: Selected articles

Articles
Coze, J.C.L. (2010) Accident in a French dynamite factory: An example of an organisational investigation, *Safety Science.* 48: 80-90.
Lekka, C. & Sugden, C. (2011) The Successes and Challenges of Implementing High Reliability Principles: A Case Study of a UK Oil Refinery, *Process Safety and Environmental Protection.* 89: 443-451.
Martínez-Córcoles, M., Gracia, F., Tomas, I. & Peiro J.M. (2011) Leadership and Employees' Perceived Safety Behaviours in a Nuclear Power Plant: A structural Equation Model, *Safety Science.* 49: 1118-1129.
Niskanen, T., Louhelainen, K. & Hirvonen, M.L. (2014) Results of the Finnish national survey investigating safety management, collaboration and work environment in the chemical industry, *Safety Science.* 70: 233-245.
Reniers, G.L.L, Cremer, K. & Buytaert (2011) Continuously and simultaneously optimizing an organization's safety and security culture and climate: The Improvement Diamond for Excellence Achievement and Leadership in Safety & Security (IDEAL S&S) model, *Journal of Cleaner Production.* 19: 1239-1249.
Vinnem, J. E. (2010) Risk indicators for major hazards on offshore installations, *Safety Science.* 48: 770-787.

Accessing effectiveness of the BSPs was adapted from a literature article by Martin Anderson (Anderson, 2005). Anderson proposes both a quantitative and qualitative approach to evaluating the programs because data on factors affecting human performance are usually insufficient. It is often the case that accidents are caused by technical problems, human error and organisation error (Vinnem 2010). As a result, the program elements for which the assessment was measured are listed in Table 3.2d in Appendix page 219.

The effectiveness of the programs in terms of how they help to reduce incidents was estimated from the evaluation of the articles with regards to whether the implementation led to the improvement of safety and minimising incidents after the implementation. Since the implementation of the elements the programs can lead to a significant reduction of accidents rates or worsen the situation than before implementation, a point ranking system of "+1" was applied where there is reduction of incidents, "0" where no change, and "-1" for worse situation than before.

The behavioural safety element together with their overall effectiveness and ranking are provided in Table 3.2e in Appendix page 219. Additional criteria such as study duration, sample size and sampling methodology were also applied with the qualifier and the point summarised in Table 3.2f, Appendix page 219.

3.2.3. Findings and Discussion of the Systematic Review of BSPs as a PSM in the HHPI's.

Three of the six selected articles, Lekka & Sugden (2011), Martinez-Corcoles et al. (2011) and Reniers et al. (2011) scored medium on the process scoring. However, Reniers et al. (2011) scored zero on the effectiveness ranking which implies no significant change in response because they were more focused on the continuous improvement of existing safety programs within the organisation. Two articles Martinez-Corcoles et al. (2011) and Lekka & Sugden (2011) had the positive effective scores of 7 and 8 respectively.

Two articles, Nisknen et al. (2014) and Vinnem (2010) had low points on the process scoring and zero effective score because the programs were not deemed effective. The lowest ranking paper for effectiveness Coze (2010) which scored -10 on effectiveness (Table 3.2g, Appendix page 220), despite scoring the highest point (12) on the process scoring (Table 3.2h, Appendix page 220). This is the only paper which fully investigate an accident by targeting the three aspects of the behavioural safety model, i.e. organisation, human and technical aspects which describe the event.

The articles reviewed were found to contain too much heterogeneity therefore making it extremely difficult to apply meta-analysis to the program elements. This heterogeneity may be found in the various elements under the aspects of the BSP covered by the articles (Table 3.2i, Appendix page 221) which also includes study duration and sample size. It was also observed from this work that not enough research has been done in relation to the HHPIs. Most of the articles also focused on the frontline staff with little or no attention to management and the organisation. The limited amount of work research done since 2007 so there is a scarcity of data for statistical analysis.

Although the articles raise awareness of behavioural change and its effect to minimise safety in the HHPIs, they did not cover other fundamentals of behavioural change including social aspect of human behaviour, which may occur outside of the organisation but could have huge impact of individual performance on a particular working day. For example, what may be classified as safe behaviour could be subjective, so workers may find it difficult to apply and uphold the program and its associated improvements without affecting social standards while working.

The review also reveals that though BSPs are expected to be positive intervention within organisations, their effect on minimising the occurrences of accidents in the HHPIs has not been well explored due to the limited research done in that field on the program since 2007. The insufficient evidence of the effectiveness of BSPs raises some concern about bias its implementation in the HHPIs. It also shows that the content of the programs may not address the purpose for which the implementation was done or a combination of these.

It would have been helpful if the effect of each parameter of the BSP is considered using large sample size with less reliance on reports which consider the main effects and its reliability but fail to measure the validity of the methods applied. Most of the incidents considered in the articles are those which were deemed to have happened just before the undesirable events or away from the event. In the case where the program was found to be less effective, the program was investigated by the authors as part of a major accident. This has led to suspicion that the true impact of the programs would be best known where similar accidents are investigated. Also, when the programs produce no effective change, the study found that the articles cover investigation which were part of continuous progress of the implementation of safety hence any positive or negative impact of the implementation of the program was not be fully obvious.

Because of the limited availability of data regarding what influence performance, not enough work has been done to quantify the aspects of human failures. Although accidents in the HHPIs are perceived to be caused by frontline staff, there is very limited consideration of security which is one of the human elements to be considered in the BSPs. This calls for better quality in the behavioural safety approach through their design, implementation, evaluation, and reporting in the industries. There have been relatively few accidents in these industries in recent years, but this could not be attributed to the effectiveness of the programs because major accidents in these industries are relatively infrequent.

The complexity of organisation safety management with regards to the HHPIs requires application of “system-oriented approach of different disciplines” (Martinez-Corcoles et al 2011) to the study of the all hierarchy management as well as frontline personnel. In this era of data deluge and the availability of IT infrastructure, there is the need for application of big data techniques to the BSPs. This may ravel the effectiveness of the existing BSPs, and any modifications that may be required.

Base on the outcome of the review of BSPs in this study, it is obvious that more emphasis must be placed on monitoring the process systems instead of behaviour of frontline personnel. As a result, the study proceeds to review of the application of big data techniques and real-time process operation data in the existing QRA methods which is the focus of this research.

3.3. *Big Data Techniques and Real-time Data as QRA Methods: A Systematic Review and Content-Analysis*

A typical QRA method involves using numerical data to identify and eliminate accident sources. For the HHPI, the QRA method applied may depend on the nature of the hazard, the risk criteria usually set by the regulator, the conditions of the facility, and the technology being used (Gadd, Keeley & Balmforth 2003). Because of the uniqueness of the process, the nature of the hazard

varies from one facility to the other. Hence for the method to provide good estimate and quantify the magnitude of the hazard, the QRA method use must be sensitive to the nature and impact of the hazard. In the context of this study, the term 'method' refers to techniques and procedures applied to obtain and analyse the data (Saunders, Lewis & Thornhill 2007).

For a QRA process to achieve set objectives, regulators have set up protocols which covers various criteria relating to the risk associated to the processes undertaken by the facilities for which the model is applied (Hart 2002). One of such criteria requires the QRA methods to be tolerable to the technology used and conditions of the facility since these conditions could be static or dynamic (Allocco et al. 2016). For instance, a facility may be static but the activities within could involve mobile elements e.g. a transport bringing in hazardous raw materials.

However, the there are some limitations associated with the methods in that although they are expected to utilise high-quality data, they sometimes fail to consider for some of specifics of the facility for which they are applied (Patel & Sohani 2013). Other limitations such as cost, schedule and performance which may satisfy fit-for-purpose requirements has also been reported (Barondes 2012). Some authors argue that QRA models lack the suitability because of the challenges with conducting controlled experiments to verify QRA predicted risks (Rae, McDermid & Alexander 2012). They also reported other limitations relating to critical data voids and daunting tasks of updating data, data validity, uncertainties and assumptions associated with data and their effects on the methods, data analysis and statistical techniques used for execution and implementation. The limitations make the QRA process time-consuming and requires the involvement of safety expert and peers which could lead to expert bias and thereby making the findings of the methods less reliable.

In the light of the limitations listed above, there is the need for an alternative approach to the QRA process which involves the application of data obtained from the operation of the process itself and more statistical and data analysis techniques which could eliminate most of the limitations. As a result, this study performs a review and systematic content-analysis of published literature on the existing QRA methods to ascertain the use of alternative technique like big data techniques and real-time data from a process operation.

This aims to address the review question, "How are big data techniques and real-time process data applied in the existing QRA methods used in the HHPIs?" To help answer this question, the study will investigate existing literature relating to the existing QRA methods applied in the HHPI's by evaluating the

- methods
- big data techniques used

- type of data used
- contents and procedures used
- overall outcomes.

3.3.1. *Problem Statement and Hypothesis of the Review*

Although the current QRA methods have improved, serious accidents (e.g. explosions, chemical releases) still happen (Chen et al. 2010). As a result, there is a need to re-assess the methods and consider other methods of dynamic dimensions or a combination of different methods (Paltrinieri et al. 2013). Some of the existing methods including dynamic risk assessment (DRA), risk barometer (RB), and dynamic procedure for atypical scenarios identification (DyPASI), have been found to provide insight into real-time events and help prevent undesirable outcomes (Villa et al. 2015). However, despite their effectiveness, they are not extensively used in the HHPIs.

Though the existing QRA methods are widely applied within the HHPIs, the adaptation of the methods to use real-time process monitoring data is either less understood or sparingly explored. In this era of automation and advanced technology where deluge of data is produced and collected as part of the activities within the HHPIs, there is the need to investigate and adapt the existing QRA methods to use big data techniques and the deluge of process operation data generated.

Due to the uniqueness of activities within the HHPIs, leading indicator of risks depends on several factors and may develop over long time periods. There are numerous sources of data including safety audits and studies therefore designing and conducting a QRA to detect possible risk events could be difficult but still possible. This is due to the web of industrial processes and the several different components which must be investigated over a long period of time; data protection issues; and companies not willing to divulge data relating to their processes.

The study intends to use real-time process monitoring data because they contain tremendous information which could provide insights into events within the process and can help detect the risk of catastrophic events. However, because the existing QRA methods have been extensively reviewed by other researchers since COMAH 1991 including the review done and published by Patel in 2013 (Tixier et al. 2002; Patel & Sohani 2013), the study considers citations from 2007 to help provide insight into any new knowledge which may not have been covered by the previous reviews.

3.3.2. Finding Previous Reviews Relating to the Review Question

The effort was to establish an existence of a previous review to ensure that the study performs an independent review of literature to help avoid creating a duplicate of existing work, save time and avoid wasting resources in producing and reporting research evidence (Chalmers & Glasziou

2009). It was revealed that there are two reviews of literature on existing QRA methods by Tixier et al (2002) and Patel & Sohani (2013). After careful consideration, it was observed that the work of Patel and Sohani was an update on that of Tixier at al. As a result, it was decided the review should perform a review to update and improve on the findings of Tixier et al. (2002) and that of Patel and Sohani (2013) by focusing on the application of big data techniques and real-time data in the methods.

The study therefore applied the search query "(Real-time process data) AND (Big data methodology) AND (Quantitative Risk Analysis) AND (Process Industry) AND (Systematic review) AND (Meta-analysis OR Content Analysis)" was applied. The databases listed under Section 3.0 were search including sources of valuable information on incident occurrences and/or their consequences such as historical databases (Prem, Ng & Mannan 2010). The list of various databases and the reasons for the search are detailed in Table 3.3a, Appendix page 222. Of the databases search, IEEE explore was found to have the highest number of review citations of 8993 (Table 3.3b, Appendix page 223). After application of the predefined filters, including additional filter for review/literature review and content analysis, it was established that there is no existing literature review and content analysis which meets all the set criteria. As a result, the study proceeds with a search for peer reviewed publications relating to the focus topic.

3.3.3. Searching for Publication for Literature Review and Systematic Content-analysis

Search strings are applied by combining the keywords (real-time and data), with the assessment terms (identify, investigate, predict, assessment, system, model), result related terms (safety, hazard, prevent, protect, danger, benefit, impact), then industry type terms (chemical, hazardous chemicals, product, manufacturing) and other related terms (environment, pollution, community). Wildcards were also used to include international spelling variations.

The search strategy was developed from a generic search “data informed risk detection in HHPI”, which was then expanded to help generate an independent review. Effort was made to ensure that generic terms which may lead to random and non HHPI citations are excluded. For instance, the terms “natural disaster” or “risk from terror attacks” were excluded as they are more likely to generate citations which are not exclusively related to HHPI’s.

Although some catastrophic incidents from the HHPIs could have an impact on the ozone layer and transportation, the terms "global warming," "atmospheric carbon dioxide" and "supply and transport" were also excluded. This decision was meant to exclude terms which are not purpose specific and help reduce the volume of citations generated from applying the search terms.

The existing QRA methods has been classified by previous reviews into deterministic, probabilistic, and a combination of deterministic and probabilistic methods, based on type of output data (Tixier et al. 2002; Patel & Sohani 2013). The deterministic methods incorporate data from the process, products, and quantification of consequences. The probabilistic methods incorporate data such as probability/frequency of accident, with the focus on failure probability of equipment/ equipment components. The combined deterministic and probabilistic methods are applied to investigate the entire process site. Table 3.3c in Appendix page 223 is a list of the class of existing QRA methodologies.

Tixier also explain that the procedures for method selection for a QRA as are based on relationship between available input and output data which has been summarised by Figure 3.3a. Thus, the selection of the methods is based on the 'user expected' outcome and available data. So based on expected result, the QRA assessor reads through the result column of the chart for output data, followed by the proposed methodologies, and then the column of input data to identify the input data for the analysis.

The criterial for inclusion and exclusion applied are provided in Table 3.3d, Appendix page 224. was applied to eliminate The final search string was (Process-specific OR real-time process OR Batch process* OR Chemical* Process OR Chemical* reaction OR Product*) AND (Data OR Inform*) AND (Identify* OR Investigat* OR Detect* OR Analys* OR Assess* OR Model*) AND (Risk* OR Safe* OR Hazard* OR Danger* OR Impact* OR Protect* OR Prevent*) NOT (Environment* OR Pollut* OR Communit*) NOT (Construction OR General health and safety OR Medical OR Biological OR Natural disaster) NOT (Disease OR Illness OR Sickness) NOT (Transport* OR Supply OR Consumer* chain OR Behaviour* OR Manager*) NOT (Climat* change OR Global warming OR Greenhouse OR Atmospheric carbon dioxide) NOT (Cyber OR Internet OR Terror attack* OR Natural disaster). The search strings and their corresponding outcome are provided as Table 3.3e, Appendix page 224.

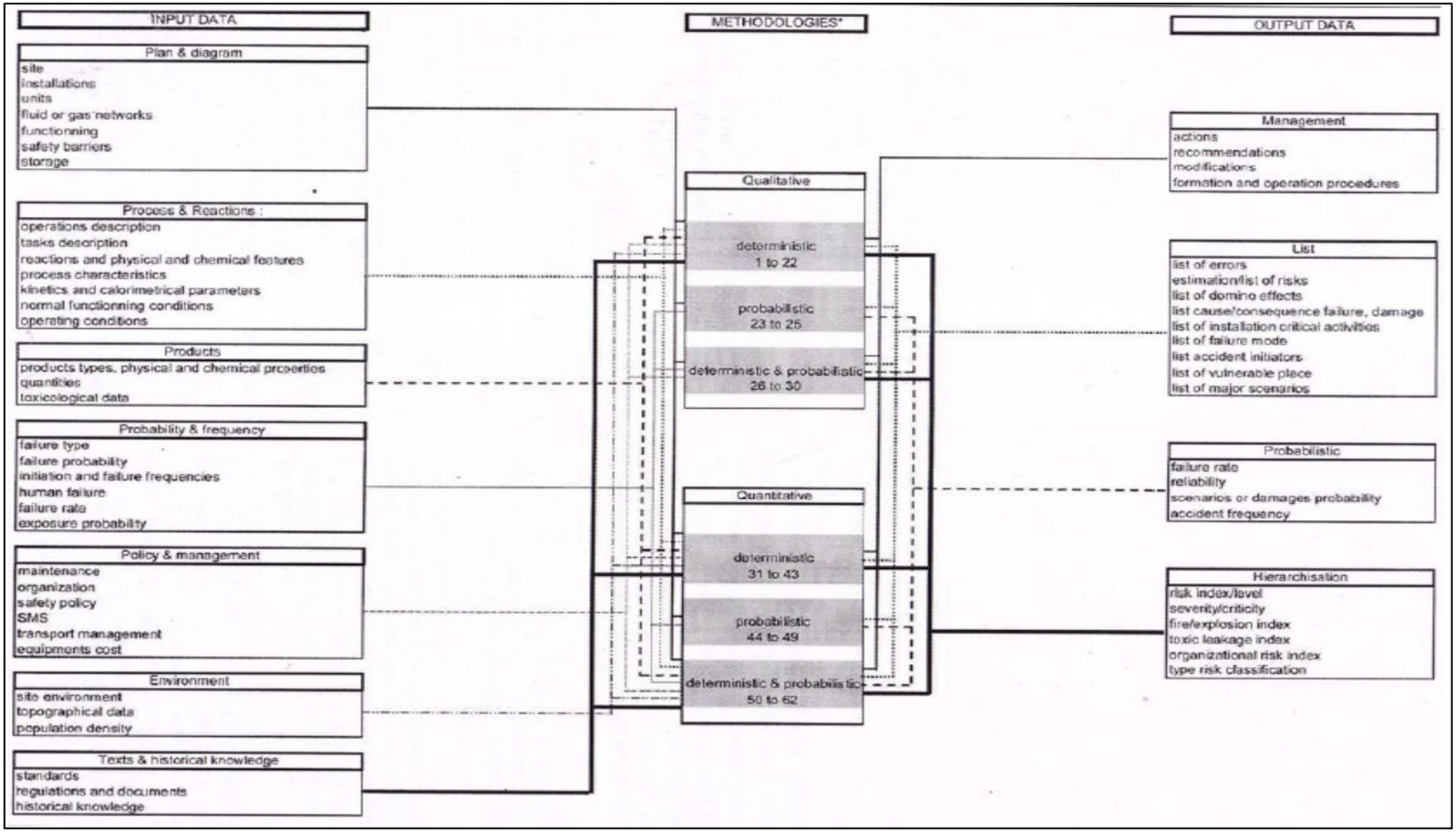


Figure 3.3a: Relationship between available input and output data and techniques within system (Source: Tixier et al. 2002)

3.2.3a. Findings from the Search for Publications for Literature Review

The search gave 5124 published articles. Of the publications obtained,

- 4856 papers were excluded after the application of filters.
- 74 were duplicates and therefore rejected.
- 177 publications were removed after abstract screening was applied.
- 6 publications were rejected after application of the predetermined inclusion and exclusion criteria.
- 11 publications obtained for synthesis and content analysis.

Figure 3.3b is the PRISMA flow diagram for the appraisal process. Table 3.3f in Appendix page 225 is a list of the journals and the corresponding number of research publications found by the study. The 11 publications obtained for synthesis and content analysis after the appraisal are listed in Table 3.3g. The small sample size of the publications obtained could be due to the QRA methods being designed to meet standard and knowledge of safety and engineering with very minimal statistical and big data applications. The methods are expected show improved engineering, procedures and supervision to prevent the calculated accidents from happening (Veritas 2001).

The study also found that combined deterministic and probabilistic (CDP) methods was the preferred class of method applied in all 11 citations (Table 3.3h, Appendix page 2269). In addition to this, there is some evidence of evolution of the QRA methods. For instance, the event coloration analysis (ECA) applied in some of the publication (Nishiguchi and Takai 2010; Noda, Takai & Higuchi 2012) is not part of the existing 31 QRA methods reviewed by Texier. Also, a combination of event tree analysis (ETA) and fault tree analysis (FTA) was applied in in 6 publications, the FTA alone was applied in 1 publication. The DEFI method, maintenance analysis, shortcut risk assessment and the WPAM were not applied. Thus, one could propose that the QRA in the publications applied were based on accident occurrence probabilities.

To recap, both Tixier et al (2002), and Patel and Sohani (2013) have classified the existing QRA methods into three as follows:

- deterministic – a method which incorporate data from the process, products, and quantifies risk consequences.
- probabilistic – a method which incorporate data such as probability/frequency of accident, with the focus on failure probability of equipment/ equipment components.
- combined deterministic and probabilistic - methods which are applied to investigate the entire process site.

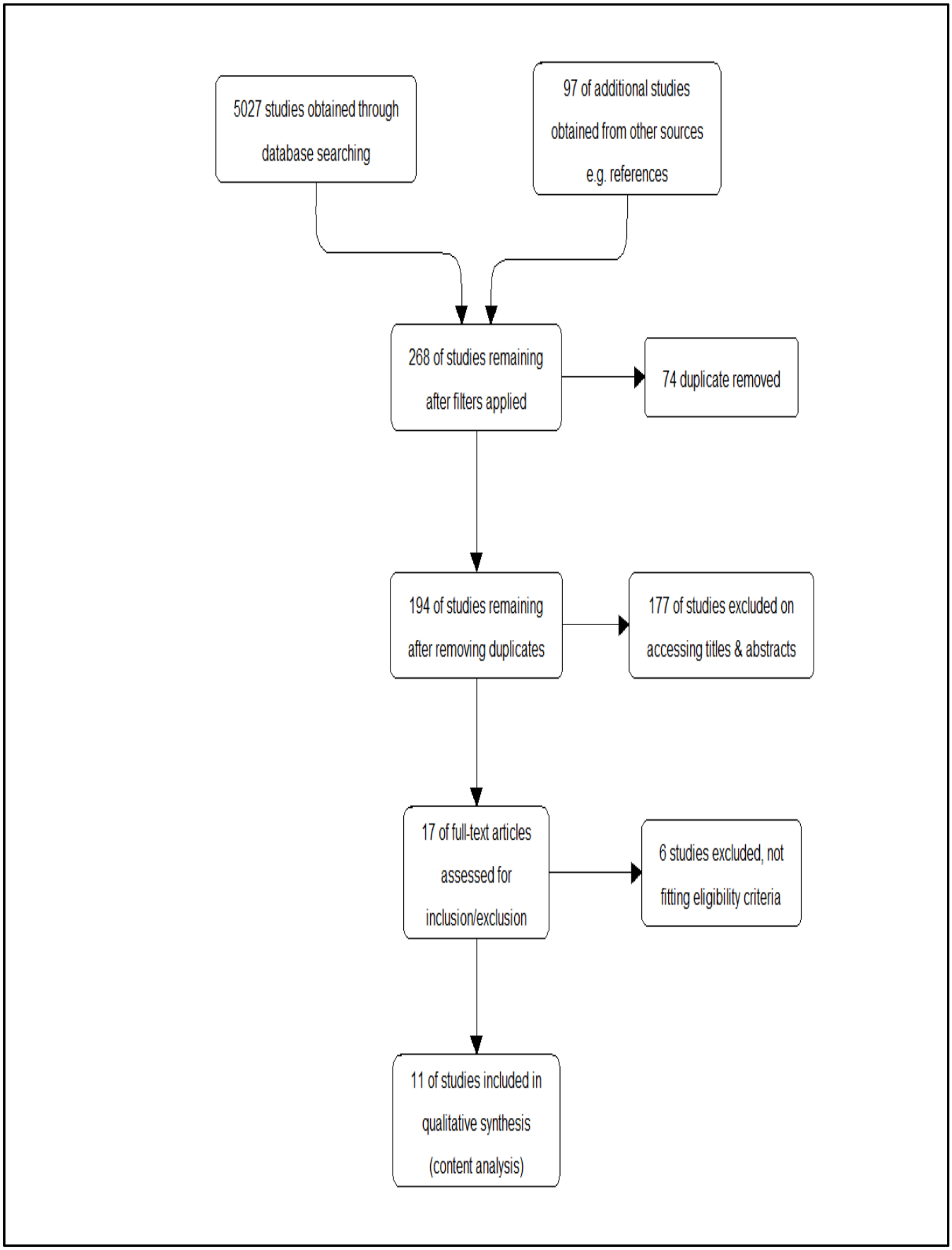


Figure 3.3b: PRISMA flow diagram of systematic appraisal of cited papers.

Table 3.3g: Selected Publication for Review

Articles
Hou, Z. and Zhao, P. (2016) Based on Fuzzy Bayesian Network of Oil Wharf Handling Risk Assessment. *Mathematical Problems in Engineering*, 2016: 1 - 10.
Kalantarnia, M., Khan, F. I., Hawboldt, K. (2010) Modelling of BP Texas City refinery accident using dynamic risk assessment approach. *Process Safety and Environmental Protection*, 88(3): 191-199.
Khakzad, N., Khan, F. and Amyotte, P. (2012) Dynamic risk analysis using bow-tie approach. *Reliability Engineering and System Safety*, 104: 36-44.
Nishiguchi, J. and Takai, T. (2010) IPL2 and 3 performance improvement method for process safety using event correlation analysis. *Computers and Chemical Engineering,* 34: 2007-2013.
Noda, M., Takai, T. and Higuchi, F. (2012) Operation Analysis of Ethylene Plant by Event Correlation Analysis of Operation Log Data. *In Proc. of FOCAPO*: 8-11.
Paltrinieri, N., et al., (2013) Dynamic approach to risk management: Application to the Hoeganaes metal dust accidents. *Process Safety and Environmental Protection,* 92(6): 669-679.
Shahriar, A., Sadiq, R. and Tesfamariam, S. (2012) Risk analysis for oil & gas pipelines: A sustainability assessment approach using fuzzy based bow-tie analysis. *Journal of Loss Prevention in the Process Industries*, 25: 505-523.
Tobon-Mejia et al. (2012) A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models. *IEEE Transactions on reliability*, 61(2):.491-503.
Wu et al (2016) A DBN-based risk assessment model for prediction and diagnosis of offshore drilling incidents, *Journal of Natural Gas Science and Engineering,* 34: 139-158.
Wang et al. (2015) Quantitative Risk Analysis of Offshore Fire and Explosion Based on the Analysis of Human and Organizational Factors. *Mathematical Problems in Engineering*, 2015: 1 - 10.
Yin, S., Yang, X. and Karimi, H.R. (2012) Data-Driven Adaptive Observer for Fault Diagnosis. *Mathematical Problems in Engineering,* 2012: 1- 21.

3.3.4. Assessing *Content and Quality*

The performance of the QRA methods were accessed base on factors including data (amount and quality), human or technical errors or physicochemical properties which may correlate to incidents (Vinnem 2010), any desirable outcomes obtained, any assumptions made. This in-depth assessment of the models was performed to help highlight any bias that may affect the quality of the research. Where a bias is suspected, the evidence was carefully considered or redefine to help assess its impact on the study conducted by the publication.

Since this study focuses on data use, the assessment of the methods considered factors like data collection methods, data storage and management to ensure the accuracy of the data was not compromised. Appropriate presentation of data, data validity and analysis (big data and statistical) were also examined for clarity of presentation by the publications. Some researchers have proposed that QRA studies must determine exposures and outcomes with minimal misclassification of risk (Khan et. Al. 2003). As a result, the methods in the publications are expected to detect or predict the risk and hazard exposure characteristics. Exposure assessment are expected to include the route by which the resources, personnel, the population and environment are exposed to the hazard. This depends on the nature and source of the hazard and differ from hazard identification hence the methods are expected to track hazard and estimate the likelihood of its devastation.

The outcome of the existing QRA methods applied in the publications are also expected to include sensitivity and uncertainty analysis so the study will assign values for this feature based on desirable/undesirable outcomes. Understanding uncertainties and their causes helps to effectively interpret risk detected (Gadd, Keeley & Balmforth 2004). Uncertainties associated with the method selection may also have effect on the outcome hence must be managed via sensitivity analysis (Freeman 1990). Because the QRA methods are generally prescriptive base on expert knowledge in safety, engineering and other engineering disciplines, this study expects that confounding factors to may be missed.

Confounders have the potential for biasing relationship between variables (Sullivan 2016; McNamee 2005). However, the QRA methods applied in the publications may include factors for reducing the effect of confounder variables. This will help to interpret the outcomes with greater clarity (Pourhoseingholi, et al. 2012; Sullivan 2016; Weinberg 1993) and help explore alternate explanations of observed relationship if the cofounders are controlled (Christenfeld 2004).

Although the study focuses on the HHPIs, other industries covered by the publications will be considered and grouped to aid the comparison like an onshore process activity which involves hazards that may not be found in a batch manufacturing process in the HHPI. Hence the review would compare the impact of the QRA within and between the groups. The methods are also expected to generate results which could allow comparison with criteria set by the regulator or the operator (Ristec 2008).

3.3.5. Content and Quality Assessment Results

To aid with the effective quality assessment, 9 objectives based on the quality categories and other parameters were set and assessed.

Objective 1: Objective of the Research Covered by the Citations

The study investigates the research objectives of the Publications and applied a point of “+1” for each objective. This is because the objectives of a research provide the understanding to observations, findings, conclusions of the work done. For this, the study reveals that the objective of the work covered by the publications varied from complicated as in Hou et al. (2016) to a single objective as found in the other 10 publications. In all, the publications focused on a total of 21 objectives. The study found that Hou et al. (2016) registered 8 of the objectives which constitute 38.1% of the total objectives, Paltrinieri et al. (2013) recorded 3 (14.3%), while each of the remaining publications recorded 1 (4.8%) objective each. Due to the diverse number of research objectives observed, no points were awarded for this objective. The publications and their corresponding objectives are listed in Table 3.3i, Appendix page 226.

Objective 2: Research Methodology

Because the methods may involve surveys, case studies, field studies, experiments or a combination of all four research approaches to capture data as part of the investigation to help minimise process risk. A point of "+1" was assigned where three or more research methods were used, a point of “0” where two methods were used and "-1" where less than two methods were applied. The study reveals that the case study approach was applied in 9 (81.8%) of the 11 publications with the remaining 5 (36.4%) publications using experimental, surveys and a combination of other methods as detailed in Table 3.3j, Appendix page 227.

Objective 3: Risk Detection

Assessing risk detection of the methods as covered in the publications, the study assigns a point of "+1" for risk detection, "0" where the method did not detect risks. The study reveals that 6 of the publications (54.5%) have risk detection characteristics as detailed in Table 3.3k, Appendix page 227.

Objective 4: *Data Use*

Generally, the type of data from the HHPI may range from operation data; process design diagrams; process site plan; real-time data type and other data that is relevant to activities of the industry (Table 3.3l, Appendix page 228). However, we focus of this review is on processes activities data such as data generated while the process was in operation. These include data relating to the operation, production/reaction, toxicity and instrumentation conditions during operation. Table 3.3m Appendix page 228 is the type and amount of data with the corresponding publication.

Assigning "+1" for use of process data and "0" for data other than that of the process operation, the study found that 5 of the publications (45%) applied data from process operation for the QRA. The study also found that 6 of the publications (54.5%) applied operation and P&ID data, 4 publications (36.4%) applied only operation data, and 3 publications (27.3%) applied operation, production and P&ID data. In terms of amount of data use (sample population or size), the study found that 6 of the publication (54.5%) use sample size of up to 50, 3 publications (27.3%) applied sample size > 50 <1000 and 4 publications (36.4%) applied sample size > 1000. The study therefore assigning points of "+1" for sample size greater than 1000.

Objective 5: Application of Statistical Analysis Techniques

Investigating the use of statistical analysis techniques used as part of the QRA method by the publications, the study found that all together a total of 22 different statistical analysis techniques were applied. They include multiple-criteria decision analysis (MCDA), cluster analysis (CA), event correlation analysis (ECA), Poisson point process (PPP), Pearson correlation (PC), Spearman's rank Correlation (SRC) Bayesian statistics (BS), Prior distribution (PD), Posterior Probability (PP), Binomial Statistics (BS), Multivariate Probability theory (MP), hidden Markov model (HMM), Entity-Property-Relationship (EPR), and the kernel density estimation (KDE) method, each of which represent 4.5% of the statistical analysis method applied. Linear Regression method (LR) which represent 13.6%, Maximum likelihood estimation method (MLE) and Baye's theorem (BT) were45% each, Gaussian process (GP), Bayesian network and Probabilistic methods (PM) were 9% each.

The study also reveals that Paltrinieri, et al. applied the highest number of 5 statiatical analysis techniques which constitutes 22.7% of the total statistical methods used by the publications. this is followed by Khakzad et al. (2012) and Shahriar et al. (2012) applied 3 techniques (13.6%) each, then Tobon-Mejia et al. (2012), Wang et al. (2015) and Wu et. al. (2016) who applied 2 techniques (9%) each. Hou & Zhao (2016), Kalantarnia et al. (2010), Nishiguchi & Takai (2010), Noda et al. (2012) and Yin et al. (2012) applied the least number of statistical techniques each i.e. 1 statistical technique (4.5%) each. The study assigns a point of "+1" was applied for each of the statistical methods applied. The details of the findings are presented in Table 3.3n, Appendix page 229.

Objective 6: Method Validation

Investigating method validation applied in the publications to understand how the methods help to achieve their intended purpose, the study considered the objective of the research and the procedures used. The study found that 81.8% of the publications use single case studies to test tested the validity of their method, 9.1% use experimental data, and the remaining 9.1% did not clearly explain how their method was validated. The study assigns points for clarity, definition and

explanation of the data and procedure used in the models using "-1" where data analysis and validity approach were not applied, "0" if the approach was found to be relatively unclear; and "+1" for the well explained or defined data analysis and validity method. Details of the outcome of this objective are presented as Table 3.3o, Appendix page 229).

Objective 7: Handling Uncertainty

Investigating how the method handle and manage uncertainty to help understand their reliability, the study found that only 18.2% of the publications address the issue of uncertainty as part of their method. The rest of the studies did not consider issues relating to uncertainty. The study therefore assigns "+1" where method in the publication considers uncertainty as part of their methods and "0" where information on uncertainty were not provided (Table 3.3p, Appendix page 230).

Objective 8: Perceive event (Accident scenarios)

Investigating nature of perceive events and their definition which help understand the application of the methods, the study found a total of 20 different events reported in the publications. Of the 20 events, 15% relates to fire or explosion, 15% relates to fuel and gas release; 15% relates to faults in design or equipment, 10% relates to maintenance of equipment, 10% relates to inadequate housekeeping, 10% relates to deficiency in safety measures, 10% relates to alarm events, 10% relates to degradation or deterioration effect, and 5% relates to tripping activities.

Khakzad et al.(2012) reported the 18% of perceived events, Paltrinieri, et al.(2013), Shahriar et al. (2012) and Wang et al. (2015) reported 13.6% events each, Wu et al. (2016) reported 9% of the events, and Hou & Zhao (2016) Kalantarnia et al. (2010), Nishiguchi & Takai (2010), Noda et al. (2012), Tobon-Mejia et al.(2012) and Yin et al. (2012) reported 4.5% of the events each. Since one dataset was used by the authors of all the articles investigated, no points were assigned for this objective.

Investigating the number of perceive events as a population by the publications due to the impact of the hazards, their undesirable consequences, the amount of data available, the study assigns a point of "+1" was for all of the events reported by the publications (Table 3.3q, Appendix page 230).

Objective 9: Research Limitations/implications

Finally, the study investigates if the publications reported any research limitations and any unanswered questions or recommendations for future research by considering any impact of the limitations on the outcome and conclusions. This was performed by investigating(a) studies limitation and (b) method limitation, for which the study assigns a point of “+1” where research

limitations are explained and “0” if the authors did not provide any information regarding research limitations. The study found that method limitations were recorded by 41.7% of the publications, data limitations were reported by 8.3% of the citations, both method and data limitations were reported by 25% of the citations, while 25% of the publications did not provide any information relating to limitations of their work (Table 3.3r, Appendix page 231).

3.3.6. Ranking

The citations were ranked by the balance of positive and negative points. Where the total negative cancels the total positive or produces a negative, the article was given a low ranking. Where a positive non-zero result was obtained but less than 5, the article was given a medium ranking. For differences greater than 4 the article was rank as high (Table 3.3s, Appendix page 231). It was observed that 9% of the articles recorded low ranking, 72% of the articles were ranked as medium and the remaining 18% of articles obtained high ranking.

The study proceeds with investigation of how the publications are linked using text analysis with much emphasis on the most frequent words which may describe the research projects and any potential classification

3.3.7. Text Analysis

The study initially applied R-package ‘tm’ to manage and scan text in the files, convert the text into a corpus, and create term-document matrix (Feinerer 2017). However, it was discovered that ‘tm’ could not read pdf files so an external package *xpdf* engine which has been recommended (Ford 2016) was acquired and installed. After the installation, it was noted that every attempt to apply the functions ’*readPDF*", ’*pdftotext*’ and ’*pdfinfo*’ to read the texts from the pdf files produce error messages. After several unsuccessful attempts, further information was sourced from various internet sites from which a R script written by Ben Marwick on the procedure to resolve the issues was found on the web page – Electric Archaeology (Marwick, 2015; Graham, 2017).

The PDF files were read into R, then converted to text and renamed. The text files were inspected to ensure that the conversion was successful. The text data was exported into a data frame of 11 observations (representing each article) and two variables (for source and text), then saved as a csv file. A corpus was created and inspected after which corpus cleaning was performed to strip whitespaces, remove single characters, and numbers. A stop-word list prepared and applied to remove unwanted words. The cleaned corpus text was compared with the source corpus to help check whether the cleaning was successful.

After the successful corpus cleaning, a document text matrix (DTM) was created and analysed. The final cleaned textual data shows that the DTM has 11 documents and 3877 terms which

appeared at least once, with a sparsity of 81% which suggests that words which appear only 19% has been removed (Figure 3.3c).

```
DocumentTermMatrix (documents: 11, terms: 3877)
Non-/sparse entries: 8155/34492
Sparsity           : 81%
Maximal term length: 25
Weighting          : term frequency (tf)
```

Figure 3.3c: Text mining DTM data after corpus cleaning

Investigating the DTM with a word cloud (Figure 3.3d) reveals that the most featured words are risk, analysis, and failure, all of which describes the documents within the matrix. As a result, the text DTM was deemed satisfactory.

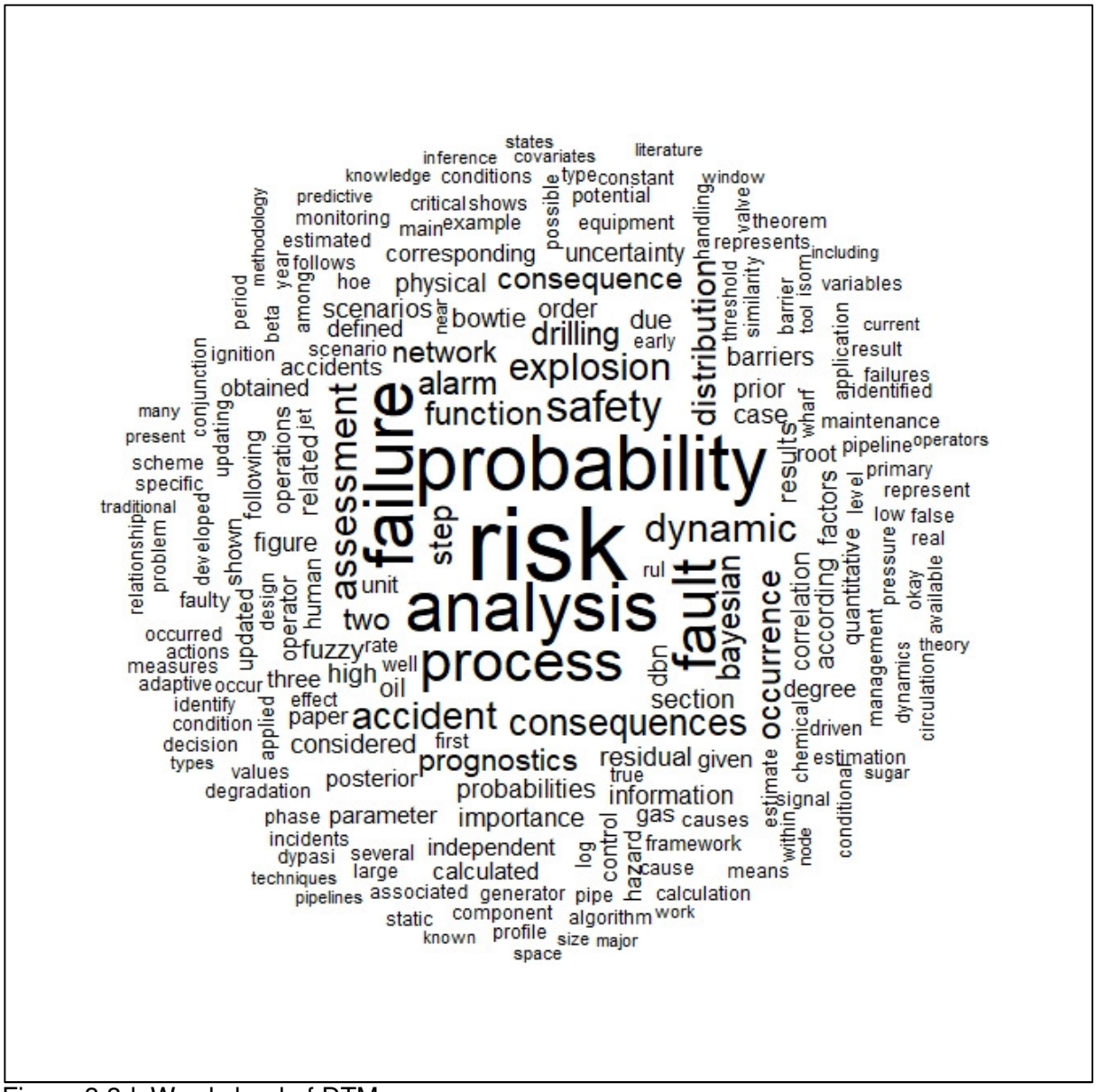


Figure 3.3d: Word cloud of DTM

3.4. Conclusion

This review publications on BSPs as part of safety management within the HHPIs which gave the insight that instead of monitoring behaviour of personnel working the focus must be on the process itself. The chapter then proceed with a systematic review and content-analysis of literature on use of real-time process data and big data techniques for QRA methods. Initially effort was made to find previous publications about the chosen review question. However, the search reveal that there has been no previous review and content analysis.

As a result, the systematic review and content-analysis was necessary prior to performing the actual research. From the review and content analysis perform by this study, it was evident that very little use of real-time data and big data techniques has been applied for QRA in the HHPI. Where statistical methods were applied, there is evidence of different approach to the use of data even when similar statistical were applied. For instance, this study found that although Wang et al. and Wu et. al. applied the BN for QRA for the same target industry, there was the difference in the number of target events and the amount of data used, with one researcher considering model uncertainties and their handling while the other didn't.

There is no evidence of the use of enough process monitoring data with most of the publications using only one dataset for their research. This means the methods were not applied to other datasets to explore their robustness. In most cases the data were presented in pictorial format with a very limited explanation on data collection, storage and handling. The study also found that the publications applied one case study approach with pre-defined phenomenon for method validation. Although the use of single case study validation of the models provides a satisfactory relationship to the researchers, the question of researcher subjectivity may apply. Also, the use of single case studies as a research method has been criticised by other researchers in the past due to its limitations (Cavaye, 1996). For instance, it has been explained that a typical single case study investigation aims to contribute to knowledge by relating findings to generalizable theory and does not necessarily define a priori constructs and relationships (Cavaye, 1996).

Considering the amount of data generated from the operations of the HHPI and the findings of this review, the study conclude that there is evidence of knowledge gap in the use of real-time data and big data techniques for QRA. Next is Chapter 4 - Real-life Case Histories, where two real-life case histories of industrial incidents are presented as part of the study to provide insight into some risk events within HHPI's. The report of one of the case histories will be critiqued to help ascertain whether available dataset for the research can be justified prior to its application.

Chapter 4 - Real-life Case Histories

4.0. Introduction

In Chapter 3, the study presents a review of research publications on behavioural safety programs (BSP) as applied for managing safety in the HHPI and found that the focus of safety must be on the process itself. As a result, the study performs a systematic review and content analysis of research publications on how big data techniques and real-time process data are used in the existing QRA method. The study now presents two real-life case histories relating to process safety incidents to provide an insight into some of the risk events within HHPIs. The study will critique the reports of the investigation of the incidents to help ascertain whether there is a justification for using available dataset for the research. Figure 4 is an illustration of the framework for this Chapter.

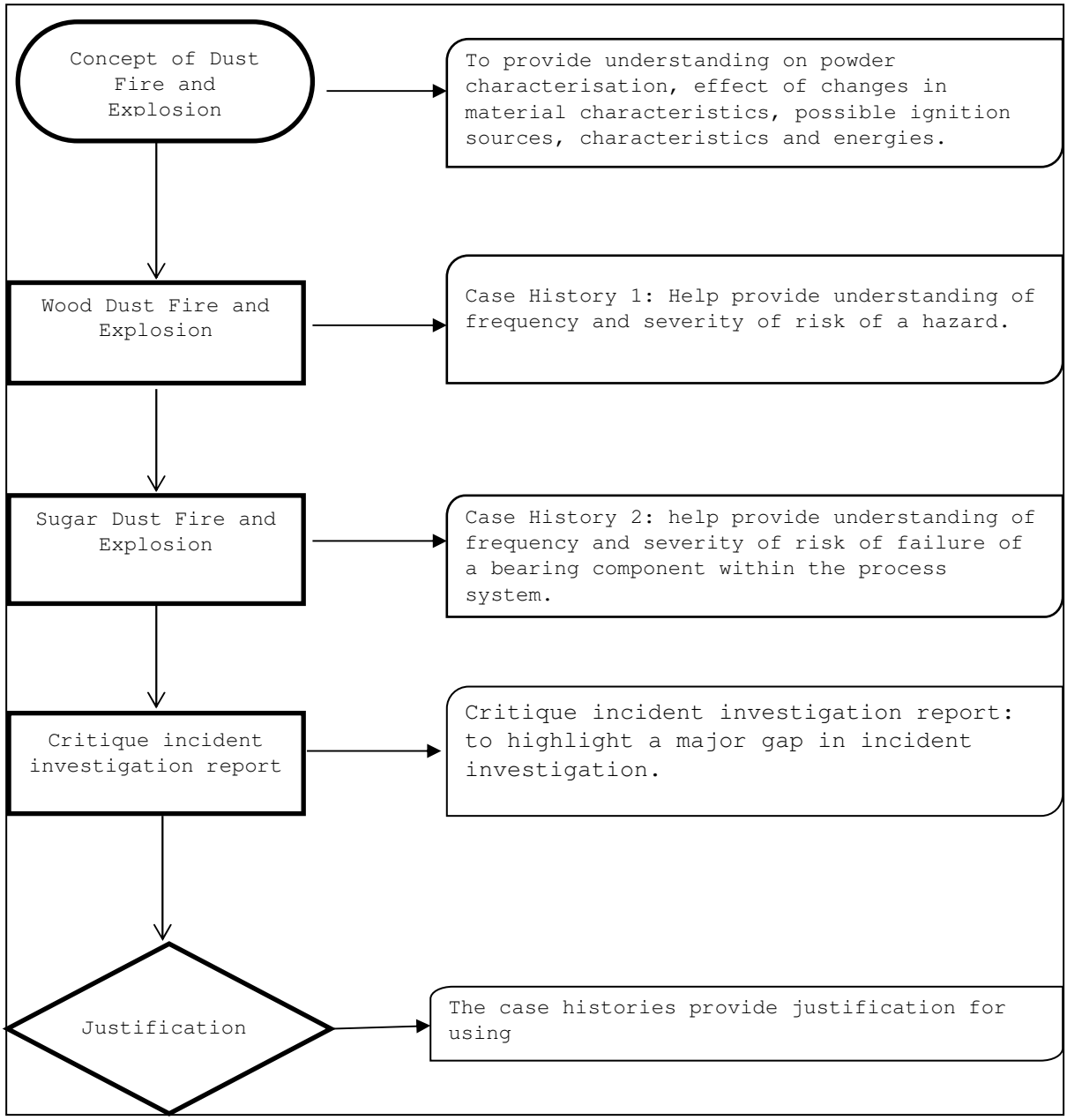


Figure 4: Flowchart for the framework of Chapter 4

4.1. Basic Concepts of Dust Fire and Explosion

Solid material can be combustible depending on their characteristics. Even where the material is known to be not normally combustible, they can burn or explode when their physical or chemical characteristics including size and concentration changes. As a result, this study covers some of the basic concept of dust fire and explosion to provide some understanding to the reader about why some known non-combustible materials can burn or explode.

4.1.1. Dust Fire and Explosion

Dust explosions occur when combustible dust obtains the three elements required for fire and two other elements as listed below (OSHA 2018b)

- Elements required for fire - Fuel, oxygen and ignition source.
- Other elements - appropriate dispersible dust particle concentration and confinement for the dust cloud.

These can be represented by the dust pentagon (Figure 4.1a). the terms 'dispersion and confinement' refers to a suspension of the dust particles in the air in an enclosed space which causes pressure to build up thereby increasing the likelihood of an explosion. Thus, dust explosions require suspension of fine particles released at an appropriate concentration within explosive range (i.e. release enough heat energy) that can sustain combustion fire in the presence of optimum oxygen concertation in a confinement that allows enough pressure to build. Thus, removing any of the elements from the classic fire triangle or explosion pentagon eliminates the possibility of a dust fire or explosion.

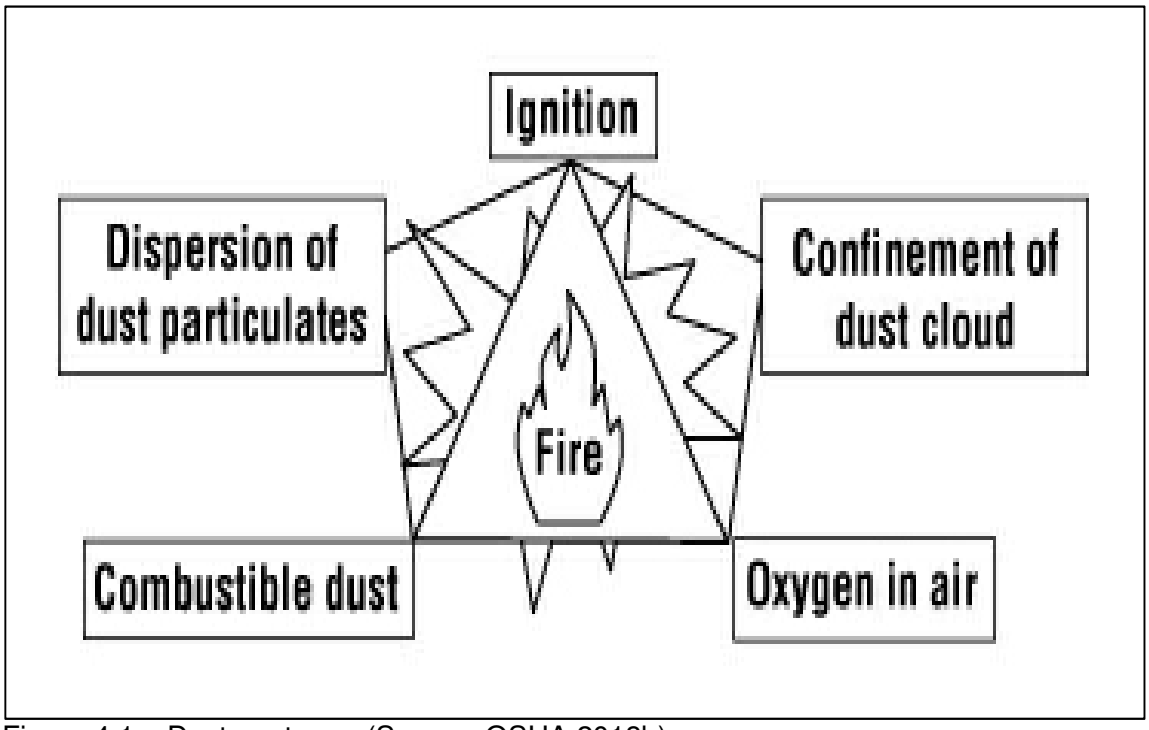


Figure 4.1a: Dust pentagon (Source: OSHA 2018b).

Explosible dust cloud has a potential hazard even in minute layers or concentrations. Ebadat (2017) has proposed that even a 1 mm layer of a dust of bulk density 500 Kgm^{-3} can generate a cloud of average concentration 100 gm^{-3} when dispersed in a room of 5 m height, and a partial dispersion of combustible dust to 1m may result in a 500 gm^{-3} (Ebadat 2017). Ebadat uses equation 1 below to expresses the concentration of a combustible dust cloud using the powder bulk density, the thickness of the dust layer and the height of the dust cloud in a confined environment as:

$$C = P_{bulk} \times \frac{h}{H} \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where:

C = concentration of a combustible dust cloud

P_{bulk} = powder bulk density

h = thickness of the dust layer

H = height of the dust cloud in a confined environment

4.1.2. Combustible Dust

Combustible dust refers to any dust that has the potential of becoming combustible under specific situations. They include agricultural products (e.g. wood dust, pesticides), metals (e.g. aluminium), chemical and pharmaceutical products, etc. The list of combustible dust can be extensive, some of which can be found in the list prepared by the US OSHA (OSHA 2018b). A sample of the list is provided as Table 4.1a.

Table 4.1a: Some examples of combustible dust. Source: OSHA (2018b)

Products	**Examples**
Agricultural Products	Egg white; Powdered milk; Starch (corn, rice & wheat); Sugar (milk & beet); Tapioca; Whey; Wood flour
Agricultural Dusts	Cocoa (bean dust or powder); Cottonseed; Garlic powder; Grass dust; Green coffee; Hops (malted); Lemon (pulp or peel dust); Linseed; Locust bean gum; Malt; Oat (grain dust or flour); Olive pellets; Onion powder; Parsley (dehydrated); Peanut (meal or skins); Potato (flour or starch); Rice (flour, starch or dust); Rye flour; Soybean dust; Spice (dust or powder); Sugar dust; Sunflower seed dust; Tea; Tobacco blend; Walnut dust; Wheat (flour, grain dust or starch)
Carbonaceous Dusts	Charcoal (wood or activated); Coal (bituminous); Coke (petroleum); Lampblack; Lignite; Peat (22% H_2O); Soot (pine); Cellulose or pulp cellulose; Cork; Corn
Chemical Dusts	Adipic acid; Anthraquinone; Ascorbic acid; Calcium (acetate or stearate); Carboxy-methylcellulose; Dextrin; Lactose; Lead stearate; Methyl-cellulose; Paraformaldehyde; Sodium (ascorbate or stearate); Sulphur
Metal Dusts	Aluminium; Bronze; Iron carbonyl; Magnesium; Zinc
Plastic Dusts	(poly) Acrylamide; (poly) Acrylonitrile; (poly) Ethylene (low-pressure process); Resin (epoxy or melamine); Moulded Melamine (e.g. phenol-cellulose, wood flour & mineral filled phenol formaldehyde); (poly) Methyl acrylate; Resin (phenolic or terpene-phenol); (poly) Propylene; Urea-formaldehyde/moulded cellulose; (poly) Vinyl (acetate/ethylene copolymer, alcohol, butyral, chloride/ethylene/acetylene/vinyl, acetylene suspension/emulsion/copolymer

4.1.3. Primary and Secondary Dust Explosions

When dust ignites, any explosion resulting from the ignition can be primary or secondary. A 'primary dust explosion' is the initial explosion which occurs when the ignited suspended dust particles within a confinement (e.g. container or piece of equipment) explodes. One such example

is the 2003 aluminium dust collector explosion at Hayes Lemmerz International facility in Indiana (USA) which resulted in one fatality and injured several others (CSB 2006). A 'secondary dust explosion' is an explosion occurs when the primary explosion cause dust on a surface to aloft and ignite and can sometimes be more destructive depending on the extent and type of dust deposit. The blast wave from the secondary explosion can also cause accumulated dust on other surfaces to generate additional explosions. The CSB suggest that some initiating events for secondary explosions are not themselves dust explosions. One such event is the massive dust explosion at the West Pharmaceutical Services facility in Kinston, North Carolina (USA) which resulted in six fatalities and destroyed the facility (CSB 2006). Figure 4.1b is a schematic diagram of events which involves a primary and a secondary dust explosion.

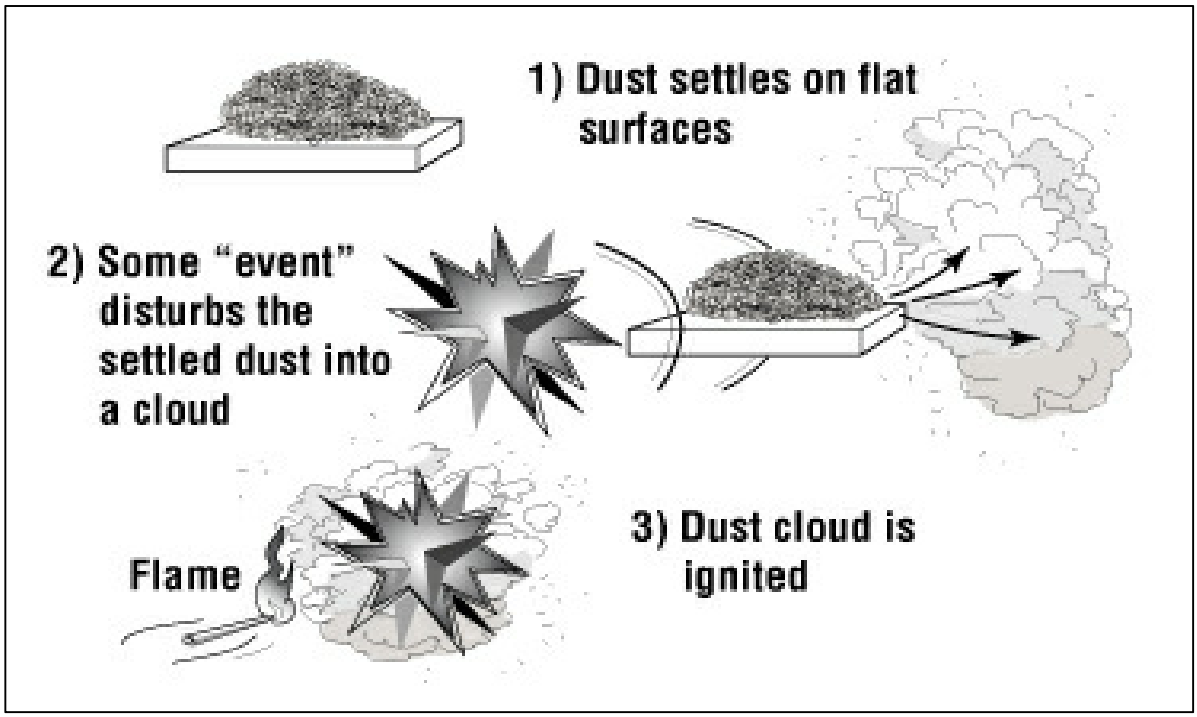


Figure 4.1b: Events in a primary and secondary dust explosion. Source: Anon (2018)

4.1.4. Factors which Determines Dust Explosivity

Dust explosions are caused by several factors including density and particle size which determines the ease of ignition and explosive severity a combustible dust. Dust density refers to the concentration of dust per cubic meter that could support explosive combustion. In terms of particle size, the finer the dust particles the more stable the dust cloud produces and thereby sustaining combustion and propagation. Per the NFPA specification, any dusts with particles less than 420 microns are considered explosive unless testing proves otherwise (Brazier 2017). Other measurable properties affecting as dust explosivity are detailed in Table 4.1c.

Table 4.1c: Measurable properties of dust. Adapted from Anon (2006) and Brazier (2017)

Property	Definition	Application
K_{st}	Dust deflagration index- a measure of explosibility of a dust cloud in units of bar.ms^{-1}	Measures the relative explosion severity compared to other dust. Determined with a 10 kJ ignition energy.
P_{max}	Maximum explosion overpressure generated in the test chamber- a measure of the power of combustion.	Used to design enclosures and predict the severity of the consequence. Determined with a 10 kJ ignition energy.
$(dp/dt)max$	Maximum rate of pressure rise.	Predicts the violence of an explosion. Used to calculate K_{st}
MIE	Minimum Ignition energy - the minimum energy which can ignite a combustible dust.	Predicts the ease and likelihood of ignition of a dispersed dust cloud. Maximum energy a typical MIE test apparatus is capable of discharging is 1000 mJ.
MIT	Minimum ignition temperature - the lowest temperature of a hot surface that will cause a dust cloud (not a dust layer) to ignite and propagate flame.	Help prevent a dust explosion occurring as a result of contact with a hot surface.
MEC	Minimum explosible concentration- the minimum concentration of suspended dust in the air that will explode	Measures the minimum amount of dust, dispersed in air, required to spread an explosion. Analogous to the lower flammability limit (LFL) for gas/air mixtures.
LOC	Limiting oxygen concentration	Determines the least amount of oxygen required for explosion propagation through the dust cloud.
ECT	Electrostatic charging tendency	Predicts the likelihood of the material to develop and discharge sufficient static electricity to ignite a dispersed dust cloud.
T_i	Ignition temperature - the lowest temperature of a surface that that is able to cause a dust/ air mix to ignite.	Depend on the shape of the vessel.
T_s	Smouldering temperature – the lowest temperature of a surface on which a 5 mm dust deposit may smoulder.	Describes the characteristics of thin dust layers. Thicker layers may cause an increase in thermal insulation which could change the smouldering temperature. Where thermal insulation lowers the smouldering temperature could trigger an exothermal reaction.
MC	Moisture content which indicates the amount of moisture in the dust.	An important factor for potential ignitions and explosions.
$MMPS$	Mass median particle size -size at which 50% of the particles by mass are larger and 50% are smaller.	An important factor for dust deflagration.

4.1.5. Identifying and Preventing Combustible Dust Hazard

Apart from particle size and density, other variables to consider for combustible dust hazard identification includes characteristics of ventilation and dispersion systems, the mode of ignition, potential ignition sources, air current, and confinement of the dust cloud, to mention a few. It is therefore imperative to have a system that avoids accumulation of dust on surfaces. This includes:

- Good housekeeping practice,
- Having appropriate safety system on process equipment,
- Education or training about combustible dust and its properties,
- Literature on the dust materials and final products use in the process e.g. safety data sheets (SDS).

4.2. Real-life Case History 1: New England Wood Pellet Dust Collector Fire and Explosion

On the 20th October 2011, a massive fire and explosion occurred at New England Wood Pellet LLC, Jaffrey New Hampshire plant in the United States which was captured in the Baghouse Editorial report of May 2012 (Baghouse 2012). According to the report, the incident was not the first combustible dust related incident at the plant. Before this incident the plant has already experienced a much severe incident in 2008.

The fire is suspected to have begun as a smoulder which was started by a spark or an ember from a pellet hammer mill in the wood pellet cooler. The fire then spread throughout the plant causing the dust collector to explode. The explosion was suspected to have vented through the explosion vents into adjacent storage silos then setting them ablaze thereby causing further spread of the fire throughout the plant. Figure 4.2 represent the block flow sketch of the rotary dryer building.

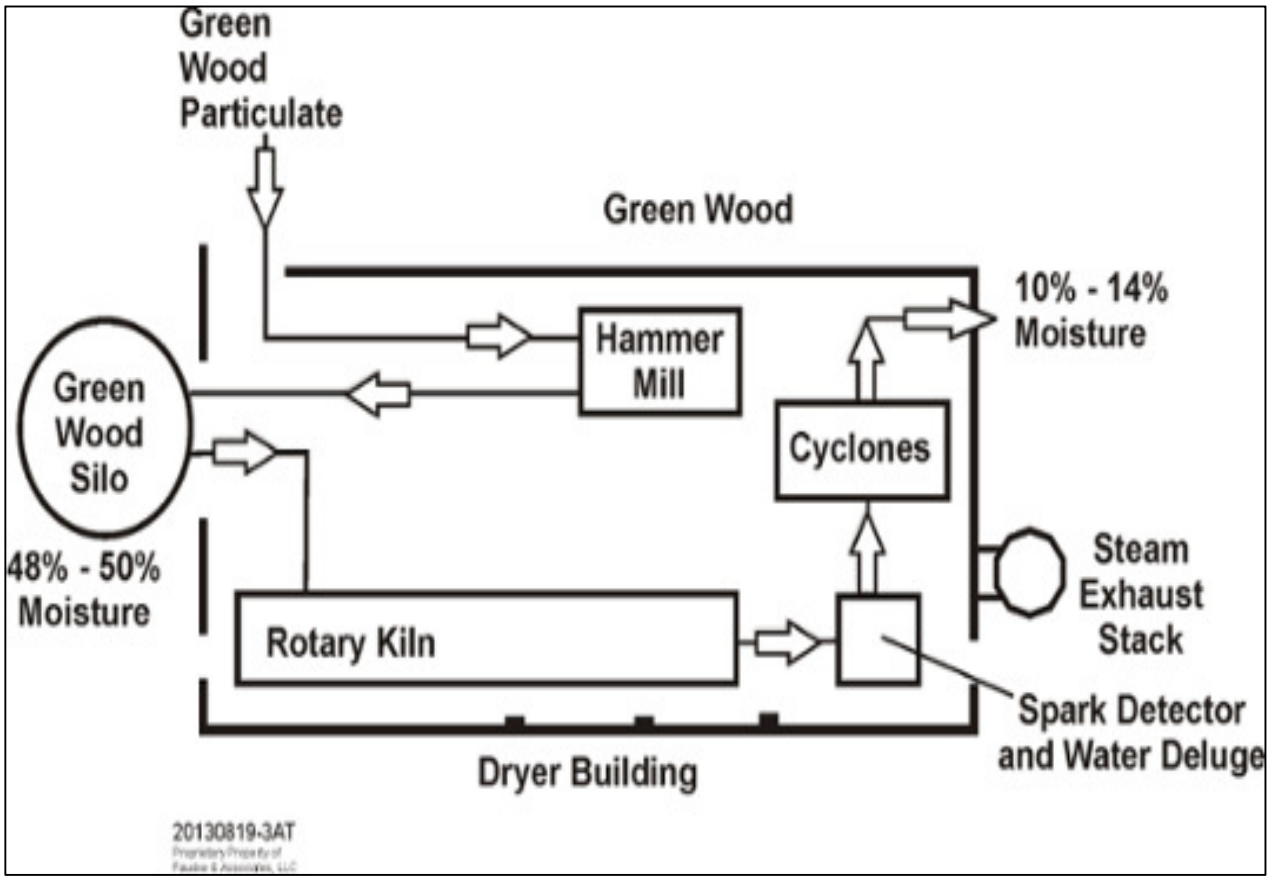


Figure 4.2: Sketch of the rotary dryer building. Source: Cullina, Dastidar & Theis (2017)

After investigating the incident, OSHA issue investigation citation and notification of penalty report which provides a detail description (OSHA 2012) of causes of the incident including

- Poor housekeeping
- Poor designed and installation in part of the process
- Poor operation at part of the dust collection system, which could have either initiated and or intensified fire and explosion, and
- Violation of the General Duty Clause of the OSHA Charter.

Under poor housekeeping, the report cites that excess amount of combustible dust was allowed to accumulate over surfaces including potential hot surfaces like conveyor belts which transports combustible dust to a hammer mill, overhead and wall surfaces in production room where the previous ignition occurred. The dust accumulations are in quantities that expose workers to fire and/or explosion hazards which could have a detrimental effect on the worker's life if any incident of fire and/or explosion occurs.

For poor design and installations, the report finds no spark detection, fire suppression, or explosion isolation systems on ductworks between most of the machines operated within the plant. One example reported was that the hammer mills and most of the dust collectors had no spark, fire, or explosion detection system. Even where the device has the required detection system installed

(e.g. duct between the pellet cooler and baghouse), the device didn't function properly. These are violations to the National Fire Protection Association (NFPA) Regulation 66 (NFPA 2017) which requires that

- For processes with particulate conveying and dust collection systems with fire and deflagration hazard, the hazard should be isolated to prevent propagation in either direction of the conveyor system.
- For the prevention of fire extension, the particulate processing system must premeditate to prevent fire or deflagration from spreading from one process system to the other.
- The ducts which carry wood dust from pellet cooler to the pellet collector were not designed to withstand the maximum explosive pressure expected for their intended load which contributed to the fire by-passing the isolation system. However, the NFPA 664 regulation requires the ducts to have sufficient strength with appropriate protection devices which can handle the extreme due to explosion pressure.

Under poor operation of the dust collecting system, the report establishes that the newly installed explosion vents on the baghouse were not fit for purpose (i.e. lack explosion suppression system per NFPA 68 so cannot withstand explosion hazards per NFPA 69) which again violates NFPA 664. For this, the regulation requires that dust collectors with fire or deflagration hazards must be directed outdoors unless they are

- equipped with deflagration suppression system
- equipped with deflagration relief vents with relief pipes extending to safe areas outside the building, and
- are of sufficient strength to withstand the maximum expected explosions pressure.

Part of NFPA 69 requires that piping, ducts, and enclosures protected by an isolation system must be designed to resist pressures estimated by the system manufacturer and must be verified by the appropriate testing procedure under set deflagration conditions to demonstrate system verification performance.

For violation of the general duty clause of the OSHA Charter, it was expected that the employer must provide to each of his employee's hazard free employment and employment environment and shall comply with the OSHA standards promulgated under the act. The explosion protection vents are therefore supposed to prevent combustible dust fire and/or explosion and thereby protecting employees and properties. Unfortunately, the vents fail to allow pressure from the explosion to vent out and away from combustible dust and the employees on the site.

This real-time case history has been used as a case study by some researchers who applied a risk-based approach to address the hazards of the combustible dust at the facility (Cullina, Dastidar & Theis 2017). They perform

- Investigations of the plans for sampling to obtaining appropriate test data
- Gap analysis, and
- Process hazard analysis.

As a result, this study will not critique the incident investigation report but present some of the findings made by Cullina and his team as in the section below.

4.2.1. Findings of the Risk-based Approach to Address Hazards of Combustible Dust by Cullina and his Team

4.2.1a. Hazard Analysis

The raw materials for the manufacturing of the pellets at the facility are green wood particulates (GWP) and kiln-dried wood particulates (KDWP). GWP are primarily large wood chips with some amount of smaller chips and sawdust. The GWPs are reduced and dried as part of the pellet manufacturing process. The GWPs are store out in the open. The KDWPs on the other hand are much smaller and drier than the GWPs. The KDWPs are fed onto a conveyor to the KDWP receiving area where they are mixed with various sources of green material a covered area, then loaded onto feed conveyors which transport the materials to a Dryer Building. The GWPs and KDWPs are used as fuel for the rotary kiln. Tables 4.2a. and 4.2b. are lists of measurable properties of GWPs and KDWPs at the various stages of the pellet manufacturing process and combustible dust test results at the aforementioned stages.

Table 4.2a: GW and combustible dust test results. Source: Cullina, Dastidar & Theis (2017)

GW	$MMPS$ **(µm)**	MC **(wt.%)**	P_{max} **(bar)**	K_{st} **(bar-m/s)**	MIE **(mJ)**	MIT **(℃)**	**Particle Size (µm)**
Storage Pile	>500	> 50%	-	-	-	-	-
After 1st Size Reduction	>500	> 48%	-	-	-	-	-
After Drying	>500	10.4 - 14	-	-	-	-	89% > 425
After 2nd Size Reduction	>500	9.4 - 13.4	-	-	-	-	77% > 425
After 3rd Size Reduction	>500	1.3 - 3.1	-	-	-	-	70% > 425
After Fuel Size Reduction Sample point 5, As Received, 5kJ Igniter	> 500	-	No ignition	No ignition	-	-	-
After Fuel Size Reduction Sample point 5, As Received, 10kJ Igniter	> 500	-	6.6 ± 10%	31 ± 30%	> 1000	-	-
After Fuel Size Reduction Sample point 5, After ASTM E1226 Processing	27 95% < 75	-	7.9 ± 10%	159 ± 12%	10 < MIE < 30 *Es = 19	440	-

**Es = Establish self-ignition*

Table 4.2b: KD and combustible dust test results. Source: Cullina, Dastidar & Theis (2017)

KD	$MMPS$ (μm)	MC (wt.%)	P_{max} (bar)	K_{st} (bar-m/s)	MIE (mJ)	MIT (℃)
Post KD Hammer Mill Sample point 6	27 95% < 75	2.6	7.9 ± 10%	152 ± 12%	10 < MIE < 30 Es = 12	440
Pellet Mill Feed (Mixed KD and Green particulate, Sample point 7)	24 97% < 75	2.6	7.9 ± 10%	174 ± 12%	10 < MIE < 30 Es = 16	440
Pellet Fines	26 97% < 75	2.5	7.9 ± 10%	182 ± 12%	10 < MIE < 30 Es = 19	440

From the parameters in the tables, GWPs on its own poses a lower risk until mix with the KDWPs because the size of KDWPs show that they can easily form combustible dust cloud which can be ignited particularly by electrostatic discharge. As a result, the baghouse serving the enclosed conveyor, hopper, hammer mill and the silos and the employee unloading the KDWPs are all at the risk of exposure to deflagration from the process system.

4.2.1b. Gap Analysis

The gap analysis performed to identify and document discrepancies between standards and existing implementation of equipment, practices and management systems under the NFPA guidance identified several concerns with (a) protection for equipment handling combustible dust; (b) electrical area classification; and (3) non-existing and outdated safety management systems.

4.2.1c. Process Hazard Analysis

The outcome of the process hazard analysis (PHA) did not correlate with the laboratory test data presented on the rotary kiln because the PHA applied data was collected using automatic data collection system taken twice per shift of the employees.

4.2.1d. The Conclusion

By using a well-defined testing strategy, gap analysis, and PHA the researchers provided some insight into hazards and the risk associated with dust hazards in the HHPIs which are not previously known. The study therefore proceeds to Real-time Case History 2.

4.3. Case History 2: Imperial Sugar Manufacturing Facility Dust Fire and Explosion

On the 7th February 2008, a series of sugar dust fire and explosions occurred at Imperial Sugar Manufacturing facility in Port Wentworth, Georgia (USA). According to reports (Vorderbrueggen 2010), the explosion resulted in 14 fatalities and other injuries, destroy buildings and equipment. The facility refines raw sugar into granulated sugar and produce powdered sugar and other sugar

products which are packaged in various capacities from small boxes to bulk tank and hopper railcar. The granulated sugars are transported to three 105-foot tall sugar silos via a system of screw and conveyor belts and bucket elevators. The refined sugar is then transferred from the silos to a loading area in packaging buildings and then to the powdered sugar production equipment.

The transporting system consists of dozens of screw conveyors, bucket elevators, and horizontal conveyor belts. The bucket elevators and steal conveyors are enclosed to prevent releasing sugar/sugar dust into the work areas. As a result, sugar dust accumulates on the overhead conduit, piping, ceiling beams, lights, and equipment. However, the enclosure was not adequately sealed. The system is also not equipped with dust removal equipment. Figure 4.3. shows a schematic diagram of the granulated sugar supply and discharge system. The screw conveyor consists of a rotating helical screw inside a closed trough. This transport granulated and powdered sugar to a discharge chute with the aid of rotating helical screw. The bucket elevator scoops granulated sugar and transports it to the discharge chute at the top of the housing unit.

The investigation of this incident was carried out by a team from the CSB joined by investigation teams from the Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF), the Georgia State Fire Marshal, Occupational Safety and Health Administration (OSHA). The findings of the investigation are detailed in the CSB Report Number 2008-05-I-GA (CSB 2009). The study critiques the final incident investigation report in the sections below.

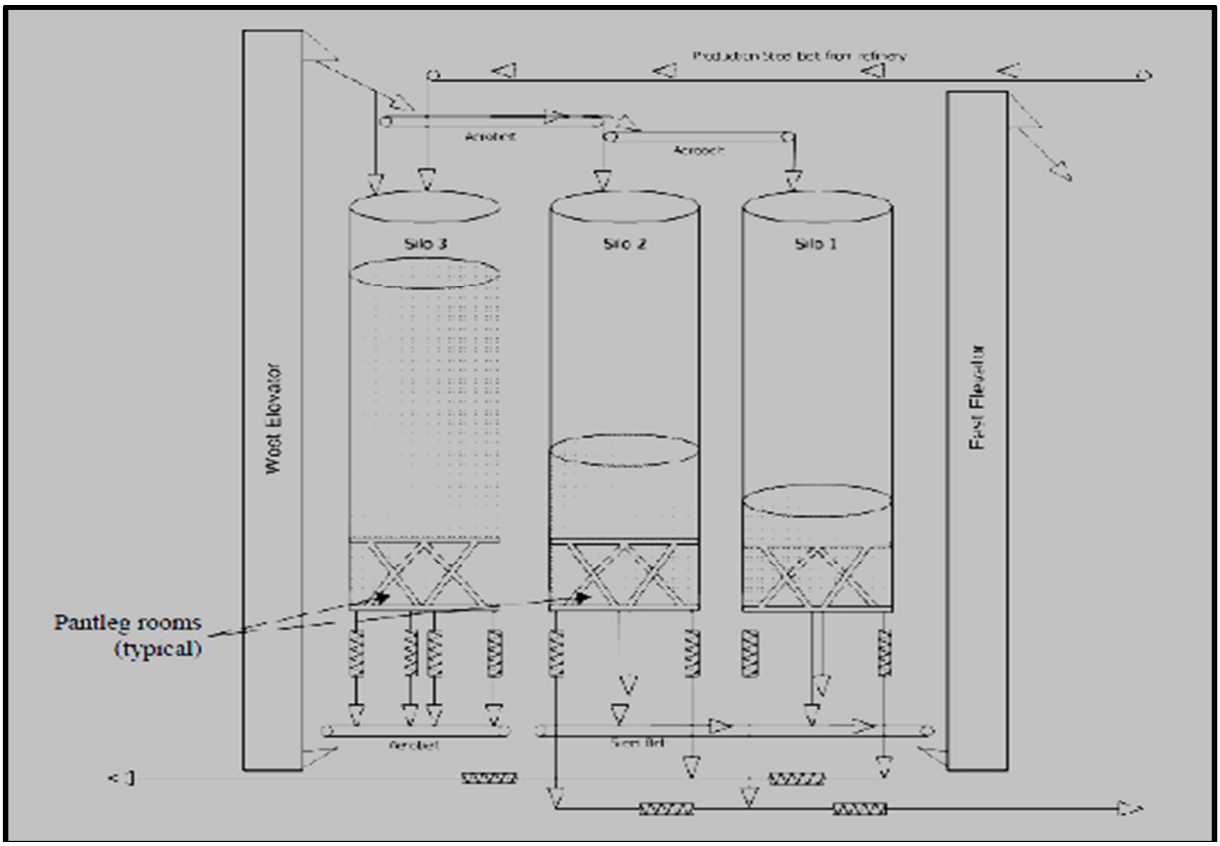


Figure 4.3: Schematic diagram of sugar supply and discharge system (CSB 2009)

4.3.1. Critiquing the Imperial Sugar Manufacturing Facility Dust Fire and Explosion Report

The final incident investigation report CSB Report Number 2008-05-I-GA of the sugar dust and fire explosion published in 2009. This report evaluation was performed from a big data perspective with the goal to provide a systematic and analytical discussion on the extent by which this investigation shows some characteristics of qualitative and/or quantitative research. The full report was obtained from the CSB website and can be access via the following link http://www.csb.gov/assets/1/19/Imperial_Sugar_Report_Final_updated.pdf.

4.3.1a: Accident Description

The accident occurred during a tour by the new CEO with three employees. The accident was suspected to have started from the packing building, destroyed the sugar packing buildings, palletizer room, and silos, and severely damaged the bulk train car loading area and parts of the sugar refining process areas. The silo fires continued to smoulder for 7 days before they were finally extinguished. The CSB identified overheated bearing in the steel belt as the most likely cause of the initial explosion.

They conclude that the first dust explosion was initiated in the enclosed steel belt conveyor located below the sugar silos 1 and 2, which then triggered the secondary dust explosions and fires. Reports on the incident (CSB 2009) reveals that the company was fined $6 million because of infringements and had to rebuild the damaged portions of the Port Wentworth refinery at an estimated cost of $220 million (NASA 2011).

4.3.1b. Evaluation of the Report

According to the report, qualitative and quantitative data were collected from interviews and other incident investigation activities. However, the following issues about the report were observed:

a. The aim of the investigation was not specified hence the study couldn't establish whether the design of the investigation is consistent with the aims.
b. Sample questionnaire and sample data were not provided.
c. The methods used for the investigation were hardly described hence it wouldn't be possible for another group to repeat the investigation using the same methods. It would have been helpful if the leadership of the investigation had ensured that a link to the data, or sample of the data extracted by both CSB and OSHA.
d. The report did not explain the task performed by OSHA, although explanation was provided that the investigation was performed by more than one organisation.

e. There report did not mention how the data collection method used was calibrated. Since the data collection was performed by more than one person, the calibration method should have been provided to provide clarity on the integrity of the data.
f. It was not possible to establish whether the questionnaire was piloted before used.
g. Details about whether the investigators have been trained to critically evaluate documents (e.g. engineering documents, equipment operation, and maintenance records) was not provided. It was expected that such details should have been provided to make the findings more credible.
h. There was no mention of the nature of the interview and interview questionnaires or pro forma used and whether they have been tested or are relevant to the investigation.
i. The data collection sheets used when the records were assessed were not included even as a figure. This study expects the report to provide sample questionnaires as figures or link to a website.
j. There was no mention on whether the records were complete enough or whether there was a presence or absence of features such as missing data, the competence level of the employees, or any unreported near-misses that should have been explained.

4.3.1c. Incident Analysis

a. This study expects the incident investigation report to capture previous dust fire and explosion incident investigation despite evidence of previous cases of similar incidents reported on the CSB website. Some information about the previous investigation of a similar incident should have been capturing to help test the methodology used.
b. It was impossible to established from the incident investigation report that the sample used was a representation of the population. For instance, the method for collection of the dust samples and analytical techniques used were not covered. The inclusion of such information would have helped in replication the investigation by another group.
c. There are insufficient details about any on sample collection techniques, no data exploration or statistical method applied, and no detail description of where and how the sample was tested.
d. Under the subtitle "Hot Surface Ignition," it was reported that hot surface ignition caused by bearing overheating was suspected to be the potential ignition source. Unfortunately, the report did not provide any information on any investigation of bearing overheating or what may have caused the bearing overheating.
e. The report also referenced Eckhoft (2003) who suggested that airborne dust, combustion gases from smouldering and decrease ignition temperature below the temperature of the pure dust could have caused the ignition. Although this information was expected provide clarity on the difference between ignition, smouldering, and their corresponding effect in the

combustion of combustible powders, this may confuse non-explosive safety experts. This should have been reworded to provide clarity for readers from diverse backgrounds and those institutions for whom the recommendations in the report were made. Additionally, the reference source is a book, so the report should have highlighted the page number to help the readers who may be interested in finding the information.

4.3.1d. Results (Key Findings and Incident Causes)

a. Sections 8.0 (Key Findings) and 9.0 (Incident Causes) should have been combined under the Section “Results.”
b. There report did not cover any statistical methods or testing applied which may be because no statistical analysis was performed as part of the investigation. Considering the nature of the investigation it was expected that some big data or statistical analysis technique could have been applied. Despite these issues, and the small sample population, the findings are justified and deemed to be acceptable for the process incident investigations.
c. Tables were not provided in the section of report even though some tables were provided under the section for method.
d. Despite the issues highlight under “b and c”, the results were presented in a clear and unambiguous manner though without tables, figures, or graphs, but as numbered points.

4.3.1e. Report’s Discussion

a. Considering the nature of the investigation, this study expects a Section for Discussions which "critique and discuss" the method used but this was not the case. This may be because the report did not explain exactly the nature of the method used under the Section for Method. A Section for Discussion of the report which explains the strength and weakness of the method used should have been presented.
b. Despite CSB's reputation as an independent national investigation body with years of investigation experience, this study expects that at least the method used in the investigation would have been explained. For instance, a weakness which may be encountered during the data collection through examination of employee training records, equipment operations, and maintenance records. There could be incomplete data such as unreported near-misses which some employee may have forgotten to report, or record of near-miss incidents reported by a different employee which may vary in the description of an incident.
c. The section for Discussion was also expected to comprehensively discuss the outcome of the investigation in relation to other incidents.

d. Although the outcomes of the investigation were clearly listed, it was difficult to establish whether it extends beyond the method used.

4.3.1f. Report's Recommendations

This section - Recommendations in the report failed to highlight what cause the bearing overheating, although bearing overheating was suspected to be the initial ignition source. This forms the bases for the use of bearing data in our research.

4.4. The justification for using the Real-time Case Histories for the Thesis

After critiquing the case study report in section 4.2 and 4.3, this study found a justification for using the two incidents as a real-life case study for a doctoral research as explain below.

The two facilities represent a typical HHPI. The QRA in the facilities involves the following phases:

- Identification phase which aims to identify the accident scenario, or the hazard associated with the equipment's, products, and other activities within the facility. This involves data and/or information from the process that could help to develop the necessary method. Characterisation phase which aims at evaluating the nature of adverse effects associated with the hazard and the quantitative assessment of the relationship and/or association between the magnitude of the exposure and the likelihood of adverse effects.
- Assessment phase which involves estimation of the likelihood and magnitude of exposures associated with the hazard.
- Risk characterisation phase which involves estimation of the risk for the set of input data.

Since the accident in the facilities has already occurred, the identification of the accident scenario was achieved by reviewing the accident investigation report. For instance, the incident investigation report on the Imperial Sugar Factory incident the initial source of ignition was overheated bearing in the steel belt. Four possible issues may have caused the bearing overheating.

- Incorrect lubrication problems.
- Mechanical issues like problems within the bearing mechanical components.
- Bearing housing-related issues.
- End of useful life of the bearing.

The characteristics of the investigation suggests that the method applied were from engineering and chemical engineering discipline. Nevertheless, there is evidence from elements of the investigation to suggest that the investigation could be performed as a data-driven incident investigation. This is because all the stages of investigation exhibit characteristics of a data-driven

investigation. These characteristics include collection and storing of appropriate data, data analysis, using the data to make inform findings and communicating the findings to the appropriate authorities.

It is evident from the reports that there were technical challenges and data quality issues which may have the potential of influencing the results. Applying big data techniques as part of the method for the investigation could have help with the successful implementation and execution of the investigation. For instance, big data techniques could have been applied to investigate the risk of bearing failure and overheating which was identified as the initial source of the ignition that led to the explosion.

The case histories and their investigation highlight the number of repeated occurrences of the incidents and the periods between the occurrences which indicates that the frequencies of the risks are already known. The extent of loss and probability of disruption event are all covered by the incident investigation report. Thus, a method for detection of the risk of failure of the bearing and overheating is one element which is required. This would require the use of bearing operation monitoring data, some of which are available for this research.

Thus, there is justification for using the two real-life incidents as case histories because they carefully defined a real-life accident which has already been investigated and with the findings available in the public domain. They, therefore, provide an opportunity for a practical analysis to help provide a QRA method which relies entirely on big data techniques and real-time data application in the HHPIs.

4.5. Conclusion

The research has presented two case histories relating to incidents within HHPI's as part of this study. A description of the incidents together with the reasons for their selection has been provided. The incident investigation reports were reviewed and critiqued after which the justification of their selection as case histories for the research was established. These justifications include

- number the number of repeated occurrences of the incidents and the periods between the occurrences which indicates that the frequencies of the risks are already known.
- the extent of loss and probability of a disruption event are already covered by the incident investigation report.
- providing a method for detection of the risk of failure of a bearing component (e.g. bearing overheating).
- the bearing operations dataset which is already available can be used for this research.

Next is Chapter 5 – Data, where the study introduces the available datasets for the research. The study provides the source and description the datasets as well as their storage and handling.

Chapter 5 - Data

5.0. Introduction

In Chapter 4 – Real-life Case Histories, two real-life case histories relating to fire and explosion incidents within HHPI's were presented. The study provides these case histories to

- Provide understanding of incidents associated with risk within the HHPIs.
- Provide an opportunity to evaluate possible risk sources within a typical process system.
- Introduce the available datasets for the research.
- Provide a justification for the use of available datasets for the research.

In this Chapter, the study introduces the datasets for the research and provide a link to the source where the data can be obtained. Attributes of the data and data handling and storage process shall be covered. Because the datasets are open source data, this study will provide

- A list of the publications of previous research work for which the data has been applied.
- The purpose for the research work done in the publications.
- Any big data techniques use by the method used in the research in the publications.
- The contributions their research in the publications made to knowledge.

Figure 5 is a flowchart illustrating the process by which the datasets were sourced to the justification for using the datasets for the study. The study will also highlight some of the issues experienced during the search for the datasets.

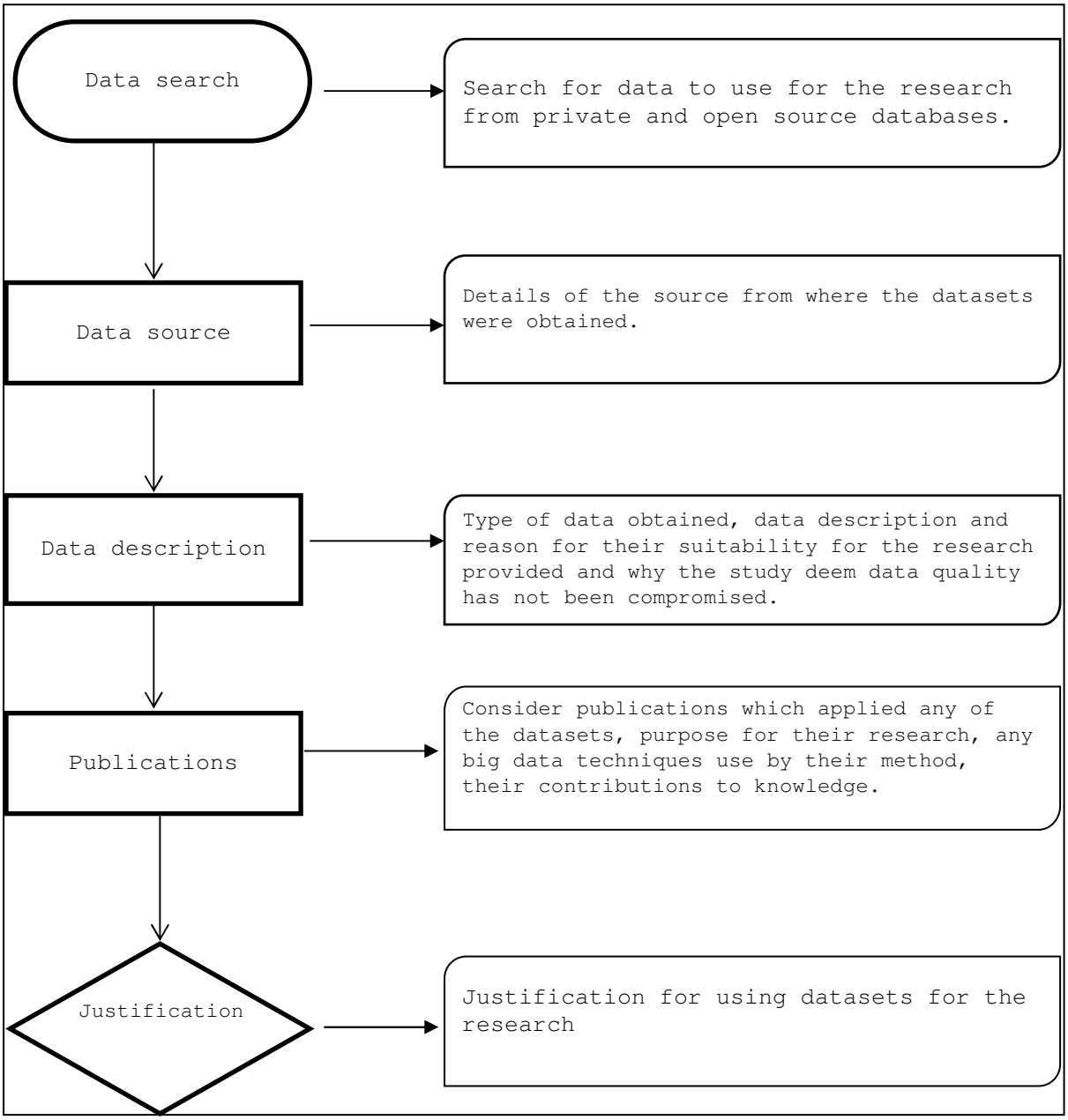


Figure 5: Flowchart for the process of sourcing dataset and justification for use.

5.1. Searching for Data

The search for data for the research began with a consultation with some representatives of the company for which I am currently employed. At the planning stage of my doctoral research, these safety experts were interested in the propose research topic. With the company being a process safety company, they agree to provide a collaborative support by providing data from the company's Industrial Explosion Hazard and Chemical Process Evaluation databases, since the research could be beneficial to the company.

Unfortunately, the interest of these individuals weaned due to several reasons including

- An opinion that data can be extracted from statistical and graphical presentation in reports.

- Privacy and data compliance regulations which became more prevalent so the company and client's data must be kept private, safe and secure in compliance with the General Data Protection Regulation (GDPR) 2016/679 (EU 2016).

As a result, the help promise by these individuals did not materialise. For instance, there was a general assumption by some of the individuals who promise to help secure data for this research that data and statistics are similar so this study could use extracts from some of the statistical information already published reports. Management of the data protection department of my organisation were also concern about releasing confidential clients' data for this research because it violates GDPR 2016/679, although GDPR 2016/679 focuses on "the protection of natural persons with regard to the processing of personal data and on the free movement of such data" (EU, 2016 p.1 & 33 - 35). Hence to provide some clarity on data and statistics, the study presents a distinction between data and statistics as used in the context of this research because the two terms are sometime used interchangeably.

5.2. Data and Statistics

Data is the raw information consisting of both signal and noise which can be published or stored as datafiles of various format. It can be analysed using several different procedure and statistical software's. Statistics on the other hand is a well-established methodology of science and mathematics which can be applied for analysis and interpretation of the data.

5.3. Data Sources

In light of the difficulties at obtaining data from my organisation database, some efforts were made to obtain the data by contacting some researchers whose publication were found to be related to the domain area of this study, but this was also not successful. The study therefore opted for data from open sources data repositories. A comprehensive search for data from open source data repositories includes the UCI Machine Learning Repository, Google, Dataverse and Kaggle was conducted. Finally bearing operation datasets were obtained from the NASA Prognosis (https://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/) and PMM Challenge data repositories (http://data-acoustics.com/measurements/bearing-faults/bearing-6/). A total of five bearing operation datasets were obtained from the two databases for the research.

Three of the five datasets were obtained from the NASA prognosis data repository. A publication by the donors of the data (Lee et. al. 2007) reveals that the process system has four bearings with a rotation speed was kept constant at 2000 RPM by an AC motor coupled to the shaft via rub belts as detailed as Figure 5.3a. The remaining two datasets were obtained from the PHM challenge data repositories. The data and detailed description were downloaded from the PHM Challenge

website via the link. The vibration data was obtained from two accelerometers placed in vertical and horizontal directions with a load applied in the horizontal direction. Figure 5.3b is a schematic representation of the process system.

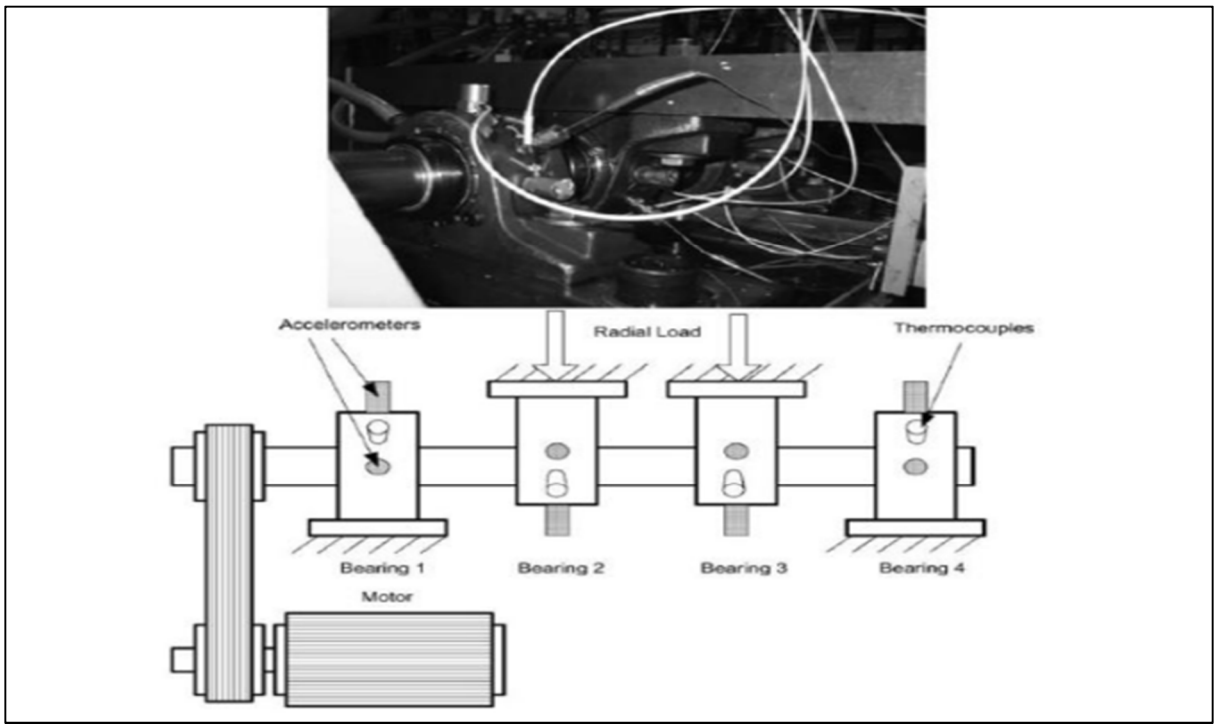


Figure 5.3a: Process system with schematic arrangements of bearings by Lee et. al. 2007

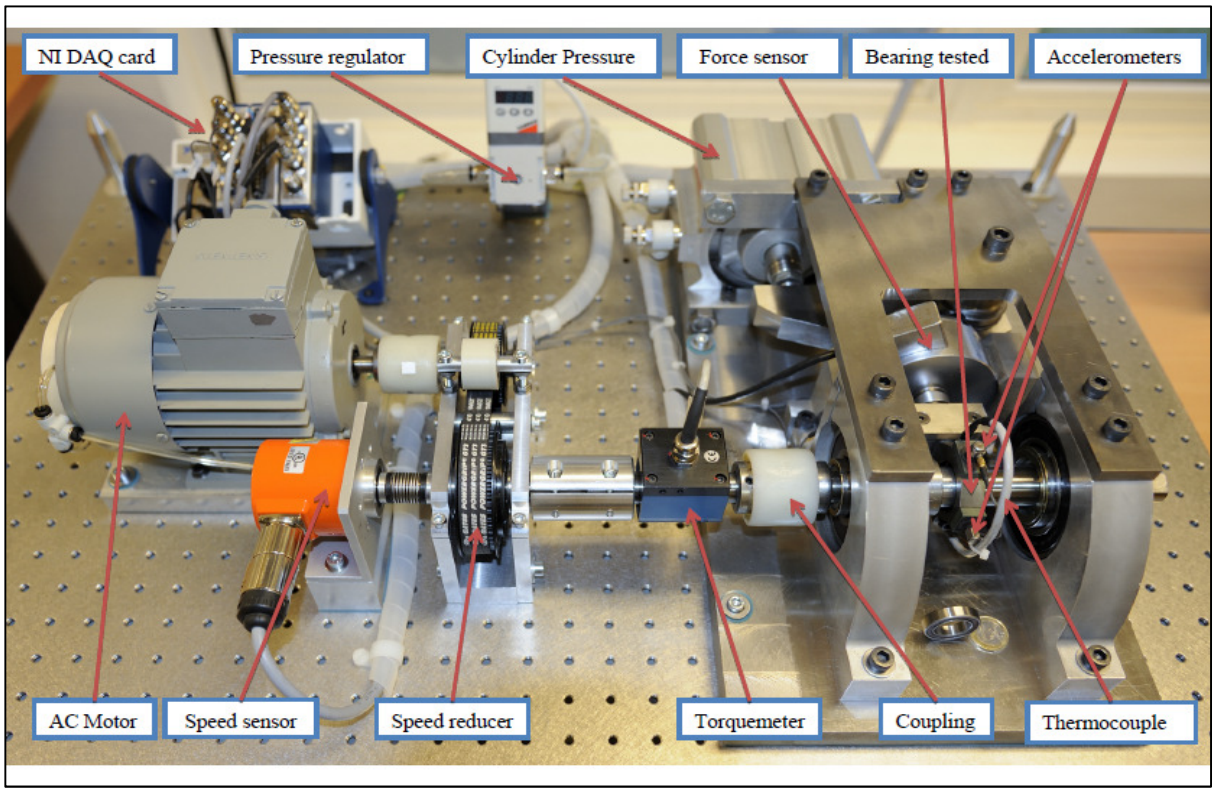


Figure 5.2b: Overview of IEEE PHM process system (Source: IEEE PHM 2012 Challenge)

5.4. *Description of Datasets*

The datasets were stored on the two sources as compressed text files to enhance its management and thereby maintaining its quality and easy storage. The compressed zipped files were downloaded, unzipped using the open source software package 7-Zip v.16.04 and stored on a local hard drive. The data files in the datasets obtained from the NASA prognosis database was found to contain only vibration data even though temperature of the bearings and flow of the lubricant were monitored as part of the operation process (Qui et al. 2011). Also, one of the datasets obtained from the PHM diagnostics database has both vibration and temperature data which were obtained under variable conditions (Nectoux et al. 2012).

Key information about attributes of the datasets will be discussed in Chapter 6. Because the data is open source the study performed a literature search for peer reviewed research publication for which any of the datasets has been used. The study then proceeds with a review of publications to understand which big data techniques were used as part of the methodology and the corresponding major contributions to knowledge.

5.5. Citations *which Applied the Available Datasets*

Several bearing health monitoring techniques have been investigated and applied by researchers (de Azevedo, Araújo & Bouchonneau 2016). They include Acoustic Measurement (AM); Electrical effects monitoring (EEM); Oil Debris Monitoring (ODM); Power quality (PQ); Temperature Monitoring (TM); and Vibration Analysis (VA). For clarity, the study provides a summary of the bearing health monitoring techniques in Table 5.5a.

Table 5.5a: Bearings health monitoring methods (adapted from de Azevedo et al 2016)

Method	**Summary**
AM	Using higher-frequency components from noise readings from sound data to predict possible bearing health issues.
EEM	Using changes in readings from electrical resistance data (e.g. discharge measurements) to predict possible bearing conditions.
ODM	Using the presence of metallic debris in lubricants to detect possible bearing health conditions.
PQ	Using power quality data (e.g. electric or mechanical power measurements) to detect any changes in bearing health.
TM	Using any changes in temperature data to detect bearing health conditions.
VA	Using changes in vibration to obtain information on bearing health.

Some scientists who have use some of the available datasets have suggested that the bearings failed after exceeding 100 million revolutions because of a crack on the outer race (Qiu et al. 2006). Others have also made claims that some parameters applied in physics-based modelling of bearing health as part of a process component monitoring process get missed so bearing vibration datasets alone are not sufficient when compared with other datasets of similar sample sizes found in the literature (Eker, Camci & Jennions 2012).

Despite the differences in opinion, bearing monitoring datasets has been applied for research with some major contributions to knowledge. Owing to the large number of research publication which has applied the datasets available for this study, and the relevant variables collected, and contributions made to knowledge, the study concludes that quality of the datasets has not been compromised. As a result, the quality of the available data was good and fit to be applied for the study.

Thirty-three of the publications were previously reviewed by de Azevedo (de Azevedo, 2016). As a result, the study reviewed another twenty-two publications found to have use some of the datasets. Table 5.5b is a summary of the review of all 55 publications, the big data techniques used and their major contribution to knowledge. The review reveals that 26 big data techniques were applied in the 52 publications. The big data techniques applied includes:

- Alpha-stable graphical models (AGM),
- Energy entropy decomposition (EMD),
- Intrinsic mode functions (IMFs) and artificial neural network (ANN),
- Discrete wavelet transformation (DWT),
- Genetic Algorithm (GA),
- Data envelope analysis (DEA),
- Discrete fourier transform (DFT),
- Extreme learning machine neural network (ELMNN),
- Hidden Markov Models (MMs),
- Fuzzy logic (FL),
- Contribution analysis (CA),
- Kernel independent component analysis (KICA),
- Cluster analysis (CA),
- Kohonen neural network (KNN),
- Linear autoregressive (LA) and extended Kalman filter (EKF),
- Local and nonlocal preserving projection (LNPP) based feature extraction,
- Local mean decomposition (LMD),

- Logical analysis of data (LAD),
- Lyapunov exponent (LE),
- Machine learning and numerical analysis (MLNA),
- Mahalanobis–Taguchi–Gram–Schmidt (MTGS),
- Prognostic Integration Architecture (PIA),
- Proper orthogonal value (POV),
- Wavelet decomposition analysis (WDA),
- Support vector machine (SVM),
- Self-organizing map (SOM), and
- Maximum likelihood estimation (MLE).

Table 5.4b: Citations which have used bearing vibration dataset to make contribution to knowledge. Extended from de Azevedo (2016)

Author(s)	Major Contributions	Machine Learning Approach
Li (2016)	Assess bearing health condition using the Alpha-stable Probability Distribution Model	(AGM)
Kulkarni et al. (2016)	Use ANN based prognosis to classify the health state of bearings.	ANN
Zarei et al. (2014)	Use Time-domain vibration signal and ANN to detect and classify bearing faults.	ANN
Ziani et al. (2012)	Bearing fault detection by ANN and GA.	ANN and GA
Unala et al. (2014)	Fault estimation algorithm based on ANN, envelope analysis (EA), HT/FFT to detect bearing faults.	ANN, EA, HT/FFT
Li et al. (2017)	Rolling Element Bearing Performance Degradation Assessment Using Variational Mode Decomposition and Gath-Geva Clustering Time Series Segmentation	CA
Dalvand et al. (2014)	Detect bearing defects using Instantaneous Frequency method with DEA and condition indicators.	DAE
Guo et al. (2014)	Envelope extraction and independent component analysis for faults detection on rolling element bearing.	DAE
Ming et al. (2015)	Cancellation method through iterative calculation of the envelope of the multi-component signal to detect bearings faults.	DAE
Waters et al. (2013)	Real-time vibration-based method to detect, localize, and identify a faulty bearing.	DAE
Zhao et al. (2013)	Tacholess envelope order analysis technique to diagnose bearings faults.	DAE
Abdussiam et al. (2011)	Monitoring bearings with Time Encoded Signal Processing and Recognition (TESPAR), vibration and envelope analysis.	DAE
Miao et al. (2013)	Health Assessment of Cooling Fan Bearings Using Wavelet-Based Filtering	DFT
Miao et al. (2011)	Zoom interpolated discrete Fourier transform based on multiple modulations to identify faults.	DFT
Kumar et al. (2013)	DWT and ANN to detect bearings faults.	DWT and ANN
Ali et al. (2015)	Identify and classify bearings defects using EMD, EED, IMFs and ANN.	EED, IMFs and ANN
Duan et al. (2018)	Three steps cumulative transformation algorithm for data processing and fusion technique for bearing degradation assessment.	ELMNN
Wang et al. (2014)	Four steps signal processing method to directly extract signal components relating to the rotational speed of a rolling bearing.	EMD
Liu et al. (2013)	Frequency domain, percent power, peak RMS, sequential forward search algorithm and adaptive neuro-fuzzy inference systems to detect and identify faults on bearings.	FL
Zhang et al. (2015)	Develop effective degradation indicator of rolling bearings for residual life prediction using a combination of continuous hidden Markov model and various bearing features to construct an effective degradation indicator.	HMM
Yu (2012b)	Hidden Markov model (HMM) and contribution analysis to monitor bearing health degradation.	HMM and CA

Author(s)	Major Contributions	Machine Learning Approach
Tobon-Mejia (2011)	Use HMMfor Failure Diagnostic and Prognostic	HMMs
Dybala et al. (2014)	Diagnose bearings' faults by decomposing raw vibration signal into several IMF with the aid of EMD.	IMF and EMD
Ma et al. (2014)	Combine KICA and LS-SVM to achieve fault monitoring and classification of bearings.	KICA and LS-SVM
Qiu et al. (2006)	Robust degradation indicator based on a self-organizing map neural network to evaluate the bearing degradation performance.	KNN
Jin et al (2016)	Anomaly Detection and Fault Prognosis for Bearings	LA and EKF
Yacout (2012)	Analyse maintenance and performance of bearing.	LDA
Mohamad-Ali, et al. (2011)	Investigates advantages of using LAD for fault diagnosis of bearings.	LDA
Caesarendra et al. (2015)	Largest LE algorithm for faults detection and deterioration track of low-speed slew bearing.	LE
Liu et al. (2014)	Local mean decomposition technology and multi-scale entropy to diagnose roller bearing faults.	LMD and EA
Yu (2012a)	Use local and nonlocal preserving projection (LNPP)-based feature extraction algorithm, to discover the global structure of Euclidean space for assessment of bearing performance degradation.	LNPP
Lei et al. (2016)	A Model-Based Method for RUL Prediction of Machinery	MLE
Sassi et al. (2008)	Developed two new scalar indicators to evaluate the severity of bearing degradation.	MLNA
Tobon-Mejia (2010)	Use Mixture of Gaussians Hidden Markov Model for failure diagnostic and prognostic	MoG-HMMs
Shakya et al. (2014)	Use Mahalanobis Distance by application of Gram-Schmid to orthogonalization process and Chebyshev's inequality to analyse Vibration data in time-frequency and time-frequency domain.	MTGS
Shakya et al. (2015)	Integrated MTGS method to fuse damage identification parameters to detect and classify bearings' defects.	MTGS
Dempsey et al. (2011)	Determine if a diagnostic tool for detecting fatigue damage of helicopter tapered roller bearings can be used to determine RUL.	None
Bolander et al. (2009)	Determine the value of incorporating vibration derived condition indicator (CI) data for predicting RUL of gearbox bearings.	PIA
Ahn et al. (2014)	Use wavelet denoising scheme and POV of intrinsic mode function covariance matrix to detect the Fault of a roller bearing system.	POV, IMF and EMD
Wang et al. (2011)	Wavelet packet sample entropy method and EMD to forecast the operating state of rolling element bearing.	SE and EMD

Author(s)	Major Contributions	Machine Learning Approach
Qiu et al. (2003)	Robust performance degradation assessment methods for enhanced rolling element bearing prognostics.	SOM and ANN
Borghesani et al. (2013)	Squared envelope spectrum and the kurtosis of the corresponding band-pass filtered analytic signal for bearings defect diagnosis.	SVM
Saidi et al. (2014)	Identify different fault patterns of bearing using a combination of higher order spectra analysis features and support vector machine.	SVM
Liang et al. (2014)	Intelligent bearing fault detection method based on a calculus enhanced energy operator to extract bearing fault signatures.	SVM
Martinez-Rego et al. (2011)	Uses one-class-v-SVM paradigm to analyse bearing degradation data under noisy circumstances.	SVM
Porotsky (2012)	RUL estimation for systems parameters with the non-trend ability behaviour.	SVM
Zimroz et al. (2014)	Bearings diagnostic using peak-to-peak, root means square of vibration acceleration and generator power.	SVM
Nizwan et al. (2013)	Vibrational analysis and discrete wavelet transform for bearing fault detection using wavelet decomposition.	WDA
Jayaswal et al. (2011)	Detect early fault on bearings using vibration	WDA
Khanam et al. (2014)	Discrete wavelet transformation assisted by sym5 wavelet to detect and estimated effect size on bearings	WDA
Li et al. (2013)	Morlet wavelets filter and spectral kurtosis to detect bearings defects.	WDA
Roulias et al. (2013)	Wavelet denoising with neighblock threshold technique for condition monitoring of roller bearings.	WDA
Sarvajith et al. (2013)	Use Fourier and DWT to determine bearings condition.	WDA
Sun et al. (2013)	DWT and envelope analysis to diagnose rolling bearing faults.	WDA
Sun et al. (2013)	Multiwavelet denoising technique with a data-driven block threshold to detect rolling bearing defects.	WDA

The study then applied text analysis to extract the most featured text from the major contributions to knowledge to find most featured words in the contribution by the publications. Figure 5.5 is the word cloud obtained which shows words like faults, degradation, defects and diagnostics as the most featured words in the major contribution to knowledge. All these words refer to events within process systems and therefore require some clarity for their usage within the context of this study. The study therefore applies some explanation provided by Parhami (1997) to explain the featured words.

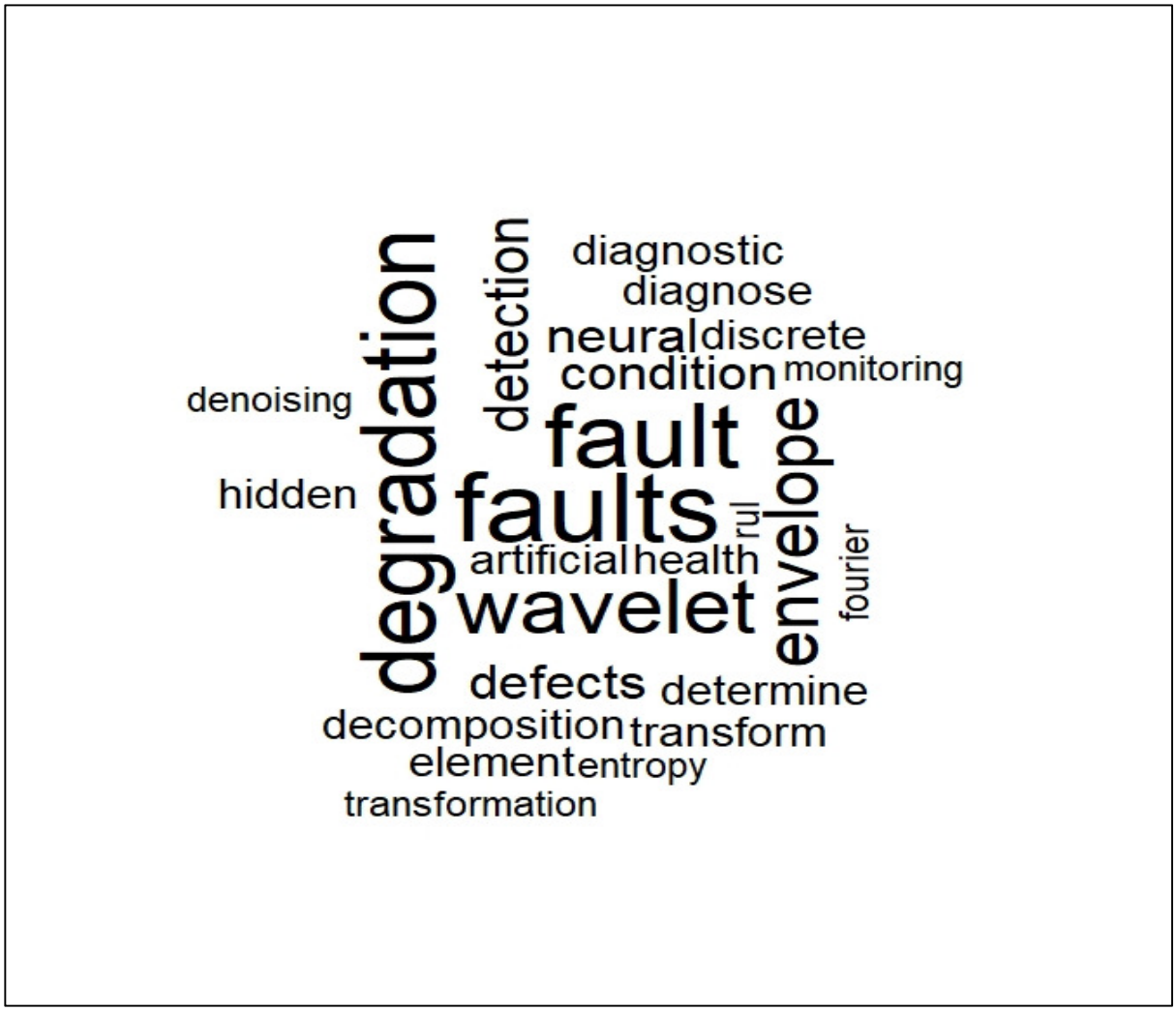


Figure 5.5: Word cloud of DTM for major contributions by the citations

Failure refers to a functional state of a system and only relates to the purpose of operation, irrespective of whether the system has a defect or not. Thus, if a component of the system is defective, part of the system could produce an incorrect signal data which defines and exposes the fault for the necessary actions to be undertaken. The presence of the fault could cause deviations in required operation data leading to errors, i.e. the process by which the presence of the fault becomes obvious.

Errors within the system sometimes do not cause any system malfunction unless a set threshold is exceeded. Thus, the system malfunction may not necessarily lead to system failure or a catastrophic effect, unless the design threshold has been exceeded. However, the degradation of the system or a system component caused by the error may cause the entire system to deviate from normal operation and eventually leading to system failure. The fault, degradation, and defects of the system may require a correction, and this is called diagnostics, which has already been covered in previous sections of this chapter.

5.6. Justification for Using Dataset for this Research

It could also be inferred that though some elements of QRA were observed in the publications which has use some of the available datasets for their research, none of the research specifically covered QRA as a main objective. This finding in section 5.5 and that of the review and critique of the incident investigation report of the real-life case histories in sections 4.2 and 4.3, provides a justification for the available bearing operation datasets to be used for this research.

5.7. Conclusion

This Chapter which forms the last chapter of Part 2 of this thesis, introduces the available datasets for the research and highlights some of the issues experienced during the search for the datasets. Details of the datasets, links to the source of the datasets, attributes of the data, the process by which the datasets were handle and stored were also covered. Because the datasets are open source datasets, the study went on to provide a list of publications which has use the data for their research, the big data techniques used as part of their research method and their major contribution to knowledge.

Although the list provided is a list of publications obtained from database search, the study acknowledges that the list may not be extensive since there could be other publications which may not have been found. However, from the objectives of their research, big data techniques used and the major contribution to knowledge reveals that there is a niche for a QRA method which relies entirely on big data techniques and real-time datasets which is the aim of this research. Next is Part 3 – Methodology, where the study provides chapters which cover the systematic approach for selecting the method for the research.

Part 3
Methodology

Part 3 - Background

In Part 2, the study performs a literature review on behavioural safety programs (BSP) as a way of checking whether the focus of safety in the HHPIs should be on behavioural elements or on the process operation itself. From accessing effectiveness of the BSP, the study found that the focus of safety in the industry should be on the process itself. As a result, the study performs a systematic review and content-analysis of research publications relating to existing quantitative risk analysis (QRA) methods, focusing on the use of big data techniques and real-time process operation datasets, to find a niche for additional contribution to knowledge. This was followed by the concept of dust fire and explosions. The study defines some of the terms of dust properties and explosion in the context of their use. Two case histories of real-life dust fire and explosion incidents were also presented to help illustrate the ideas that some dust explosion incidents are cause by the risk of failure of a process component including bearings. The final chapter provides an introduction of available datasets for the research and the justification for their use. The study now presents Chapters 6 - 9, which together form the systematic approach that led to obtaining the QRA method and detail procedure as the Part 3 - Methodology.

The study begins by providing clarity between a research methodology and a research method by adapting the explanation provided by Harding (1987, p. 19) who suggests that one problem encountered in a research relates to identifying a distinctive method, i.e. one which provides a distinction between the methods, methodologies and epistemologies and show how to pursue the research. Sandra explains that, the research methodology is the rational/theory and analysis of the process by which the research is conducted (Harding, 1988, p. 2). The research method on the other hand, is the technique by which evidence for the research are gathered (Harding, 1988, p. 2). To provide further clarity, Sandra added a third terminology - epistemology, which she expressed as the theory of knowledge (Harding, 1988, p. 3). Thus, the methodology is the justification for using the steps in the research approach while the research method is simply a research tool or component of research such as the method by which this study will obtain and provide the proposed QRA method and detail procedure. The epistemology is the theory of knowledge regarding the methods applied for the research which cover its scope and validity.
Reddy (2018) also explains that research methods are the strategies, tools, and techniques used by the researcher to collect the relevant evidence needed to create theories which must be credible, valid, and reliable. Hence every research must be accomplished by a methodology, which consists of a systematic and theoretical analysis of the research methods.

Part 3 as presented in this thesis is therefore base on the general advice to researchers by Thompson (2013) which includes the following:

- a methodological discussion, including discussion of epistemology and ontology as relevant

- the research design, including a discussion of methods with due recognition of their blank and blind spots.
- a clear audit trail of what the researcher has done,
- how much data was produced and how it connects to the research question(s),
- how the data was analysed, and
- a pointer to any particular problems/issues that arose.

Chapter 6 - Methodology

6.0 Introduction

In Chapter 5, the study introduces the data available for the research together with a review of 55 publications which has applied some of the datasets for their research. In this chapter, the study explains and presents a methodology to thoroughly solve the research problem by investigating different big data techniques and software packages to obtain the propose QRA method and procedure for data analysis. The study expects this to provide clarity of the methodology in terms of its correctness, reliability, and reproducibility in a clear, systematic, logical, and replicable manner with the aim of providing the answer to some of the research questions. The study also provides an outline of these steps which includes the concept of presenting data in visualise format to help communicate any insight and patterns within the data.

To achieve the above objectives, the study will use NASA Bearing Dataset No. 2 as a training dataset because:

- it has fewer variables and dimensions than the other NASA datasets,
- the type of risk suffered by the process is known, and
- the component bearing which suffered the risk is also known.

Issues encountered in data explorations and the solutions that help resolve the problems shall be covered in Chapter 7 -Data Exploration and Challenges. The rationale behind the path of data management and exploration process used, the various PC software packages used and the reasons for their selection for this study will also be explained.

The overall approach for arriving at the proposed big data QRA method, which cover all the steps undertaken in Part 3, may be illustrated by the flowchart of Figure 6. As detailed in the flowchart, the process will begin with consideration and selection of big data techniques to apply. This will be followed by data pre-processing, followed by data exploration and data management processes (like problems with data exploration and how they are resolved). Next will be investigation of big data techniques and software packages, followed by selection of the techniques and packages deemed appropriate (with justification) for the big data QRA method. The selected big data techniques and software packages will then be applied to the Training dataset to help obtain the big data QRA method. The study will then provide a detail step-by-step procedure for the QRA method obtained in through reporting of its findings as a justification that big data technique can be applied to real-time process monitoring data for QRA within the HHPI.

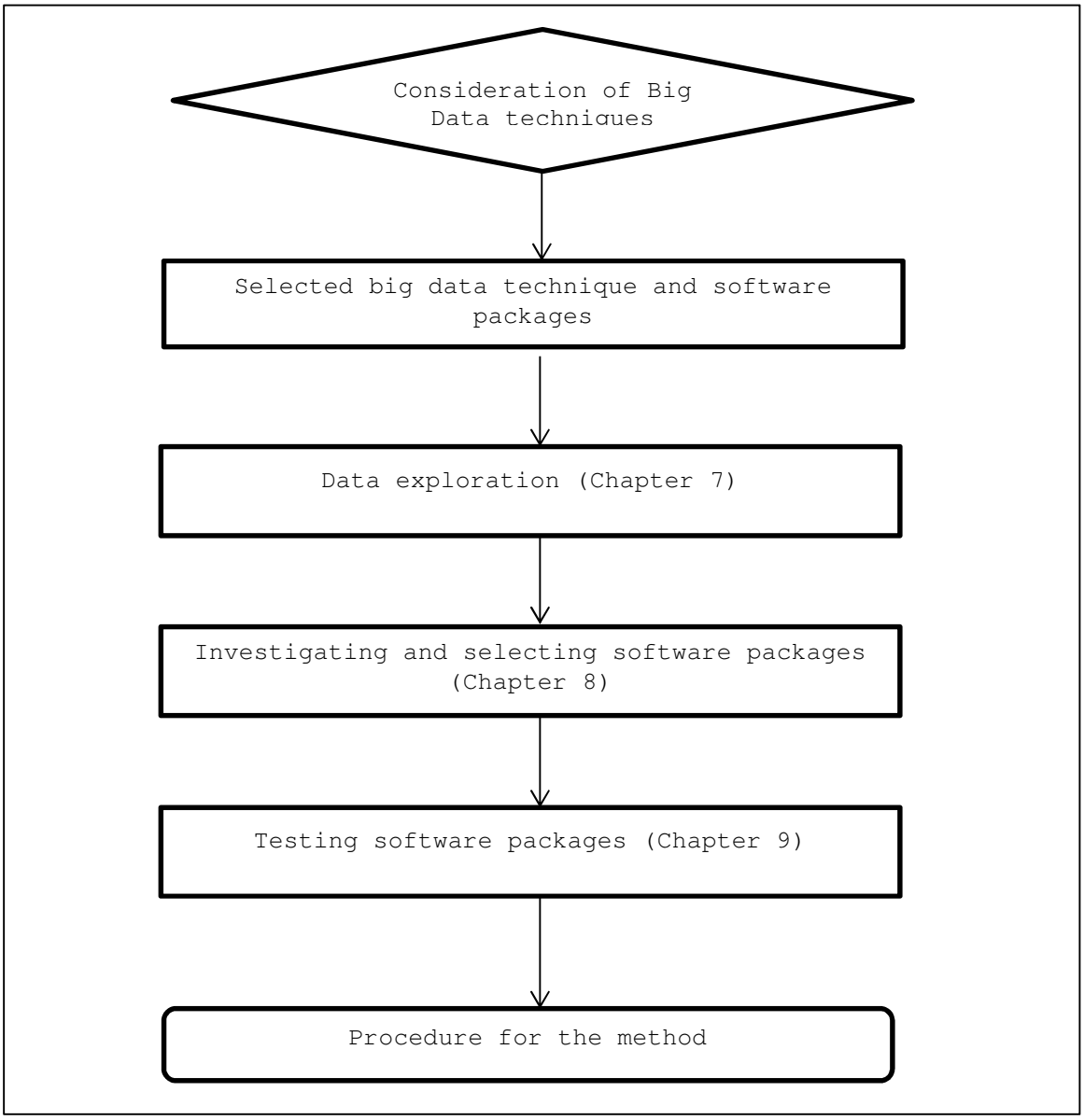


Figure 6: Flowchart for approach to investigating and obtaining the propose QRA method

6.1. Considering Big Data Techniques

The study considered the concept of Machine learning (ML), Statistics, artificial intelligence (AI) and data mining as the first step towards obtaining the data analysis method. Although it is possible to apply all four concepts to the training dataset to obtain the method for data analysis, the study acknowledge that such process would be laborious because they differ in the approach used. It has been suggested that in theory there is no difference between the four techniques. As Wasserman (2014) states, both statistics and ML refers to the science of learning from data and though the two fields differ in their history, conventions, emphasis and culture, they are identical. Hence as part of the process for selecting the concept to apply, the study carefully considered the similarities and differences in the four as below.

Statistics generally involves the collection and organization of data, analysing the data to determine any existing relationships, as well as interpretation and presentation of the findings. Thus, statistics deals with all aspects of data, including the planning and design of the experiments from which the data is generated.

AI has been applied as a research tool “for specific scientific and engineering problems and maintains a distance to the cognitive sciences” (Muller & Bostrom 2016, p.1). It involves the principle that every aspect of learning or any other feature of intelligence within a set of data can be simulated by a machine (McCarthy, Minsky, Rochester, & Shannon 1955, p. 1). Thus, the AI approach could either be applied to monitor the process operation itself to detect risk with the aid programmed computers as an intelligent agent without learning from data generated from the process operation.

ML involves extracting and/or predicting events using information derived from a training dataset. It applies data mining technique like ‘supervised or unsupervised learning’ and data pre-processing steps to improvement of learner accuracy (ARE 2014). Thus, ML tends to focus on inference in non-standard situations which often forms the basis for the algorithms used (Marsupial 2010).

Data mining focuses on the unknown and involves extracting information from a dataset, then transforming it into understandable structure for further use (ARE 2014). As mention above, a typical data mining process involves various statistics and ML techniques and can be applied to large datasets.

Owing to the above considerations and the size of the data available for the research being time stamped, the study selected data mining as the technique for the risk detection as analytical method to be investigated. To proceed with the investigations, the study considered two statistical model, Weibull distribution and time series analysis, to select the most suitable statistical model for context of the research.

6.2. Time Series vs. Weibull Analysis

Weibull distribution has been extensively applied in engineering as a “statistical model for studying fatigue and endurance life in engineering devices and materials to analyse the breaking strength of materials” (Aryal & Tsokos 2011, pp 89). It has also been applied in other areas of research including time-to-failure distribution analysis (Rajkumar et al., 2011), traffic modelling (Ageyev & Qasim 2015) and other useful application including capacitor, survival analysis, ball bearing, communication systems engineering, reliability testing, relay and material strength failures, theoretical maximum value, description and analysis of energy from wind turbines and wind speeds, the weather and in finance and general insurance(Salh 2014).

A Weibull model allows monotonically increasing or decreasing hazard functions within the distribution. It is a general-purpose reliability-engineering model, which assumes exponential, normal, as well as other types of distribution depending on the shape parameters (Herbert, Iniyan & Goic 2010). Herbert explains that, estimation of Weibull distribution parameters can be done graphically using probability plotting or analytically, using least square or maximum likelihood techniques. Weibull models incorporates time series data for the fitting of a parametric distribution to characterise time-to-failure probability in manufacturing systems and components (Zhaiet et al. 2013). A typical model can also use only time series data which is normally distributed but with no representation (Kadhem et al. 2017).

Time series analysis on the other hand, has been applied in various applications to obtain meaningful insight into the characteristics of data and for predicting future events from predetermined events. It has been applied together with other statistical methods like regression analysis to test theories including interdependence of variables and time series and to detect further events in situations where data exhibit some form of distinct ordering. Typically, a time series events can be (a) stationary, i.e. without any time-variance in the underlying process or with opposing or smooth sudden changes, or (b) nonstationary, where there are abrupt changes in the events data.

The available datasets were recorded over given time periods for which the study could provide a compact description and explanatory variables which can be processed and used with data mining models. As a result, the study opted for time series analysis as the preferred statistical method. Of the data mining techniques which incorporates time series, the study selected change-point analysis for the research.

6.3. Change-point Analysis

Changepoint analysis has been applied for time series datasets for research over the years. The concept has its origin in the article published by E.S. Page in 1963 with a focus on sequential detection of a change in the trend in process quality control data in manufacturing process. It has been proven to be a tool which aids the efficient understanding of essential information base on abrupt changes within data obtained from various sectors including oceanography (Killick, Eckley, Jonathan, & Ewans 2010), DNA copy (Killick & Eckley 2014), process behaviour in engineering (Sharma, Swayne, & Obimbo 2016) and meteorology (Arif et al. 2017). The methods for change-point detection can be sequential (online), where the analysis may be performed as data becomes available to identify abrupt changes near the most resent observation, or (b) retrospective (offline), where one off analysis may be performed on historical time series dataset (Page 1963). As a requirement, the big data QRA method would be deem as successful if it is able to handle data of

different dimensions as well as detecting single and multiple process risk events (Aminikhanghahi & Cook 2017; Fisher & Jensen 2018). This is also a requirement of change-point models. Thus, the method must have the ability of a balanced to handle historical and real-time datasets.

Change-point analysis, quality control, anomaly detection, breakout detection, segmentation, structural change, and event detection are similar and occasionally referred to change-point detections but have some differences. Some even refer to these as change-point mining because the process involves using data mining techniques to acquire knowledge (Boettcher, 2011). For instance, anomaly detection aims at detecting outliers, quality control focuses on the stability of mean and standard deviation, while change-point estimation only interprets changes in the data. The change-point methods have been investigated, categorize, and compared (Zarenistanak, Dhorde & Kripalani 2014). Reviews of change-point methods have been covered by previous researchers (Reeves & Chen 2007; Aminikhanghahi & Cook 2017). They include maximum likelihood estimation, regression, kernel methods, all of which have a limitation on the number of changes they can detect (Truong, Oudre & Vayatis 2018). As a result, no further review of publications on change-point analysis was performed by this study.

Although the methods can be grouped into supervised and unsupervised learning, the differences in the methods and the motivation for their application needs careful consideration before their usage. Thus, any significant changes within the data may represent a transitions event so that their detection could be useful at detecting the risk event within the process system (Aminikhanghahi & Cook 2017).

6.4. Applying Change-point Analysis for QRA

Catastrophic events from risk may come from different sources within the process system, including failure of a component, a fault, an error or damage within the system. As a result, the effect of risk can be mitigated if the source of the risk can be detected so that the necessary actions can be taken. However, for data from complex systems, as those obtained from process safety operations the detection of risk can be extremely difficult because of the limitations of the current QRA methods mentioned in Part 2 – Literature Review. To help overcome some of these limitations, some of the HHPIs have employed alarm generation risk detection systems (AGRDs) which issues alarms when any abnormally within the process operations data exceeds a set threshold. Some also incorporate system shutdowns into their alarm system which stops the process operation when the process data exceeds the set threshold.

Since risk is a product of the probability of occurrence of a catastrophic event and the severity of the loss caused by the event (CCPS 2000), there is the need for a system that could eliminate acceptable (e.g. false alarms) from actual operational risk. To achieve this, some researchers have

use principal component analysis (PCA), a dimensionality reduction technique which applies correlation between observed variables in the data (Russell 2000).

A typical PCA approach compresses a large set of variables into few important variables and provide the direction of the most dominant variance, giving indications of unwanted events within the process (Imtiaz, 2007). Although this study proposes the use of change-point analysis as a method for risk detection in the big data QRA method, PCA may be investigated where necessary. This is because although PCA has been extensively applied for risk detection in process systems, this study wants to gain new insights into risk detection and analysis. Hence, the study will investigate as many big data techniques as possible to obtain the most appropriate techniques to apply for the QRA method. All data points between any detected events may be considered as normal operational risks. However, if any detected point in the data gets to the region it could lead to catastrophic event.

6.5. Selection of Software Platform for the Method

Various software packages and platforms including Notepad, Notepad++ v. 7.5.1, TXT collector v.2.0.2, Batch Text Replacer, Microsoft Excel, the MS Excel combined with Kutools v.5.00 and BinaryMark batch files v.5.0.7.0 were investigated for this study. For instance, at the initial data preparation stage, attempt to use the software packages listed above packages were not successful due to software memory limitations. Table 6.5a is the list of software packages, a description of the task for which they were applied, and the challenges detected. As a result, R language platform was considered and applied for analysis of the training dataset.

Other reasons for considering the R language platform include:

(a) it is an open source platform therefore there is no cost involve,

(b) was extensively used as part of my preparation for this research because training on the use of the software was provided as part of my training and

(c) I have also applied R language for previous projects. Also, as an open source language platform, R has a collection of organised packages, including packages for change-point analysis which are portable for various computer operating systems with readily available help documentation (Paradis 2010).

The platform also has most of the requisite tools which has been applied for performing analysis reproducibly in various fields including DNA sequencing (Snellenburg et al. 2012; McMurdie & Holmes 2013) and various structural equation modelling (Boker et al. 2011). Besides R as platform is open source for statistical computing and graphics (Gasparrini 2011). Table 6.5b is a summary of change-point packages on the R language platform and summary description of their application which the study will consider as part of the process for the selection of appropriate package or

packages for the big data QRA method. After detecting the risk event, the study will investigate interaction effect at the time period of the risk event.

Table 6.5a: Software packages used for data reduction and their task description

Software Package	Task	Observed Pro's	Challenges Observed
R Studio v. 3.3.2	Converting data files from text to csv, merging, reading and writing tables.	Ease to use for the task applied.	Insufficient memory to handle and process the large size of data
Notepad++ v. 7.5.1	Deleting and merging text files.	Easy to apply for deleting lines within each file.	Does not support deleting of multiple lines within the files. Its application for this research was very laborious as the performance of deleting multiple lines in the files within the folder had to be conducted manually for each file. Hence its application was time-consuming.
Notepad	Editing text files.	Support manual editing (deleting of lines within each data file)	Does not support batch editing of the files hence application for the research was very laborious since it had to be performed manually on individual text files.
7-Zip v.16.04	Applied for unzipping the archive files downloaded from the NASA data repository.	Able to unzip the .rar files for conversion to text.	Option to create a new archive for the decompressed files was not available.
Batch Text Replacer v.2.8.0.0	Used to delete every 20479 lines in the text files.	Help save time with editing multiple text files and has good backup capability.	Not able to delete multiple lines within the files.
MO Excel & Kutools v.5.00	To combine, merge and edit my files quickly in excel.	Selecting and deleting content was a bit easy.	Couldn't combine all my files in excel for editing due to insufficient memory capability and row limit of < 105800.
TXTcollector v.2.0.2	Merge text files in the folder into one file for editing.	Able to merge file in the folder (Test Data No. 1.	Some files were skipped, and Error messages appear.
BinaryMark Batch Files v.5.0.7.0	For batch editing and merging the text files	User-friendly interface, able to edit flat and merge multiple flat text files.	The trial version has the limitation of working on 4 batch files at a time; different task couldn't be performed on the batch files at once but by a series of steps. Merged files had to be investigated with another program e.g. Notepad++ to ensure that merging was successful and in the order of arrangement of the processed files within the file directory.

Table 6.5b: R- Language change-point packages

Package	Usage	Summary Description
AnomalyDetection & BreakoutDetection	CPD using outliers and non-seasonal variations	Uses time series decomposition to identify change-points.
bfast	CPD for raster data	Analyse raster data (e.g. satellite images) time series and handles missing data.
brca	CPD analysis of irregularly sampled data	Particularly identify behavioural changes in the animal movement data set.
Changepoint	CPD	Applies non-parametic & frequentist for finding change-points within data.
Changepoint.np	Nonparametric CPD	An extension to the change-point package.
ChangepointsHD	CPD for expensive and high-dimensional models	Allows for the efficient estimation of change-point in complicated models with high dimensional data.
ChangepointsVar	CPD for changes in variance	Detects change-point for variance piecewise constant models.
ChangepointTesting	CPD in clustered signals	Detects change-point for clustered signals and signals of low magnitude than standard methods.
qcc	CPD using the stability of mean and standard deviation	Applies quality control charts (e.g. Shewhart, Cusum, EWMA, etc.) to detect change-point.
Strucchange	CPD	Detects change-point by fitting, plotting and testing trend changes using regression models.

6.6. Interaction Effect

Interaction effect generally refers to a situation where the operation of one component of the system affects another component within the process system. It has been proposed that merely detecting the statistical significance of the interaction effect between the independent and moderating variables on the dependent variable without explanation and theoretical arguments cannot be considered a contribution to knowledge (Andersson, Cuervo-Cazurra & Nielsen 2014). So, in the context of this study, interaction effect at the period of the risk event will provide understanding to the types of systems being exhibited within the process operation at the period of the risk event.

As explain in Chapter 1, the components of the system may either exhibit organise simplicity or organise complexity (Goerlandt & Reniers 2018). To recap, for systems exhibiting organized simplicity, each component (subsystem, system element) acts independent of one another. As a result, any risk event detected in the operation of any of the components have no effect on any other component so there is no interaction effect. For systems exhibiting organized complexity, the operations of the components affect one another through non-linear interactions and feedback loops. As a result, any risk event detected in the operation of one component may be due to the contributions from the other components though interaction effect.

A note on terminology is in place, distinguishing the use of the terms 'moderators' and 'predictors' in this study. The study explains the two terms by adopting some definitions proposed by Grace-Martin (Grace-Martin 2018) as follows:

- Predictors are the components of the process system whose operations have a potential to produce an effect on the component which suffer the risk event, without any real distinction between their roles.
- Moderators are the components of the process system whose operations can make some contributions to the effect of the predictors on the component which suffer the risk.

The study applies linear regression models to determine the interaction effect at the change-points. To determine the predictor and moderator components for the regression model, the study applies a decision tree model. The study will also apply R-package "*stargazer*", to provide a summary of the regression results including the summary statistics and side-by-side comparison of the models (Hlavac 2014). Where significant interaction effect is detected in the regression, ANOVA type II will be applied and if the hypothesis valid, after which effect

plots will be produced to visually express the relationships and any uncertainties within the model measurements.

6.7. Conclusion

This chapter has introduced the steps in the framework adopted to arrive at a propose QRA method for data analysis for the research, particularly consideration of big data techniques, selecting change-point as a data mining technique, considered software packages and their usage, and method for investigating type of systems being exhibited by interaction up to the period of the risk event. Next is Chapter 7– Data Exploration and Challenges, which is a continuation of the framework for arriving at the data analysis method for this study.

Chapter 7 – Data Exploration and Challenges

7.0. Introduction

In Chapter 7, the study introduces the framework adopted to arrive at a propose data analysis method for the study and discuss the selection of change-point analysis as a big data techniques and associated software packages. In this chapter which is a continuation of the framework for arriving at the method, the study provides an insight into how big data techniques and software packages were investigated using the training dataset, the challenges encountered, and the solutions applied to resolve the challenges. As discuss in Chapter 6, the study applies NASA Bearing Dataset 2 as the training dataset to help arrive at the method.

7.1. Investigating Number of Observations within Data Files

Information on the dataset states that each test file in the dataset must have 20,480 observations. With each row within the data file representing one data point, it was expected that each test file should have 20,480 lines. Twenty data files were selected at random from the dataset at random for investigation. The files were opened and investigated using Notepad++ which confirms that the number of lines within each of the selected data file was found to be 20480. This led to the suspicion that the sampling rate or sampling time described in the information accompanying the data may be approximation.

7.1a. Challenge 1: Sampling Time and Sampling Rate

Each of the twenty test files reveals that the data has 20480 observations and 4 variables. From the sampling rate 20 kHz/sec as described in the information accompanying the data 20,000 rows are expected. Investigating the sampling time by dividing the number of rows observed in the data by the number of rows expected (i.e. $\frac{20480}{20000}$) gave a sampling time of 1.024 secs. This suggests that the 1-second sampling rate description in the notes may be an approximation. Investigating the sampling rate by comparing the number of rows determined from the data to the number of rows expected also gave a sampling rate of 20.48 kHz/sec. This confirms that the sampling time and sampling rate are approximations. The study therefore proceeds with further exploration of the data.

7.2. Data Exploration

Twenty test files were sampled at random for exploratory data analysis. As an example, the study presents exploratory data analysis of one of the test files as a sample datafile.

Exploring the data file using the R-studio gave the first six rows of the data presented in Figure 7.2a. The variables were renamed as Bearing1, Bearing2, Bearing3 and Bearing4 for V1 to V4 respectively, to reflect the bearings from which the data was collected.

	Bearing1	Bearing2	Bearing3	Bearing4
1	-0.132	0.154	0.056	-0.015
2	-0.029	-0.166	0.125	-0.027
3	-0.049	-0.107	0.122	0.002
4	-0.034	0.095	0.142	0.061
5	-0.146	0.059	0.105	0.071
6	-0.032	-0.022	0.039	0.059

Figure 7.2a: First six rows of sample datafile in Training Dataset

The descriptive statistics of the sample datafile (Figure 7.2b) reveals that there are no missing observations, no nulls and no NAs in the data. The mean vibration frequencies are less than 0, most of which are within the band 0.0 and -0.15. The highest max and min values were 1.02 Hz and -0.91 Hz. These were observed in the vibration frequencies of the data obtained from operations of Bearing 3. The standard deviations suggest that the spread of the distribution of the frequency of the vibrations of Bearing 3 is wider than that of the other bearings.

Summary Statistics

	Bearing1	Bearing2	Bearing3	Bearing4
nbr.val	2.048000e+04	2.048000e+04	2.048000e+04	2.048000e+04
nbr.null	1.630000e+02	1.630000e+02	1.710000e+02	2.950000e+02
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	-7.130000e-01	-4.590000e-01	-5.910000e-01	-2.730000e-01
max	6.980000e-01	4.440000e-01	7.300000e-01	3.050000e-01
range	1.411000e+00	9.030000e-01	1.321000e+00	5.780000e-01
sum	-5.068000e+01	-3.351100e+01	-4.995700e+01	-4.973100e+01
median	-2.000000e-03	-2.000000e-03	-2.000000e-03	-2.000000e-03
mean	-2.474609e-03	-1.636279e-03	-2.439307e-03	-2.428271e-03
SE.mean	9.219999e-04	7.748776e-04	7.975254e-04	4.791458e-04
CI.mean.0.95	1.807193e-03	1.518822e-03	1.563213e-03	9.391640e-04
var	1.740972e-02	1.229692e-02	1.302624e-02	4.701812e-03
std.dev	1.319459e-01	1.108915e-01	1.141325e-01	6.856976e-02
coef.var	-5.331988e+01	-6.777050e+01	-4.678893e+01	-2.823810e+01

Figure 7.2b: Descriptive statistics of sample datafile in Training Dataset

With the magnitude of the means frequency of the vibrations approximately equal to 0 Hz obtained from the data from operation of all four bearings, it was deemed that the means are less informative. As a result, the focus was shifted to other statistical features such as skewness and kurtosis as per advice from Upadhyay, Kumaraswamidhas & Azam (2013).

Output from investigating other statistical parameters (Figure 7.2c) shows small values of the interquartile range for the data from the operation of all 4 bearings. This could may be due to the magnitude of the frequency of the vibration of the bearings since the datafile is the first test file in the Training Dataset, i.e. data collected at the beginning of the process operation when the bearings are in their healthy state. The skewness reveals that the distribution of the frequency of the vibrations of Bearing 1 and Bearing 4 are negatively skewed while that of Bearing 2 and Bearing 3 are positively skewed. The kurtosis values suggest that the

scale-free movement of the mass of the probability from the shoulders to the centre and tails peaks more in Bearing 1, followed by Bearing 3, then Bearing 2, with Bearing 4 having the lowest peak. The highest kurtosis obtained for Bearing 1 led to the suspicion that Bearing 1 is in healthier condition (Kamaras and Dimitrakopoulos 2016) than the other three bearings.

Inter Quartile Range

Bearing1	Bearing2	Bearing3	Bearing4
0.154	0.149	0.149	0.093

Skewness

Bearing1	Bearing2	Bearing3	Bearing4
-0.18389401	0.03090508	0.01026294	-0.03160271

Kurtosis

Bearing1	Bearing2	Bearing3	Bearing4
4.728418	3.025117	3.425630	2.992294

Figure 7.2c: Other statistical parameters of sample datafile in Training Dataset

The plots of the data (Figure 7.2d) show some regularly spaced spikes in all four bearings. The plots are of a similar pattern with extreme values and a means closer to 0. Also, the vibration of all four bearings in this plot appear healthy because this is the first test file which represents data collected when all four bearings within the system are expected to be in in healthy conditions.

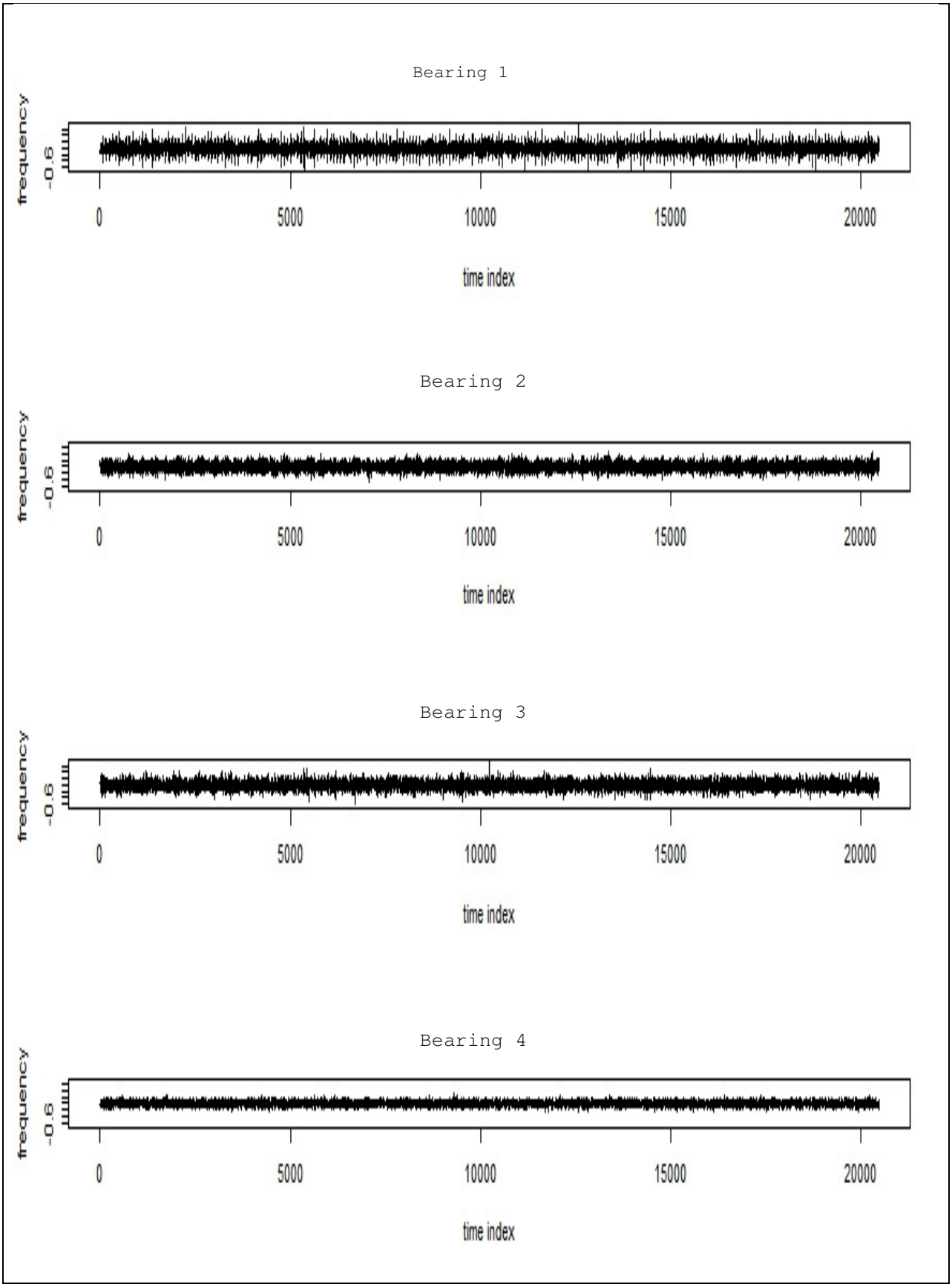


Figure 7.2d: Plots of vibrations of sample datafile in the Training Dataset

Inspecting the data with a box plot (Figure 7.2e) reveals that comparatively there are more extremes and outliers in the frequency of the vibrations of Bearing 1, followed by the frequency of the vibrations of Bearing 3, then the frequency of the vibrations of Bearing 2, and finally the frequency of the vibrations of Bearing 1 having the least extremes. The histogram plot (Figure 7.2f) shows unimodal distributions for frequency of the vibration of all four bearings. The quantile plot (Figure 7.2g) indicate lack of normality as the distributions are pulled asymmetrically towards higher values, indicating positive skewness.

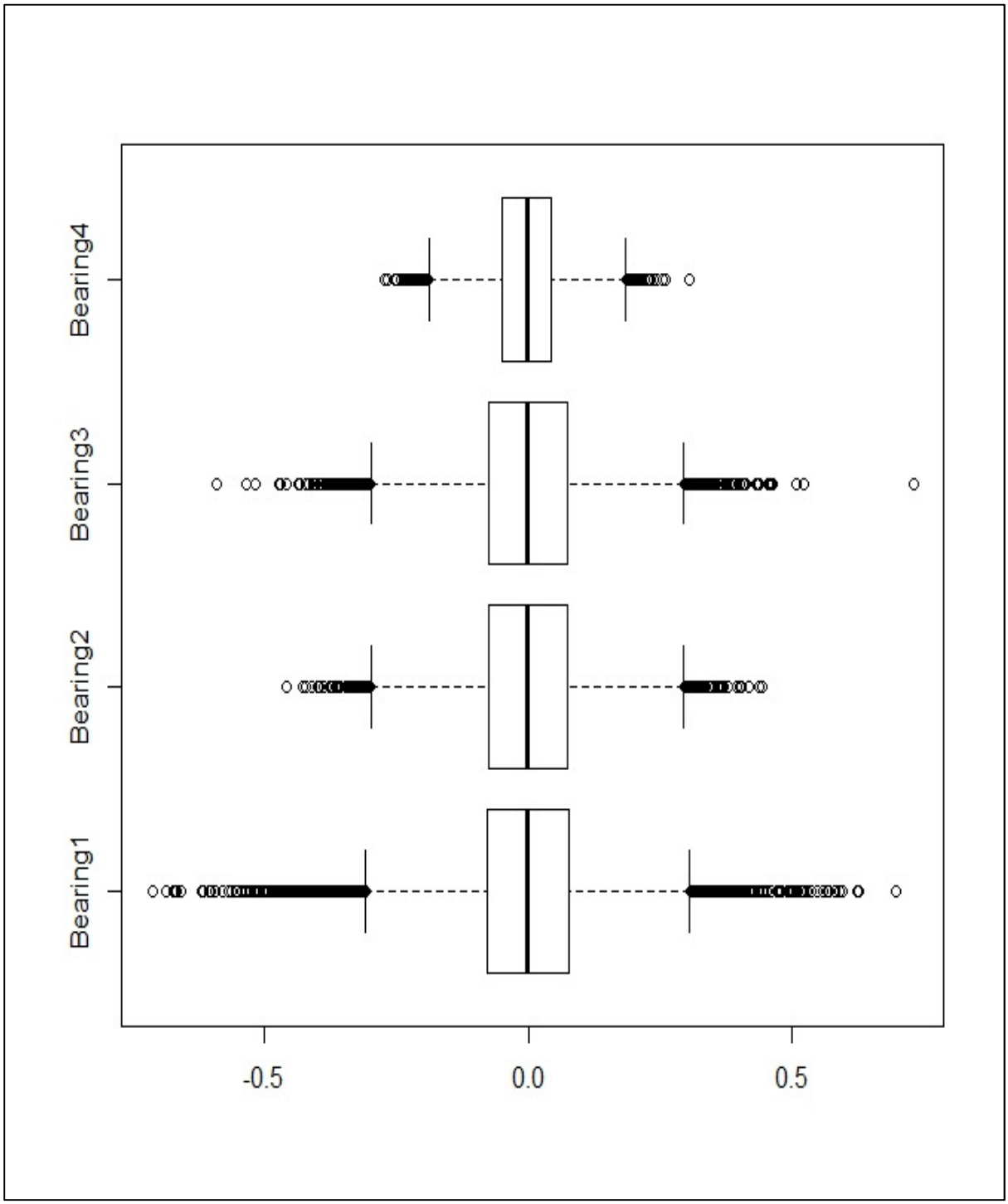


Figure 7.2e: Box plot of distribution of data in the Training Dataset.

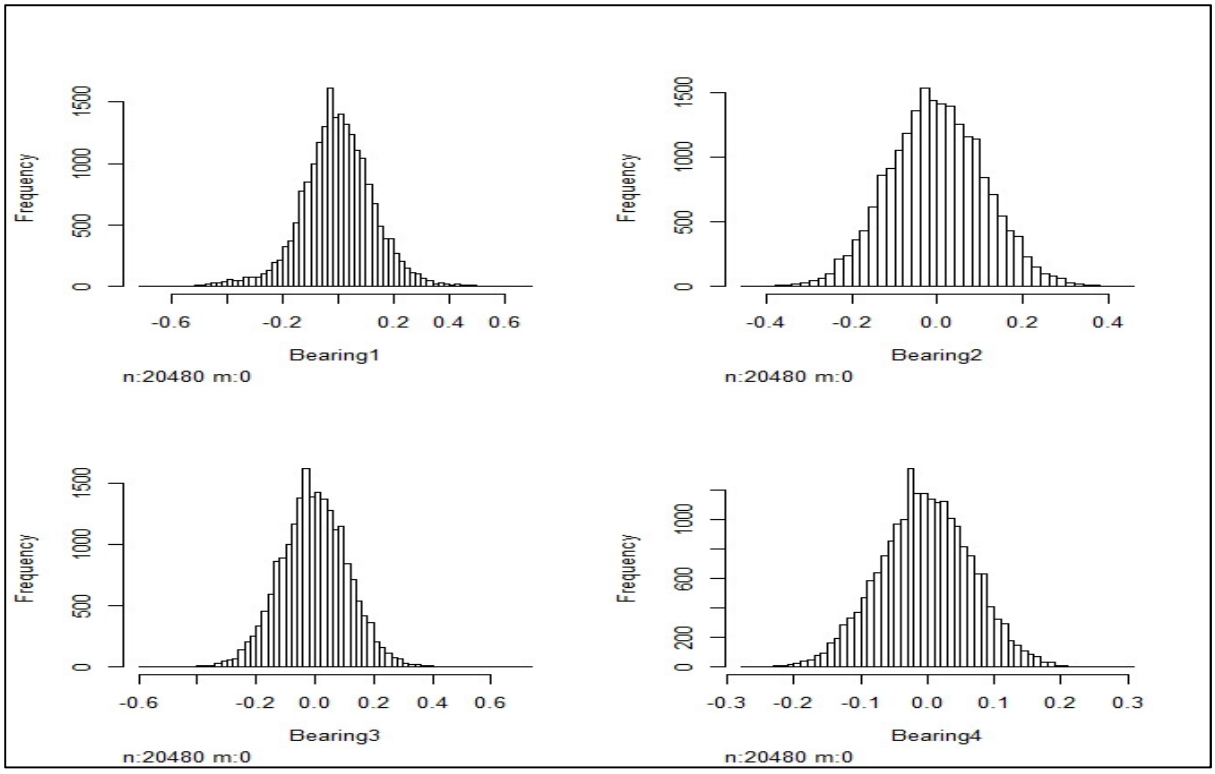


Figure 7.2f: Histogram of distribution of data in the Training Dataset.

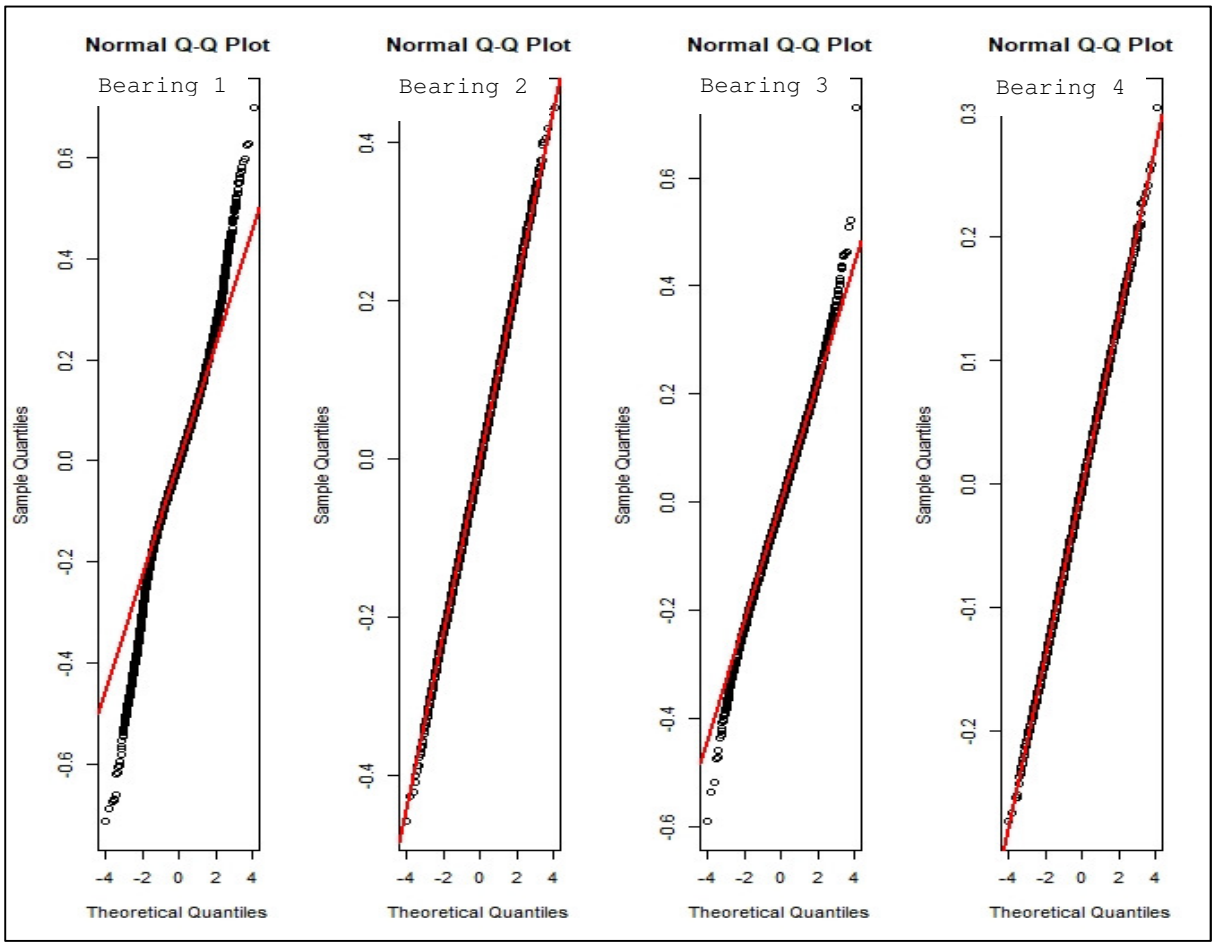


Figure 7.2g: QQ plots for the data in the Training Dataset.

It has been proposed by Karlsson et al. (2012) that undetected vibrations can seriously affect the power of the bearing signals. However, bearing vibration data has been described as one with multi-frequency components (Kwak et al. 2014). As a result, the study considered the outliers observed in the data from analysis using boxplot as part of the data.

Time sequence and lag plots were applied to investigate time effect on the frequency of the vibrations, any potential bias in the operation of the process system from which the data was generated and randomness of the data. The run sequence plot (Figure 7.2h) shows no response with time, which means there is no time effect on the operation of the component bearings within the process system. The lag plots (Figure 7.2i), shows a positive linear trend of the frequency of the vibrations of all four bearings which also indicates that the underlying data of the frequency of the vibrations are non-random. The low p-values of the Pearson's correlation test (Figure 7.2j) are suspected to be due to the large size of the observations (N) or proper relationships between the vibration of the bearings.

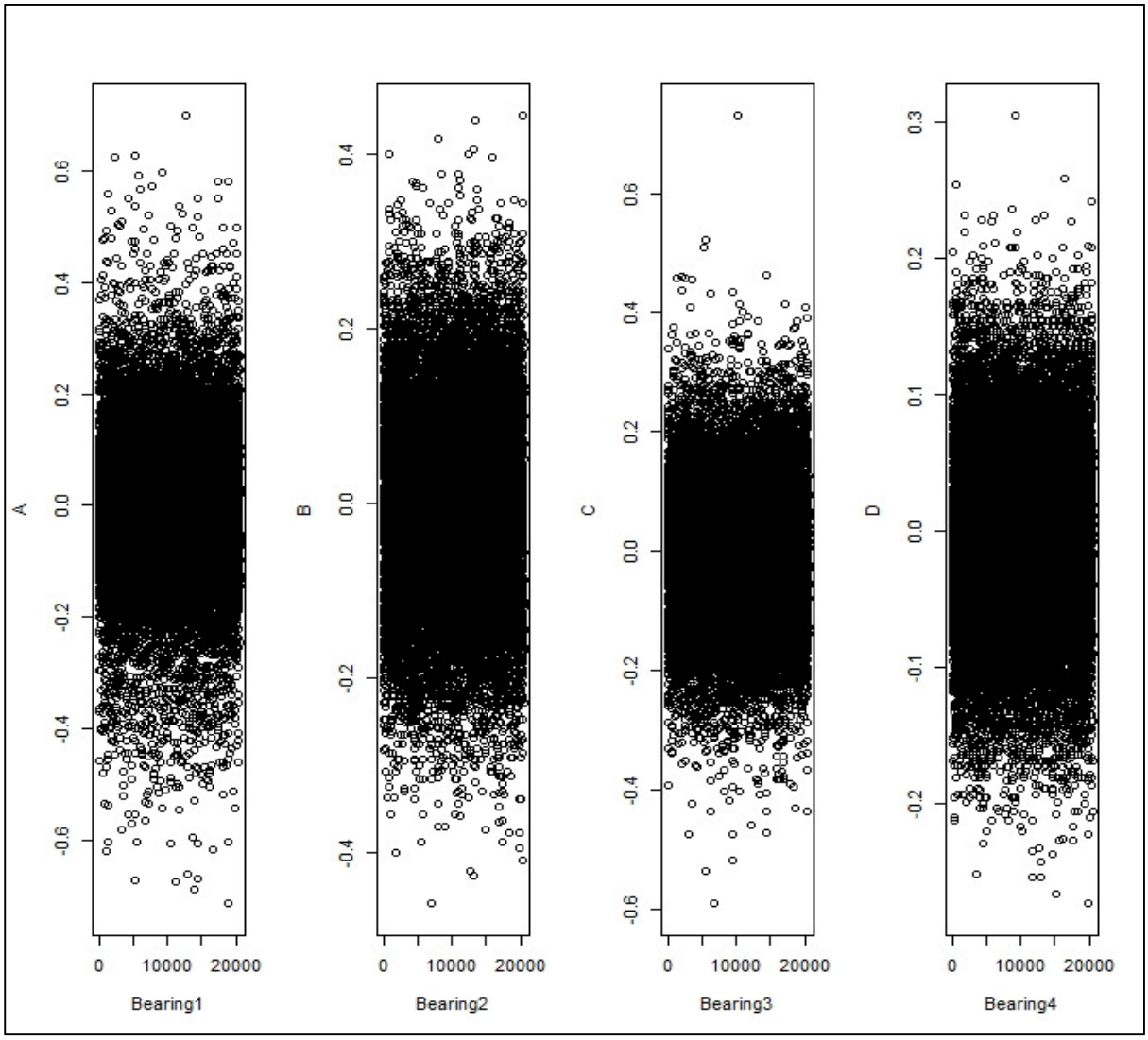


Figure 7.2h: Sequence plot of sample datafile in the Training Dataset

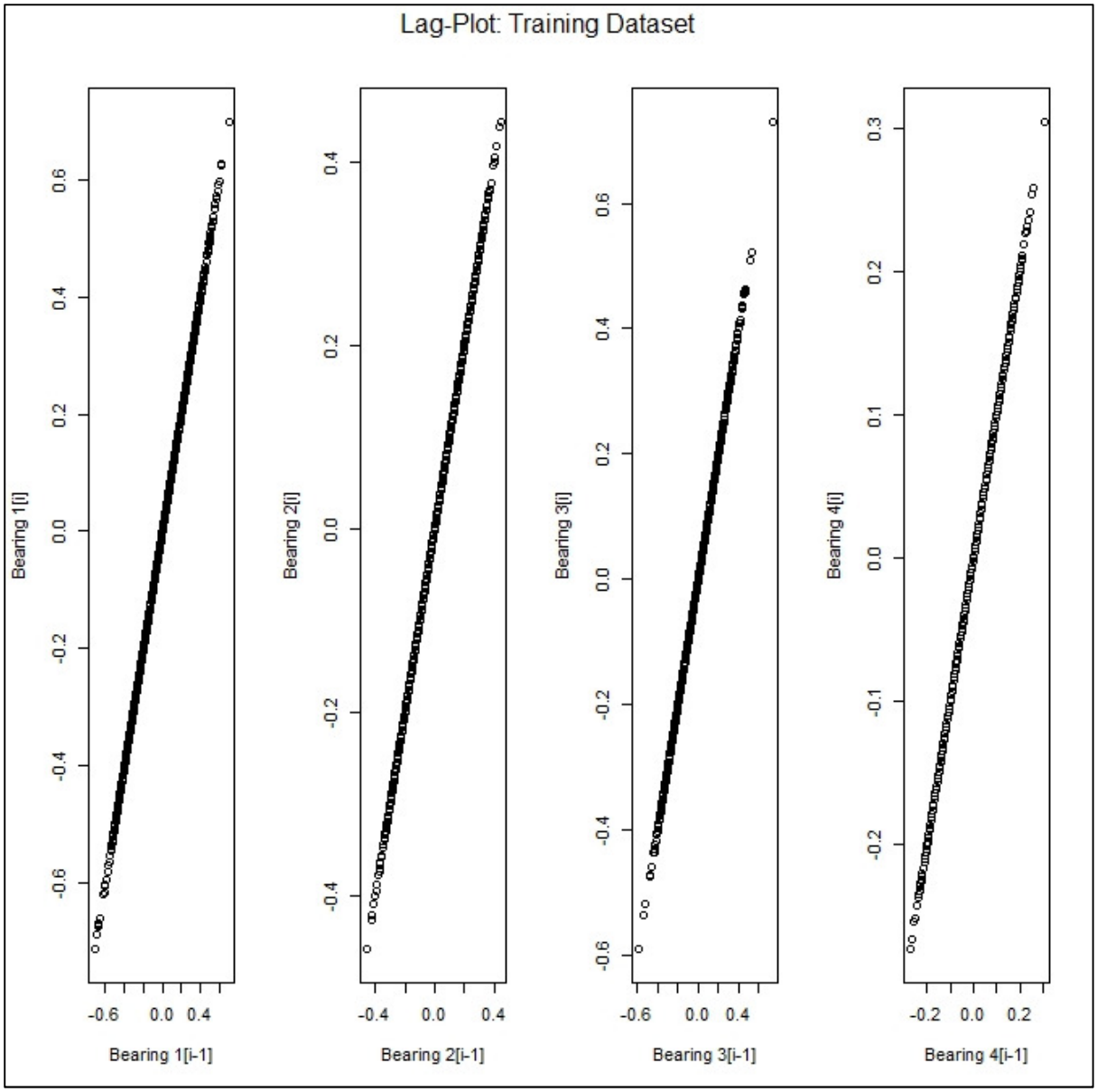


Figure 7.2i: Lag plot for data in Training Dataset

Pearson's correlation matrix

	Bearing1	Bearing2	Bearing3	Bearing4
Bearing1	1.00	0.22	-0.07	-0.07
Bearing2	0.22	1.00	-0.27	-0.34
Bearing3	-0.07	-0.27	1.00	0.24
Bearing4	-0.07	-0.34	0.24	1.00

n= 20480

Figure 7.2j: Result of Anderson-Darling test for data in Training Dataset.

This observation together with the position of the bearings within the process system led to the suspicion of some communication between the component bearings through bearing-bearing interaction. The study therefore suspects that the operation of the bearings within the process are not truly independent. The study proceeds with analysing all the test files in the dataset.

7.3. Challenge 2: Analysis of Combined Files within Dataset

The files in the dataset were combined to obtain a data file with 4 variables and 20,152,320 observations. Several attempts were made to analyse the data, but software memory could not handle the size of the combined data file causing R-studio to crash and freeze. The study therefore applies alternative solutions to manage the data for analysis as discuss below.

7.3.1. Attempted Solution 1: Reducing Data

The first solution attempted was to reduce the data volume by sampling every 1-minute observation in the data to obtain 984 observations for analysis. Of the software packages listed in Table 6.5a under Section 6.5page 102, BinaryMark Batch Files software was successful for this task. The process of reducing the data with the BinaryMark Batch Files software involves removing every 2047th lines via batch processing using the software, then stored in folders named by the number of edits. For instance, after the first process of editing the data by deleting the first 2047 lines, the data files were stored in a subfolder named D2Edit1. This process was adopted to help monitor the stage-wise data reduction process. All 984 data files were then merged after the reducing and storing process.

Variation in the data was investigated using a profile plot (Figure 7.3a) which shows that the mean and standard deviation of the variables appear similar up to about 7,000 data points beyond which some variations occur in the order of Bearing 1 > Bearing 4 >Bearing 3 > Bearing 2 registering the least variation.

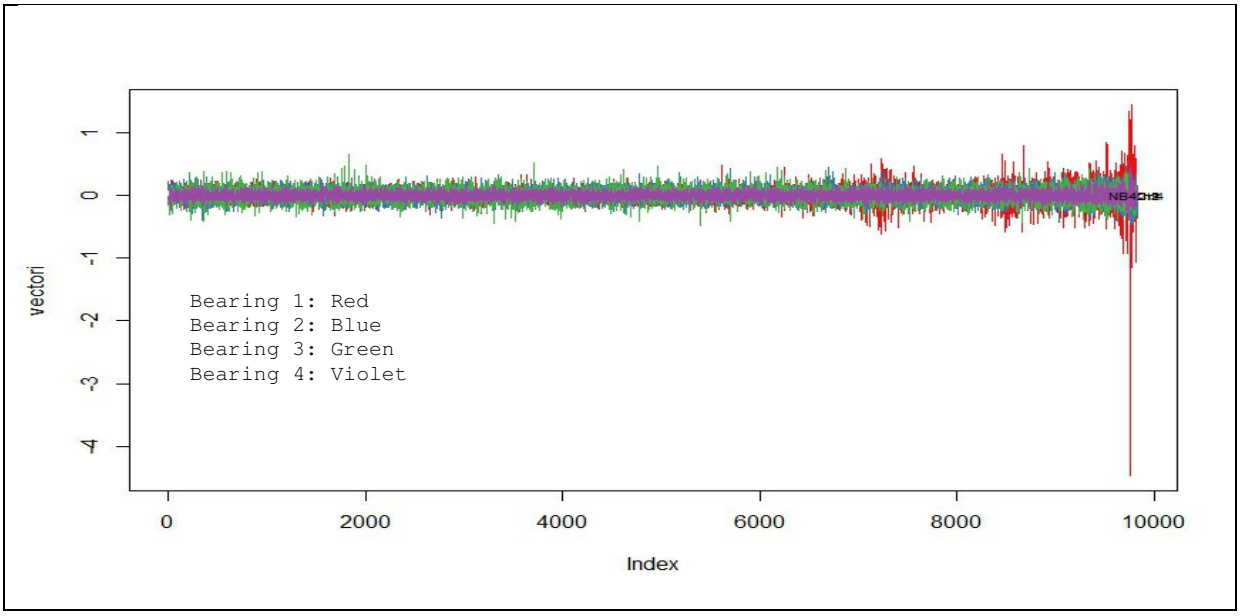


Figure 7.3a: Profile plot for data in Training Dataset.

A multivariate time series plot (Figure 7.3b) which gave a time series matrix divided into three colour patterns, with purple for low values, grey for medium values and green for high values obtained. The plot reveals there are higher values in the data obtained from the operation of Bearing 1 than that of the other bearings.

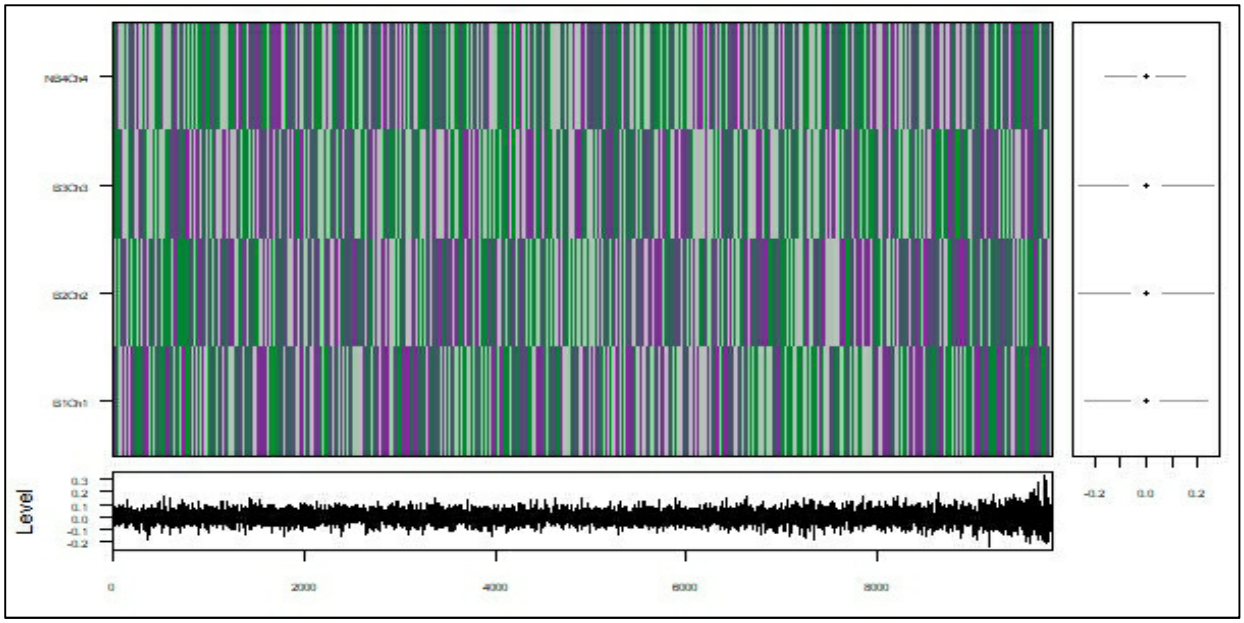


Figure 7.3b: Multivariate time series plot of Training Dataset

Time series decomposition of the individual variables using a frequency 60 to specify the hourly time series gave a plot for Bearing 1 (Figure 7.3c), which reveals no seasonality. A similar observation was obtained in the decomposition time series plot for the other three component bearings (Figure 7.3d to 7.3f, Appendix page 235).

Figure 7.3c: Seasonal trend in Bearing 1

The study created a periodogram to help obtain the highest possible frequency values and establish periodicity within the variables. The periodograms (Figure 7.3g and Figure 7.3h) shows no trend around frequency 0.05 Hz. The highest "power" frequency in the periodogram which applies to the main seasonality was obtained as 21.1 min (Figure 7.3i to Figure 7.3l, Appendix page 236). This explains the no seasonality was observed.

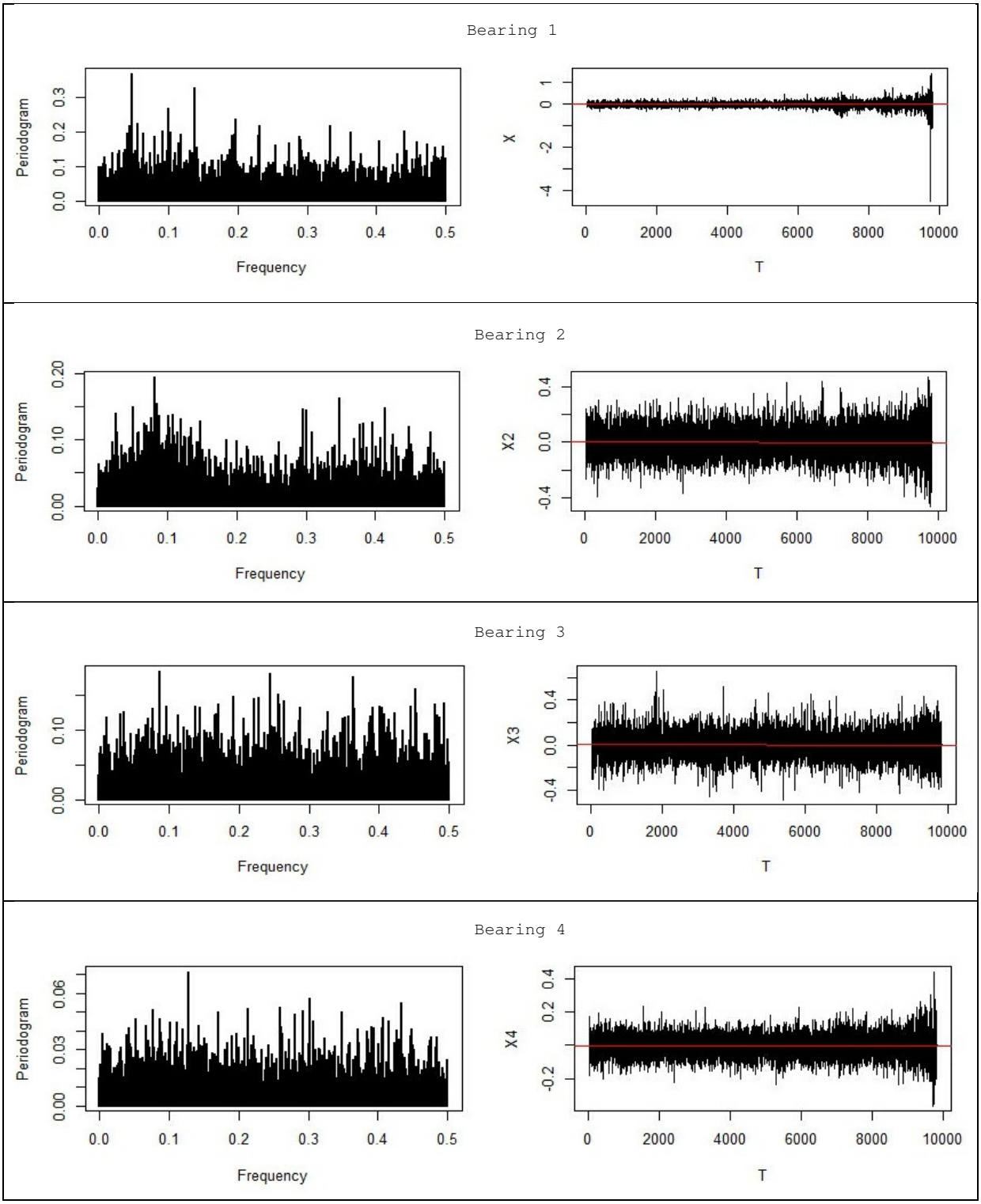


Figure 7.3g: Periodogram and trend for data in Training Dataset.

Upon careful considerations of all the investigations, it was thought that some vital information and features which could be derived from the data could be missed. As a result, a decision was made to discontinue using this approach. The study therefore considers PCA as an alternative solution.

7.3.2. Attempted Solution 2: Applying PCA

As discussed earlier in Chapter 6 under Section 6.4, PCA has been applied to compresses a large set into few important variables to provide the direction of the most dominant variance and giving indications of the unwanted events within a process system (Russell, 2000; Imtiaz, 2007). As a result, the study applied PCA to capture most of the variation between the data. From the proportion of variance of the summary of the PCA (Figure 7.3m) and screen plot (Figure 7.3n), two components applies. The PC1 variability explains 33% of the total variance of the data. From the rotations of PC1, Bearing 2 and Bearing 4 appear to be strongly related but in the opposite direction. This suggests that there seem to be a latent factor affecting all four bearings, with Bearing 1 and Bearing 2 in the opposite direction to Bearing 3 and Bearing 4. The study therefore proceeds with factor analysis (FA) to identify the latent factor affecting the operation of the bearings.

```
Summary of PCA
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.1560  1.0040  0.9505  0.8673
Proportion of Variance 0.3341  0.2520  0.2259  0.1881
Cumulative Proportion  0.3341  0.5861  0.8119  1.0000

Rotation
                PC1         PC2         PC3          PC4
B1Ch1     0.2204400 -0.9216484  0.1195625  0.29610017
B2Ch2     0.6315460 -0.1145919 -0.2747132 -0.71592669
B3Ch3    -0.4587734 -0.2332768 -0.8565108 -0.03870511
NB4Ch4   -0.5848785 -0.2881231  0.4202698 -0.63109078
```

Figure 7.3m: Summary of PCA of Training Dataset

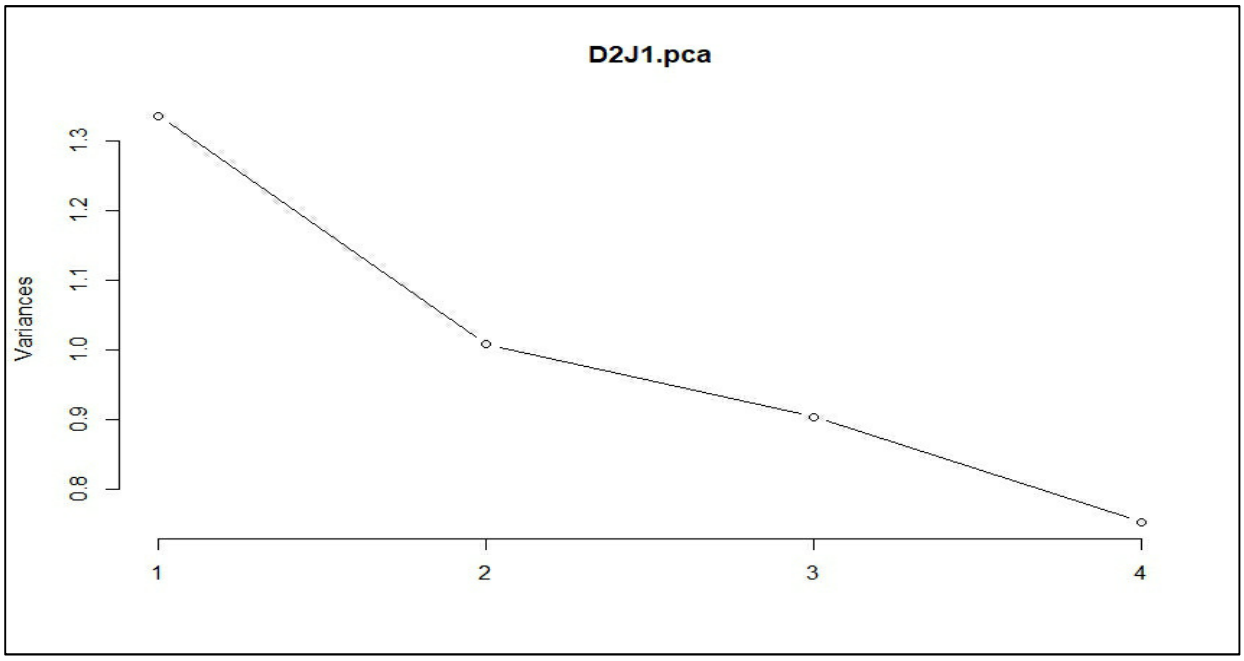


Figure 7.3n: PCA of Training Dataset

The FA plot from the FA (Figure 7.3o) suggests two factors without any components. The output of the factor's loadings (Figure 7.3p) and the FA diagram (Figure 7.3q) reveals that the 2 factors model exhibit a good fit. The FA diagram also reveals that the Bearing 2 and Bearing 4 are affected by the same factor but in the opposite direction, while a different factor affects Bearing 1. However, Bearing 3 appears unaffected by any of the two factors. As a result, the study applies adequacy test to determine the validity of the models.

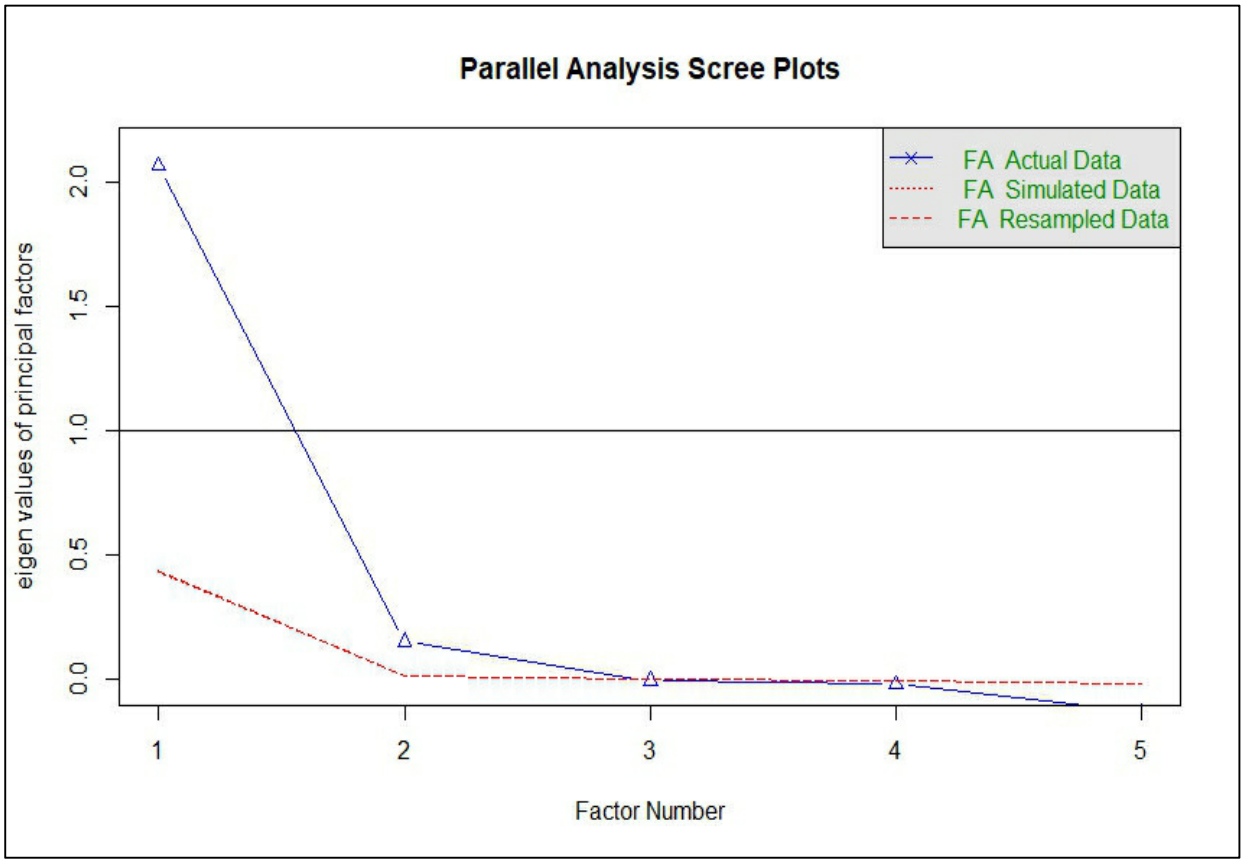


Figure 7.3o: FA screen plot for Training Dataset

```
FA with 1 factor
Standardized loadings (pattern matrix) based upon correlation matrix
         MR1    h2   u2 com
B1Ch1   0.12 0.014 0.99   1
B2Ch2   0.58 0.339 0.66   1
B3Ch3  -0.24 0.059 0.94   1
NB4Ch4 -0.39 0.150 0.85   1
                 MR1
SS loadings     0.56
Proportion Var  0.14

FA with 2 factors
Standardized loadings (pattern matrix) based upon correlation matrix
         MR1   MR2    h2   u2 com
B1Ch1  -0.02  0.35 0.127 0.87 1.0
B2Ch2  -0.48  0.16 0.293 0.71 1.2
B3Ch3   0.25  0.01 0.062 0.94 1.0
NB4Ch4  0.50  0.13 0.242 0.76 1.1
                       MR1  MR2
SS loadings           0.55 0.17
Proportion Var        0.14 0.04
Cumulative Var        0.14 0.18
Proportion Explained  0.76 0.24
Cumulative Proportion 0.76 1.00
```

Figure 7.3p: FA loadings for Training Dataset

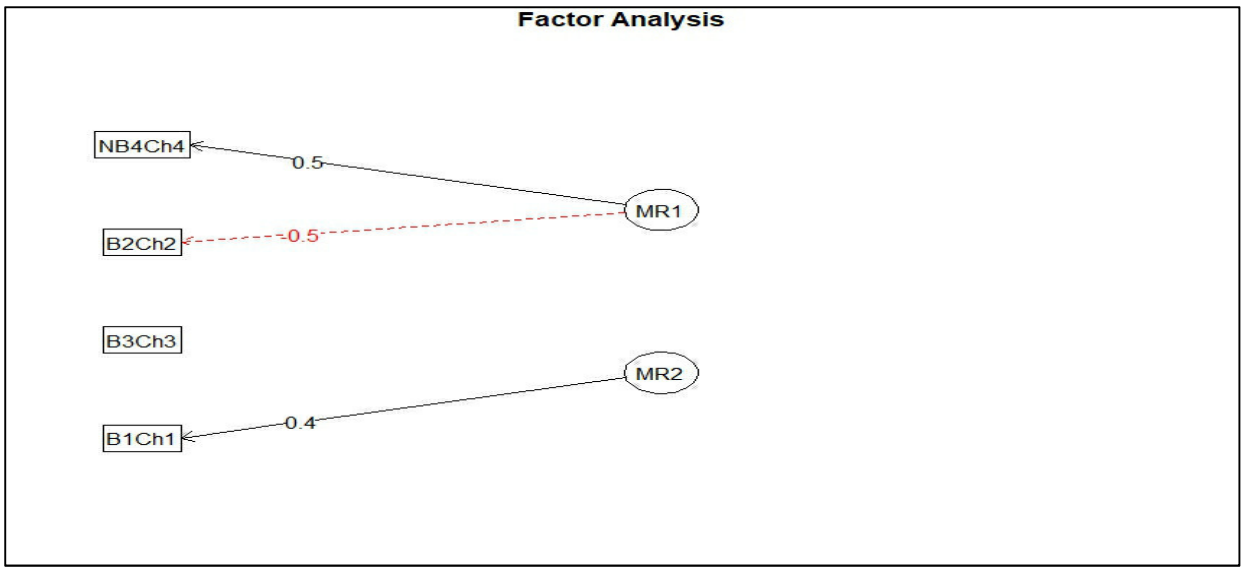


Figure 7.3q: FA diagram for Training Dataset

The root-mean-squared-error of approximation (RMSEA) index of 0.053 obtained from the output of the adequacy test reveals that the two-factor model exhibits a good model fit.

7.3.3. Attempted Solution 3: Applying Big Data Tools in R

Another solution applied was using big data tools such as R-packages *data.table* to read the data file, followed by package *bigmemory* to handle the large dataset and help alleviate the memory problems. However, this approach did not improve the issues of memory capacity as R freezes during analysis. After several unsuccessful attempts, the study applied packages *biganalytics* and *bigtabulate.* However, it was observed that *biganalytics* and *bigtabulate* couldn't handle out-of-memory datasets when applied due to memory capacity and CPU time. The study therefore considers another approach which involves slicing the combined data into chunks prior to analysis.

7.3.4. Attempted Solution 4: Slicing Data into Chunks for Analysis

The data was sliced into successive subsets of 2,000s for analysis. The first 2,000 data points were created; inspect normality using boxplot. The process was repeated by increasing the data points to 4,000 by including the next 2,000 data point. This was repeated by analysing successive increments of 2000 observation until all the data has been analysed. As a sample the boxplot obtained for the first 6,000 observations (Figure 7.5) reveals more extremes in the data obtained from the operations of Bearing 3 than in the data for the other bearings. However, this process was found to be very laborious and therefore not appropriate for the study due to time constraints. As a result, an alternative solution was sourced through various internet sources to help condense the data without losing key features was applied. One such approach selected involves applying feature extraction

technique to the data prior to analysis which has been used by other bearing vibration researchers (Caesarendra & Tjahjowidodo, 2017).

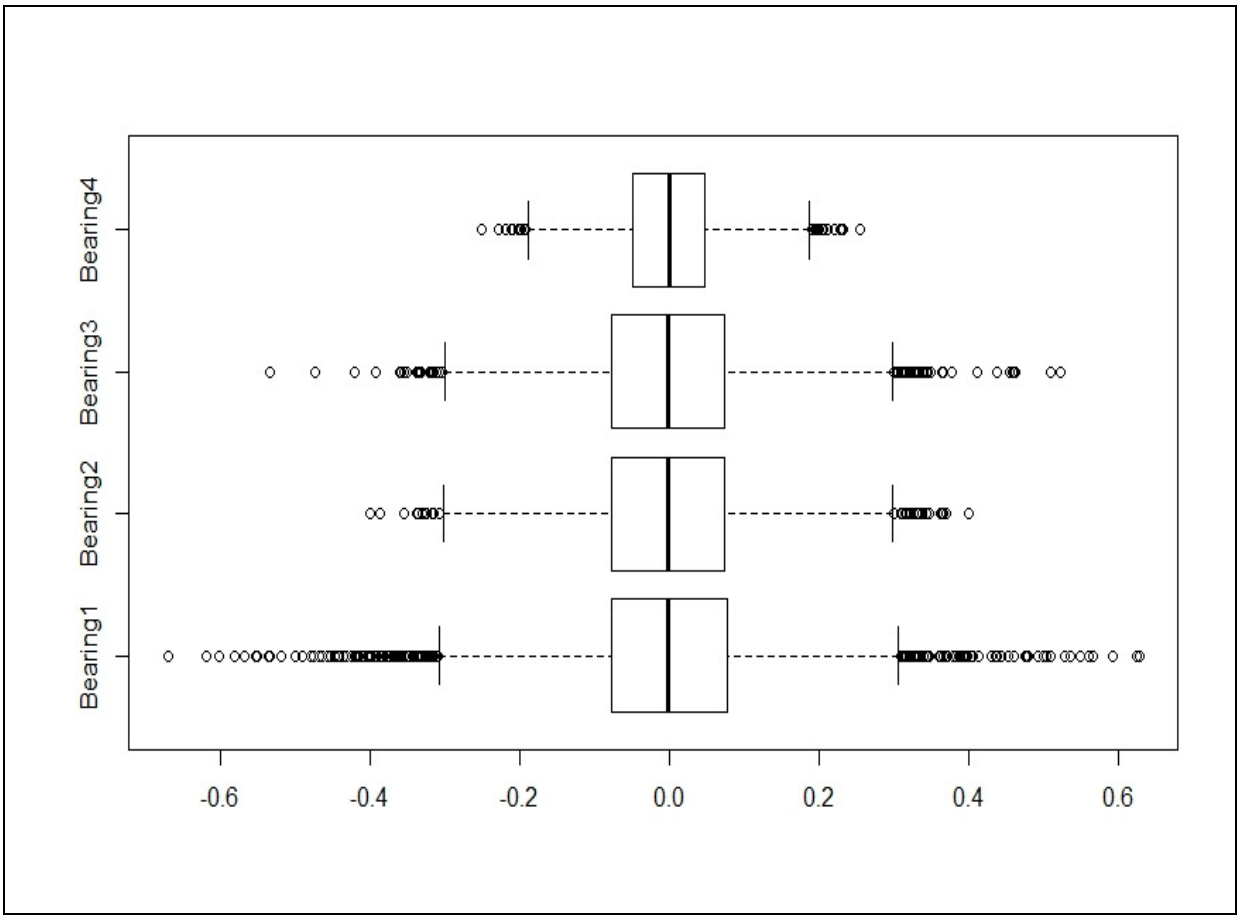


Figure 7.5: Sample boxplot obtained for first 6,000 observation in Training Dataset

7.3.5. Attempted Solution 5: Condensing Data by Applying Feature Extraction

The most dominant features from bearing vibration data are the level of the key vibration frequencies (Nistane & Harsha, 2016; Saruhan et al, 2014). These features depend on the health conditions of the bearings and can be examined by probability density function (PDF) of the vibration signal. Table 7.6a is a summary of the features and their method of determination.

The details of the bearings in the dataset was provided by the researchers who donated the data (Qiu et al 2006) as:

Ball diameter (Bd): 0.331 inc

Number of rolling elements (Nb): 16

Rotational speed (Rs): 2000 rpm = 33.3Hz

Pitch diameter (Pd): 2.815 inc

Contact angle (a): 15.17° = 0.084227 radians

Table 7.6a: Summary of bearing features

Features	Summary
Visually derived features	Use patterns in the data and develop a method to capture it.
Statistical features	Characterise the data on statistical properties like mean, median, standard deviation (SD), root mean square (RMS) skewness, kurtosis, entropy [$e(p)$], shape factor, crest factor (CF).
Time-frequency representation	Implement a map one-dimensional time-domain signals to a two-dimensional function of time. e.g. short-time Fourier transform (STFT), wavelet transform and Wigner-Ville distribution
Complexity measurements	Nonparametric tests to compare or measure the similarity of two cumulative distribution functions e.g. Kolmogorov-Smirnov (KS) and simple entropy tests.
Other features	Quantify the periodicity of the time series data e.g. singular value decomposition (SVD), piecewise aggregate approximation (PAA) & adaptive piecewise constant approximation (APCA).
Phase-Space Dissimilarity Measurement	Dissimilarity measurements used to quantify the signal complexity. e.g. Fractal Dimension, approximate entropy.

The study therefore calculated the rotational frequency (Rf) from the rotational speed as 33.3 Hz (i.e. 2000/60) and the contact angle (in radians) as 0.08427 (i.e. 15.17/180). The characteristic bearing fault frequencies (Granny & Starry, 2011; Kamaras & Dimitrakopoulos, 2016) are calculated as detailed in Table 7.6b. The frequencies include ball pass frequency for outer race (BPFO), ball pass frequency for inner race (BPFI), ball pass roller frequency (BSF) and the fundamental train frequency (FTF), and other statistical time-domain features such as shape factor (SF)Crest factor (CF), spectral kurtosis (SK) (Sohaib, Kim & Kim, J-M, 2017).

Table 7.6b: Brief description of bearing characteristic fault frequencies

Feature	Description
BPFO	The rate at which a ball or roller passes a point on the outer race $[0.5Rf(1 - Ratio)]$ = 14.69
BPFI	The rate at which a ball or roller passes a point on the inner race $[0.5Nb \times a(1 - Ratio)]$ = 0.595
BSF	The rate at which a point on a ball or roller will contact the inner OR outer race $[\frac{Pd}{Bd} \times 0.5a(1 - Ratio^2)]$ = 0.353
FTF	Rate at which bearing cage travels around the bearing $[0.5a(1 - Ratio)]$ = 0.0149
VHF	Very high frequency (> 6 kHz)
HF	High frequency (2.6 – 6 kHz)
MF	Medium frequency (1.5 – 2.6 kHz)
LF	Low frequency (0- 1.5 kHz)
CF	Calculate the magnitude of impact due to rolling and raceway contact, appropriate for spiky signals. It is the standard deviation divided by the RMS $(\frac{\sigma}{RMS})$
Entropy	A measure of the degree of randomness in the data.
Kurtosis	Quantifies the peak value of the PDF. Value is approximately 3 for a healthy bearing (Eftekharnejab et al, 2011).
RMS	Values help identify differences between vibration signals. Same applies to the mean and SD.
SF	The RMS to mean ratio. Value depends on an object's shape but independent of dimensions. $(\frac{RMS}{\mu})$
Sk	Quantifies symmetry of data, value approximates to 0 for healthy bearing., shifts to positive or negative when a fault develops.
Var	Measures the dispersion of the data around the mean.

Where

$$\text{Ratio} = \frac{Bd}{Pd} cos\ (a)$$

Frequency domain analysis using fast fourier transform (FFT) algorithm was applied to decompose the signals into their Nyquist frequency (Nf). The Nf is half the sampling rate of the of the signal (Seeber & Ulrici 2016). Applying FFT was aimed at removing noise signals from the observations without losing key features and therefore condenses the data to a size that could be handled by system memory for easy analysis in R.

Several attempts were made to perform FFT with R, including sourcing for information from various internet sources. Fortunately, the study found some work done in R by Victoria Catterson (Catterson 2013). Initial attempt using Catterson's R-code produced bearing-specific data file with 11 variables, 6 of which are duplicates. Figure 7.6a is a sample of the first 10 rows of the bearing-specific data obtained using Catterson's R-code. As a result, the study re-coded the Catterson's R-code to help achieve the study objectives.

12/02/2004 10:32	985.4477976	0	49.80955171	64.45941987	978.6111925	12/02/2004 10:32	985.4477976	0	49.80955171	64.45941987
12/02/2004 10:42	985.4477976	49.80955171	986.4244555	978.6111925	42.97294658	12/02/2004 10:42	985.4477976	49.80955171	986.4244555	978.6111925
12/02/2004 10:52	985.4477976	986.4244555	49.80955171	42.97294658	978.6111925	12/02/2004 10:52	985.4477976	986.4244555	49.80955171	42.97294658
12/02/2004 11:02	985.4477976	986.4244555	49.80955171	978.6111925	42.97294658	12/02/2004 11:02	985.4477976	986.4244555	49.80955171	978.6111925
12/02/2004 11:12	985.4477976	986.4244555	978.6111925	49.80955171	984.4711398	12/02/2004 11:12	985.4477976	986.4244555	978.6111925	49.80955171
12/02/2004 11:22	986.4244555	985.4477976	49.80955171	993.2610607	978.6111925	12/02/2004 11:22	986.4244555	985.4477976	49.80955171	993.2610607
12/02/2004 11:32	985.4477976	978.6111925	49.80955171	986.4244555	42.97294658	12/02/2004 11:32	985.4477976	978.6111925	49.80955171	986.4244555
12/02/2004 11:42	985.4477976	986.4244555	978.6111925	49.80955171	42.97294658	12/02/2004 11:42	985.4477976	986.4244555	978.6111925	49.80955171
12/02/2004 11:52	986.4244555	985.4477976	49.80955171	42.97294658	987.4011134	12/02/2004 11:52	986.4244555	985.4477976	49.80955171	42.97294658
12/02/2004 12:02	986.4244555	985.4477976	978.6111925	49.80955171	987.4011134	12/02/2004 12:02	986.4244555	985.4477976	978.6111925	49.80955171

Figure 7.6a: Sample of bearing specific data obtained using Catterson's R-code

The Nf of the data was calculated from the number of observations as 10,240 (i.e. 20480/2). The magnitude of each complex number (Mod) was then calculated from R function '*fft*' as the amplitude with the first half of the data points because the second half of the data is assumed to have the necessary and sufficient condition to achieve a complex conjugate symmetry of the first half in the frequency domain (Rippel, Snoek & Adams 2015; Catterson 2013). The study obtains eight frequency bands for each bearing by calculating the power frequency bands (Figure 7.6b). The characteristic bearing fault frequencies (first four frequencies) which are the key frequency features of the bearings appear similar because the data file being explored represent data for health bearings.

```
Calculate Power in Frequency bands

B1features
5.494612     2.775701     6.834377     6.907662 16809.401512
36822.497448 11808.692887 11906.973726


B2features
5.494612     2.775701     6.834377     6.907662 27839.564576
42054.367380 10752.759786 15279.800753


B3features
5.494612     2.775701     6.834377     6.907662 41565.290814
55881.211021 15972.941342 16945.322056

B4features
5.494612     2.775701     6.834377     6.907662 11814.912006
25310.248997  7543.014879 10849.492301
```

Figure 7.6b: Features in Frequency bands in the Training Dataset

Figure 7.6c is a plot of zoomed FFT profile the data after which reveals that none of the four characteristic fault frequencies (BSF as green, BPFO as blue, BPFI as Red and FTF as brown) has been affected. It was therefore decided that the study will apply feature extraction for all the datasets prior to all data analysis.

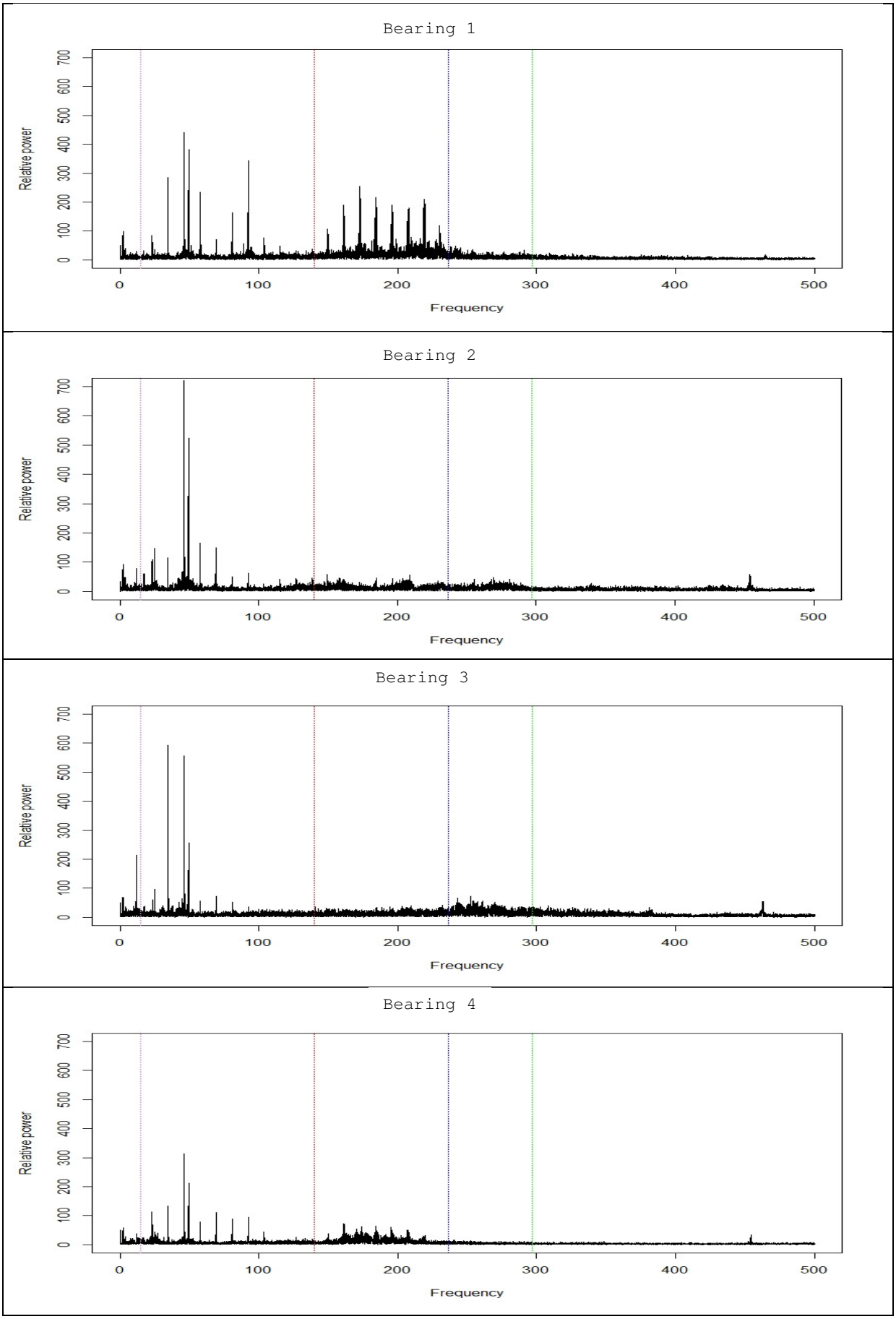


Figure 7.6c: Zoomed FFT profile plot for Training Dataset

With the aid of the feature extraction, the data was condensed by extracting the 23-key statistical and domain features (Min., Qu.1., Median, Qu.3, Max., Mean, SD., Skew, Kurt, RMS, FTF, BPFI, BPFO, BSF, F1, F2, F3, F4, F5, VHF, HF, MF, LF). The data was formatted in a time order of the source files to create four bearing-specific data files with 984 rows (i.e. one row per source file) and 23 variables (i.e. features). The files were timestamped according to the time on each test file in the dataset from which the key features were calculated. The files are written as data tables using R-function *write.table*, in '*csv*' format and saved for further use. A sample structure of the files created is presented as Figure 7.6d. The data files are inspected using descriptive statistics which reveals reveal no missing values.

```
Structure

'data.frame':     984 obs. of  24 variables:
 $ timestamp: Factor w/ 984 levels "2004-02-12 10:32:39",..: 1 2 3 4 5 6
7 8 9 10 ...
 $ Min.x     : num  -0.386 -0.388 -0.4 -0.576 -0.391 -0.366 -0.408 -0.334
-0.361 -0.344 ...
 $ Qu.1.x    : num  -0.059 -0.051 -0.054 -0.051 -0.054 -0.054 -0.054 -
0.054 -0.054 -0.051 ...
 $ Median.x  : num  -0.01 -0.002 -0.002 -0.002 -0.002 -0.002 -0.002 -0.002
-0.002 -0.002 ...
 $ Qu.3.x    : num  0.037 0.046 0.046 0.049 0.049 0.049 0.049 0.049 0.051
0.049 ...
 $ Max.x     : num  0.454 0.369 0.503 0.608 0.391 0.439 0.388 0.415 0.386
0.378 ...
 $ Mean.x    : num  -0.0102 -0.00259 -0.00248 -0.00228 -0.0024 ...
 $ SD.x      : num  0.0735 0.0753 0.0762 0.0787 0.0784 ...
 $ Skew.x    : num  0.084 0.0521 0.0328 0.0415 0.0282 ...
 $ Kurt.x    : num  0.628 0.648 0.513 1.158 0.603 ...
 $ RMS.x     : num  0.0742 0.0754 0.0762 0.0787 0.0785 ...
 $ FTF.x     : num  5.49 9.18 1.97 6.85 4.75 ...
 $ BPFI.x    : num  2.78 6.08 12.54 2.73 3.83 ...
 $ BPFO.x    : num  6.83 8.3 1.3 6.96 4.46 ...
 $ BSF.x     : num  6.91 12.96 15.3 2.32 15.05 ...
 $ F1.x      : num  985 985 985 985 985 ...
 $ F2.x      : num  0 49.8 986.4 986.4 986.4 ...
 $ F3.x      : num  49.8 986.4 49.8 49.8 978.6 ...
 $ F4.x      : num  64.5 978.6 43 978.6 49.8 ...
 $ F5.x      : num  979 43 979 43 984 ...
 $ VHF.pow.x: num  16809 18079 17380 17254 17297 ...
 $ HF.pow.x : num  36822 38358 38322 40367 39972 ...
 $ MF.pow.x : num  11809 11990 11961 12403 11967 ...
 $ LF.pow.x : num  11907 12636 13161 13214 13171 ...
```

Figure 7.6d: Sample structure of bearing-specific data obtained after feature extraction

7.4. Conclusion

This chapter has covered data exploration, challenges observed, and solutions applied to overcome the challenges. The study found that the sampling time and sampling rates

described by the notes accompanying the data are approximations. As a result, the actual sampling rate and time were re-calculated before data was used. The challenges observed during data exploration are mainly PC software memory issues. The study found feature extraction as one of the solutions which will be applied to all the datasets used in this research. Next is Chapter 8, where the study will investigate and select software packages for data analysis.

Chapter 8 – Investigation and Selection of Software Packages

8.0. Introduction

In Chapter 7, the study performs data explorations and selected feature extraction as a suitable technique for condensing the data for analysis without losing key information. In this chapter the study will investigate various change-point software packages on the R platform listed in Table 6.5b under Chapter 6, page 95. The packages will be investigated using the bearing-specific datasets obtained after the feature extraction process in Chapter 7. Based on the description of the packages and their applications, the study selects packages, *qcc*, *brca*, *changepoint* and *strucchange* for investigation. The first package selected for investigation was package *qcc*.

8.1. Investigating Package qcc

Using package *qcc* involves setting a threshold to control stability. However, the study found that setting a threshold for the bearing-specific datasets was extremely challenging. For instance, when the package was applied to detect the risk in the data for Bearing 1, it was discovered that setting a threshold to control the stability of the mean and the standard deviation was extremely difficult. As an example, the study presents the *qcc* profile of the data for Bearing 1 as Figure 8.1. Owing to the challenges observed, the investigation of the package was discontinued. The study therefor proceeds with investigation package *brca*.

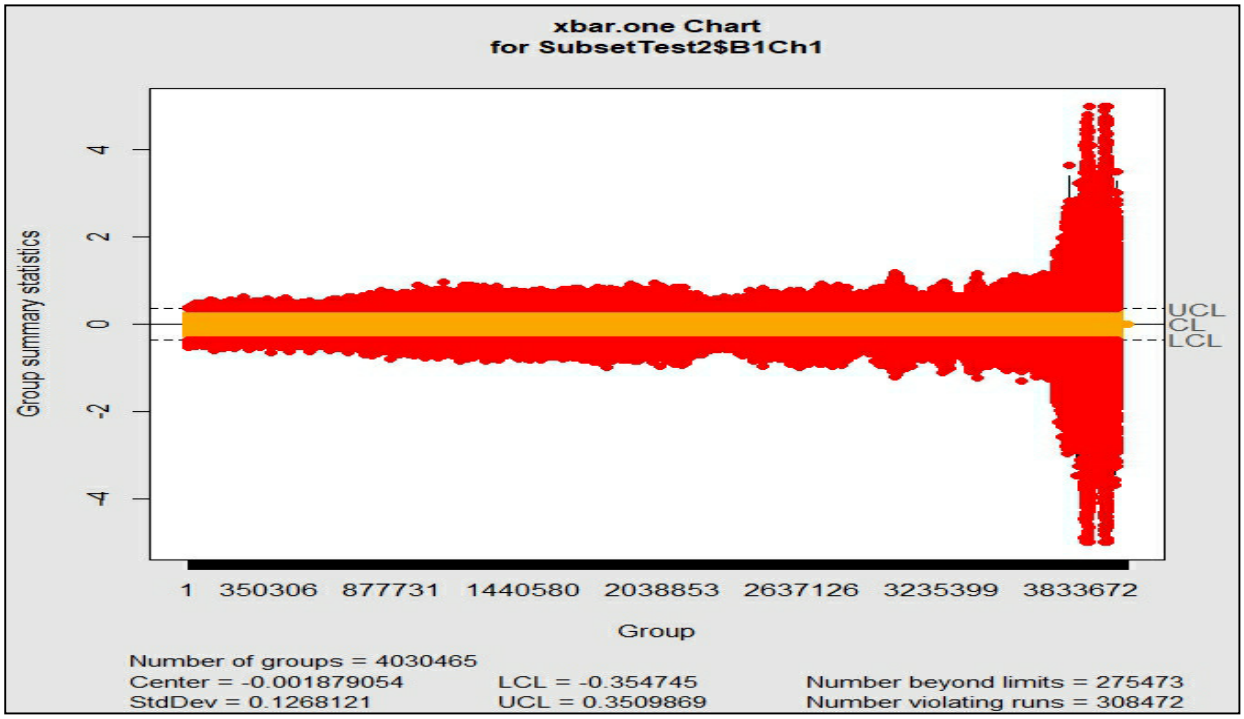


Figure 8.1: *qcc* profile for Bearing 1 in Training Dataset.

8.2. Investigating Package brca

During investigation of package *brca* with the dataset, the study found the package to be less effective for the study objectives. For instance, using the package for the data for Bearing 1, the plot obtained (Figure 8.2) reveals that the package is not fit for the purpose of the study. Further investigation of the package reveals that it has only been successful at tracking movement of animals and other objects. As a result, the use of *brca* for the research was discontinued. The study therefore proceeds with investigation of package *changepoint*.

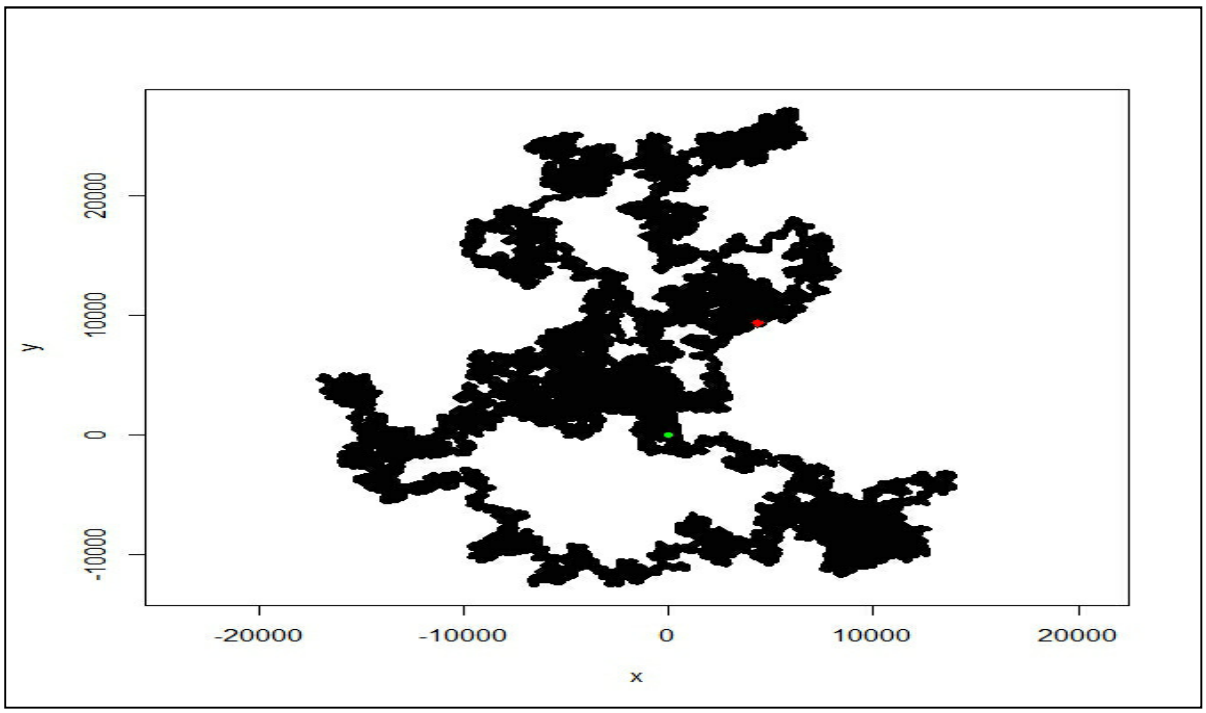


Figure 8.2: *brca* plot for Bearing 1 in training Dataset

8.3. Investigating Package Changepoint

The study found that the package has three algorithms namely the Pruned Exact Linear Time (*PELT*), Binary Segmentation (*BinSeg*) and Segment Neighbourhoods (*SegNeigh*) algorithms. A brief description of the algorithms in a previous research on the package (Killik & Eckley, 2014) has been adapted and provided in Table 8.3a.

Table 8.3a: Algorithms of R-package *changepoint* and their description

Algorithm	**Description**
BinSeg	Finds approximate change-point in data and computationally fast.
PELT	More accurate and computationally more efficient for change-point determination.
SegNeigh	Similar to the PELT algorithm but slower.

There are concerns that the '*BinSeg*' algorithm is less accurate and require high computation power compared with the PELT algorithm (Killik & Eckley 2014) hence was not investigated. Every attempt to use the '*SegNeigh*' algorithm produce an error with a warning message- "choose an alternative penalty; '*SegNeigh*' is computationally slow, use PELT instead" was observed. As a result, only the '*PELT*' algorithm was fully investigated and found to be successful. The study therefore selects package changepoint with the '*PELT*' algorithm as one of the big data techniques for the QRA method. The study then proceeds to investigate package *strucchange*.

8.4. Investigating Package Strucchange

The study found that package *strucchange* has two other algorithms namely *F-statistic*, which determines the number of boundary crossings in the data, and *SupF-statistic*, which provides statistical p-values for the significance of the crossing. The study therefore selects all three algorithms for further investigation with the training dataset. The study also selects R-package *PerformanceAnalytics* (PA) which provides an overview of the various metrics from the list of observations in the data with their significance (Peterson et al., 2018). The PA plot was preferred because it displays (a) the bivariate scatter plots with a fitted line, (b) the value of the correlation with their significance levels.

Bearing fundamental frequencies may depend on bearing geometry and rotor speed (Shah & Patel, 2014). Owing to this, investigation of fundamental train frequency (FTF) of the bearings was initially considered for investigation of interaction effect. However, it was discovered that some research has highlight normal bearing vibration to consists of a combination of separate independent effects which causes the probability density to approach a Gaussian curve (Patel & Giri, 2017). And because the study aims at detecting risk events within process operations the three features BPFO, BPFI and BSF which represent the defect of bearings must be applied.

8.5. Conclusion

In this chapter the study has provided explanation on how the packages on the R platforms were selected for testing on the training dataset as part of the method investigation. Two change-point packages, package *changepoint* and *strucchange* were selected for the method. For package *changepoint*, the study selected the '*PELT*' algorithm while package strucchange with its two associated algorithms were selected. Next is Chapter 9, where the performance of the packages will be tested with the training dataset. If successful, the packages will be used for all data analysis in this study.

Chapter 9 – Testing Software Packages

9.0. Introduction

In Chapter 8, the study selected two change-point packages *changepoint* and *strucchange* were selected for the proposed method. In this chapter, the study applies the '*PELT*' algorithm in package *changepoint and* package *strucchange* with its associated algorithms on the training as further investigation into obtaining a suitable method for data analysis for the research. Any unsuccessful performance on the training dataset by a package or algorithm provides a justification for rejection and will not be applied as part of the method. Because the notes which accompanied the dataset highlights outer race failure (BPFO) of Bearing 1, the study investigate the performance of the packages and associated algorithms at detecting the risk of failure BPFO using all four bearing-specific datasets.

Testing the packages and algorithms on all four bearing-specific datasets will help the study to ascertain whether the packages can detect risk of failure without any bias. If any of the packages and algorithms detects a risk of failure in any bearing apart from Bearing 1, or any type of failure other than failure BPFO, performance bias will be suspected. The study will therefore reject the package or algorithm because it is not fit for the purpose for its application as part of the method. The order by which the packages were applied to the data was selected at random. First package applied to the bearing-specific-datasets was the change-point package *changepoint using the PELT algorithm*.

9.1. Testing Package changepoint

The '*PELT*' algorithm of package *changepoint* was applied to the data to determine the change-points by changes in the mean and variance. The outcome of the change-point by changes in the mean (Figure 9.1a) reveals that high number of change-points. For instance, 149 risks of failure (change-points) were detected in the data for Bearing 1, 637 risks were detected in Bearing 2, and 733 risks were detected in Bearing 3 and also in Bearing 4. However, the visual plots showing the position of the change-points (Figure 9.1b), suggests that only one risk of failure was determined in each of the bearings around the time indices 700 for Bearing 1 and 550 for Bearing 2, Bearing 3 and Bearing 4.

```
Changepoint Results

Bearing 1
Mean
cpts(mvalue1):  149 change points
Variance
cpts(vvalue1):  21 Changepoints
 logLik(B1.pelt)
      -2*logLik -2*Loglike+pen
       6870.681        6882.933
Bearing 2
Mean
cpts(mvalue2):  637 Changepoints
cpts(vvalue2):  72 Changepoints
logLik(B2.pelt)
      -2*logLik -2*Loglike+pen
       17441.65        17454.03

Bearing 3
Mean
cpts(mvalue3):  733 Changepoints
Variance
cpts(vvalue3):  47 changepoints
logLik(B3.pelt)
      -2*logLik -2*Loglike+pen
       17341.27        17353.65

Bearing 4
Mean
cpts(mvalue4): 733 Changepoints
Variance
cpts(vvalue4):  51 Changepoints
 logLik(B4.pelt)
      -2*logLik -2*Loglike+pen
       17401.58        17413.97
```

Figure 9.1a: Number of risks detected by package *changepoint* (by changes in the mean) in Training Dataset.

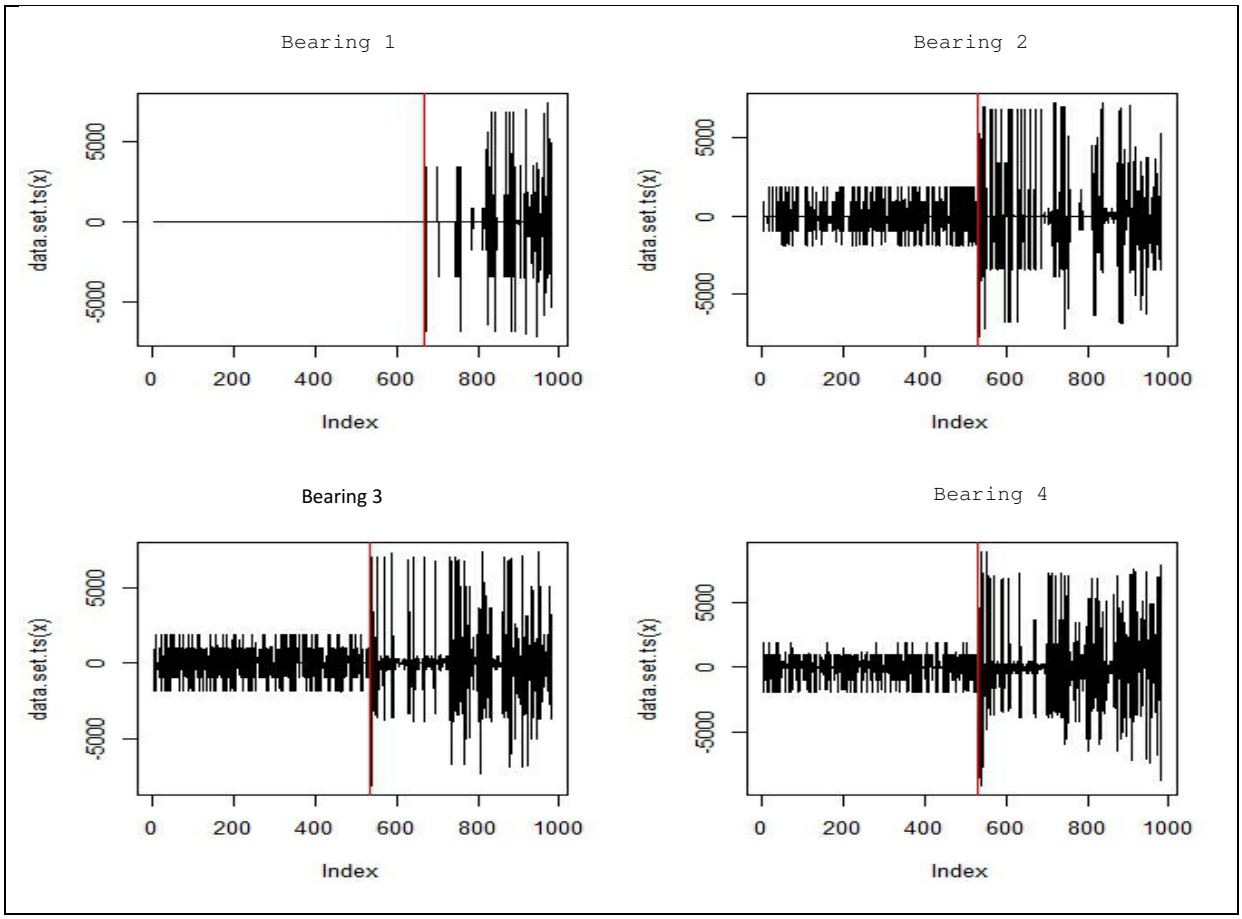


Figure 9.1b: Risks detected by package *changepoint* (by change-point by changes in mean) in Training Dataset.

However, the outcome of the investigation of change-point by changes in the variance gave a visual plot showing one risk of failure detected at time index 968 in the vibrations of Bearing 1 but no risk detected in the vibrations of Bearing 2, Bearing 3 and Bearing 4 (Figure 9.1c). Since each time index is represented by a test file, the corresponding test file was found to be the file with the time stamp “2004.02.19.03.42.39”.

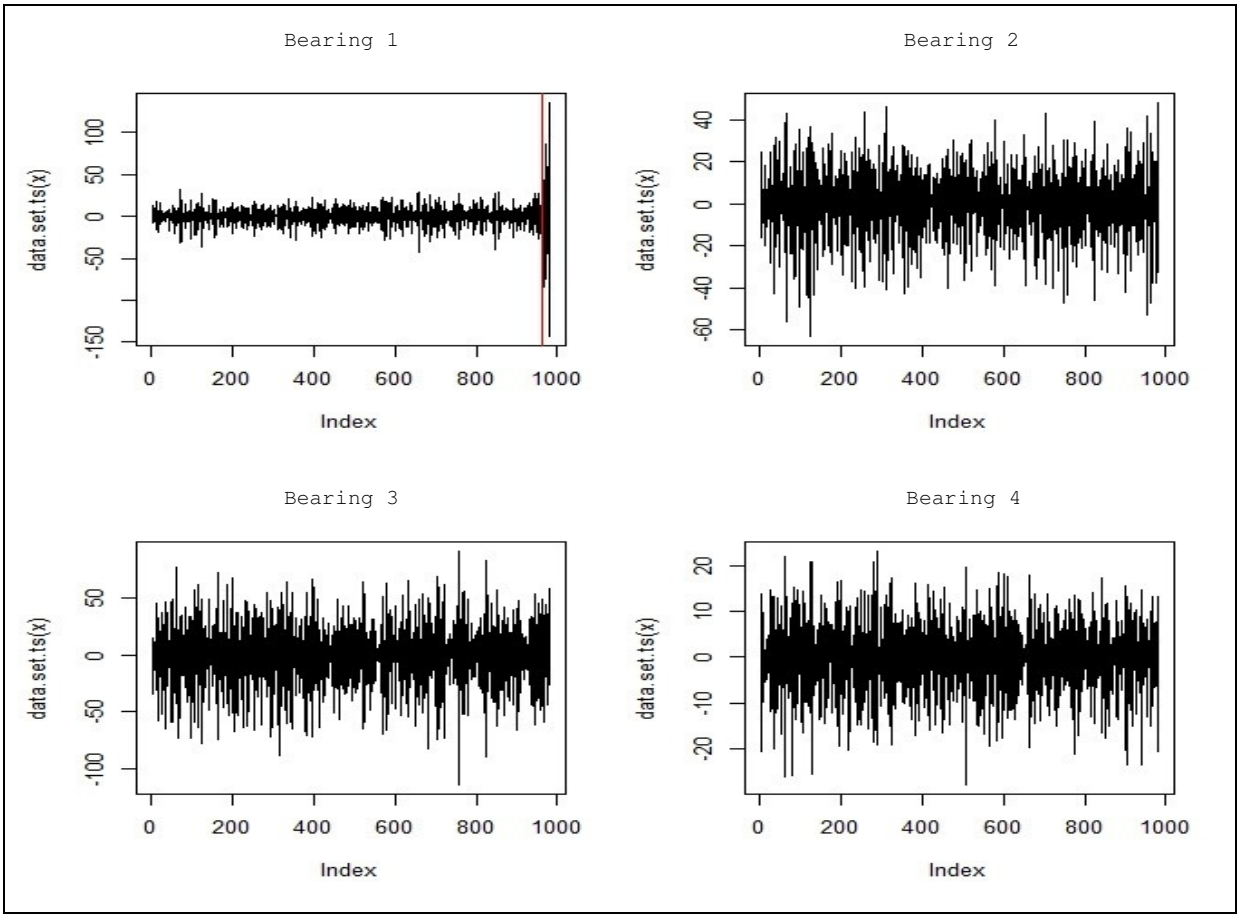


Figure 9.1c: Risks detected by change-point by changes in variance in Training Dataset.

After a careful thought of the outcome of the risks detected, it was suspected that high number of risk of failures detected by the change-point by changes in the mean could have been influence by the presence of noise. Besides because risks were detected in the vibrations of Bearings 2, 3 and 4, suspicion bias was suspected. As a result, the study rejects the application of change-point by changes in the mean as part of the method hence only change-point by the changes in the variance will be applied. The study proceeds to test package *strucchange* on the training data.

9.2. Testing Package Structucchange

The visual plot of the outcome of testing package *strucchange* detected one risk of failure (breakpoint) at time index 837 in the vibrations of Bearing 1, but no risks detected in the vibrations of the other three bearings. This corresponds with the visual plot of Figure 9.2a

which shows the time index of the risk as between 800 and 850. The corresponding test file was found to be the file with time index 2004.02.18.05.52.39. The study therefore proceeds to investigate '*F-Statistics*, which is an algorithm associated with the *Strucchange* package.

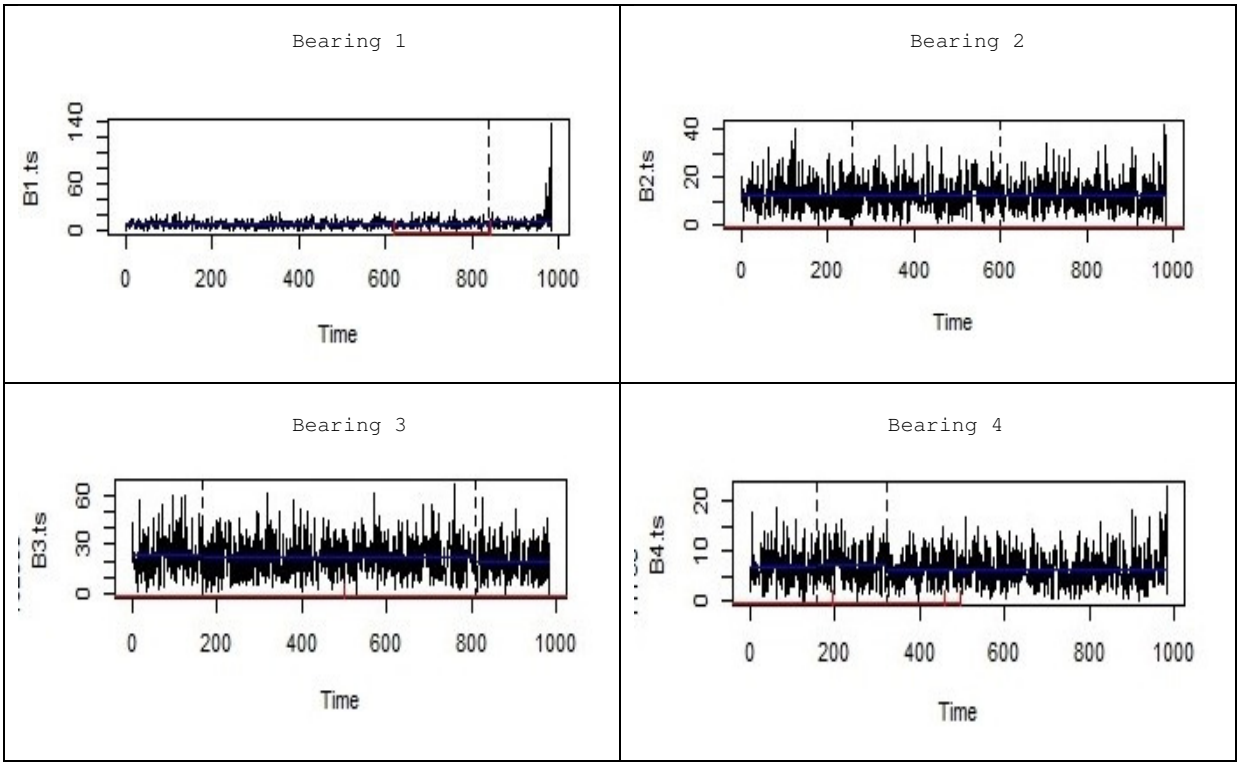


Figure 9.2a: Risks detected by package *strucchange* in Training Dataset.

When the '*F-Statistics'* was applied to the dataset, the visual plot obtained (Figure 9.2b) detects one risk (i.e. boundary crossing) in the data of Bearing 1, but no risks in the data of Bearing 2, Bearing 3, and Bearing 4 with a statistically significant '*Sup.F test'* (Figure 9.2c). However, further information from various published sources including Zeileis et al (2002) suggests that the '*F-statistics'* is sensitive to no more than one risk event (change-point) in a data and the '*Sup.F test*' is also influence by large size of N (Zeileis et al, 2002). As a result, the study suspends any further use of '*F-Statistics'* and '*Sup.F test'* for the reserach.

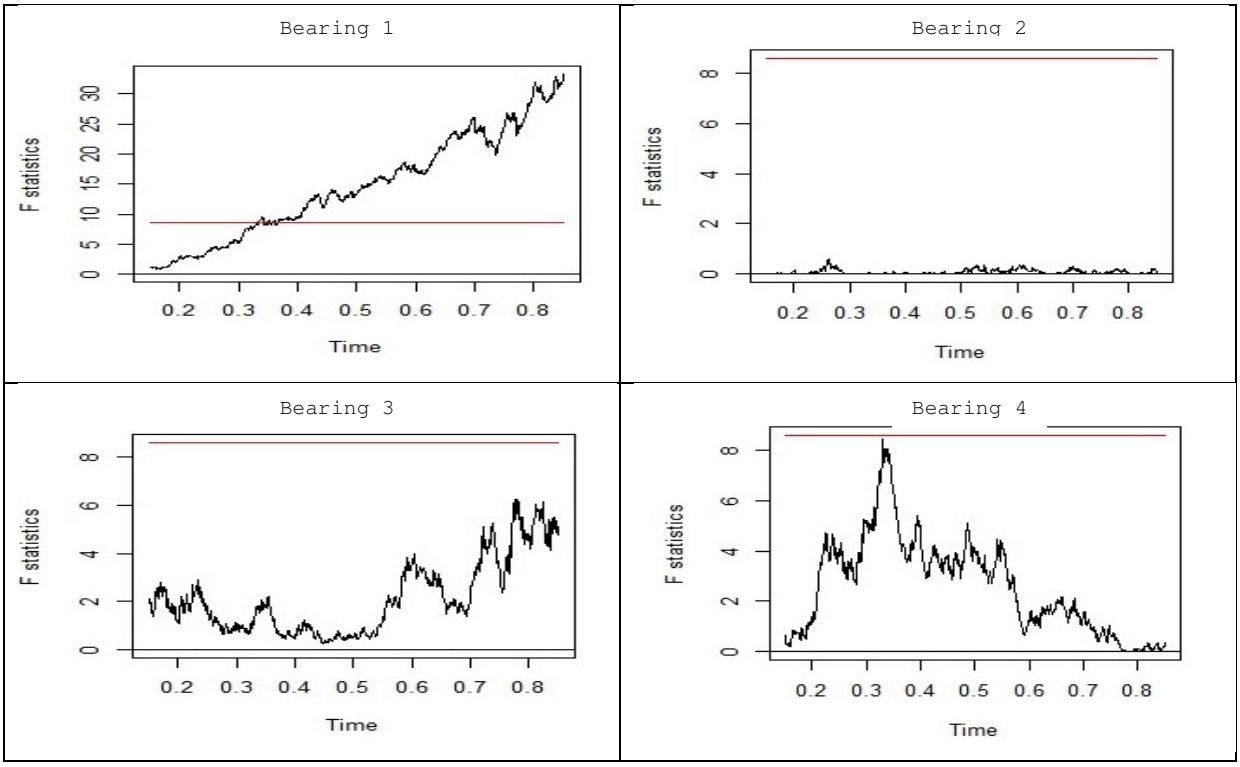


Figure 9.2b: Risks detected by *F-statistics*' in Training Dataset

```
Significant Test p-value Results

Bearing 1
sup.F = 33.332, p-value = 3.266e-07

Bearing 2
sup.F = 0.5785, p-value = 1

Bearing 3
sup.F = 6.2639, p-value = 0.1436

Bearing 4
sup.F = 8.4392, p-value = 0.05415
```

Figure 9.2c: Outcome from applying '*Sup.F'* test to training dataset

Comparing the time-domain features of the test files representing the time indices of the risk events detected, the output (Table 9.2a) it was noted that the risk event associated with BPFO and BPFI at the time index of the detection by package *changepoint* were comparatively higher than those detected by package *strucchange*. However, the risk associated with BSF shows an opposite trend. This led to a suspicion that the three type of risks may have some association. Because only one dataset was being used at this stage of the research, the study could not provide any valid explanation for this suspected associations. As a result, the study will investigate any association between the risks as part of the research since this was not found in any of the articles found to have applied any of the datasets being used for the research.

Table 9.2a: Time domain features of files at the at the Change-point

Test Files	Change-point	FTF	BPFI	BPFO	BSF
2004.02.18.05.52.39	837	9.802	12.170	7.663	8.218
2004.02.19.03.42.39	968	21.839	16.614	14.683	5.461

The study compared the period of risks detected by the packages with a plot of the root mean square (RMS) of the vibrations for the entire lifecycle of all three bearings, with more for comparison purposes with more emphasis on the plot for Bearing 1 since it is the bearing which suffered from the risk. The RMS and kurtosis has both been described as more suitable indicators for bearing degradation (Li, Li and Yu, 2019). The study therefore plots the RMS of all four bearings (Figure 9.2d, Appendix page 265) but presents the RMS of Bearing 1 which is discussed below.

The plot of the RMS of the vibrations of the bearing over its lifecycle of Bearing (Figure 9.2e) shows a change in the trends around the time index 700, where there appears a sharp rise in the data with no underlying trend, followed by a gradual drop in the trend, then a second rise just below the time index 850 which is close to the time index of the risk detected by package *strucchange*. Thus, the risk detected by package *strucchange* is the lower threshold of the risk event (i.e. onset of acceptable risk) suffered by component Bearing 1 in the process system.

Another change in the trend occur from around time index 950 which is also close to the time index of the risk detected by package *changepoint* and continue to increase to the end of the life of the bearing. This is the highest threshold of the risk (i.e. the main risk) suffered by the component bearing. Any risk events between the period between the time index of the risks detected by the two packages refers to the period of acceptable risk. Thus, the two packages provide a good detection of the risk of failure within the process system. The study proceeds to investigate the relationships being exhibited by the system components at the time indices of the risk event using interaction effect.

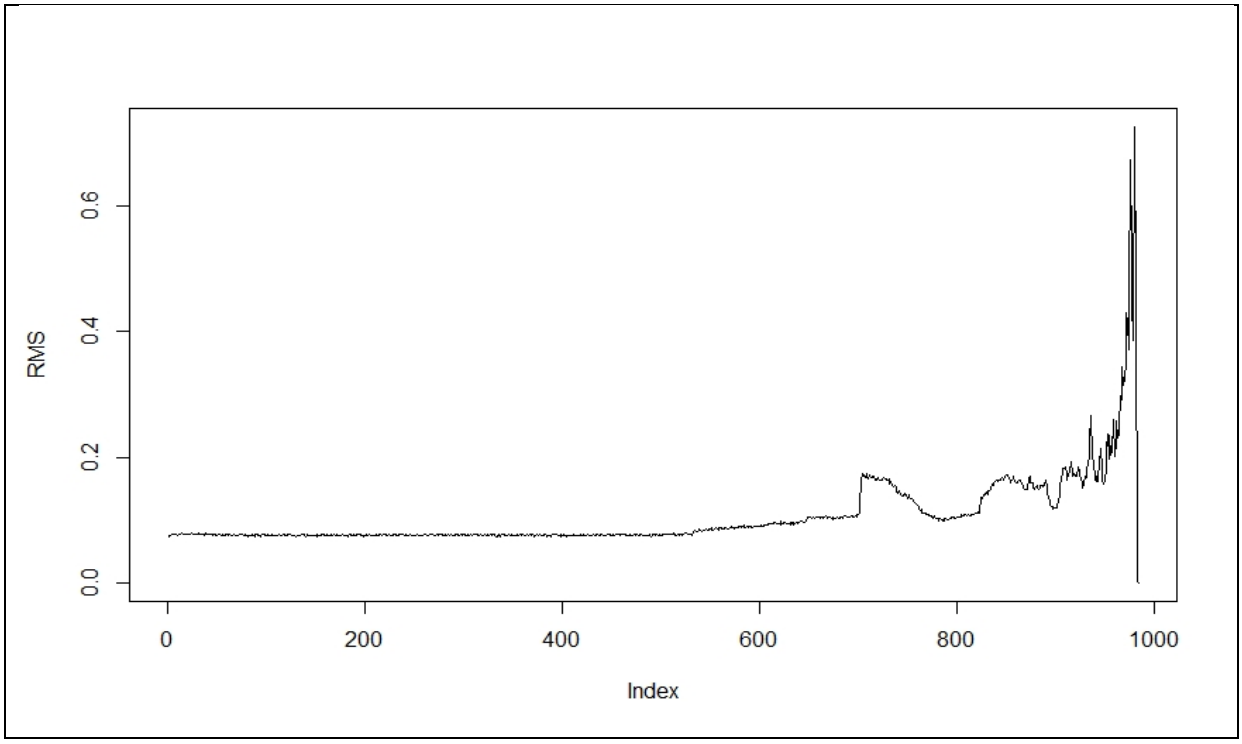


Figure 9.2e: Plot of RMS of the vibration of the lifecycle of Bearing 1 of Training Dataset.

9.3. Interaction Effect

Investigating any communication between system components through interaction up to the time index of the risks detected , the study initially investigates relationships through correlation between the component bearing which suffered from the risk (Bearing 1) and the other component bearings within the process system using correlation plots and Pearson's correlation tests.

9.3.1. Investigating Relationships

The relationships between the component bearings were investigated using package *PerformanceAnalytics*. The plots obtained (Figure 9.3a) reveals that for the interactions up to time index of the risks detected, the distribution of the vibrations of all four bearings are positive skewed. There is a significant positive Pearson correlation between Bearing 1 and Bearing 2; Bearing 1 and Bearing 3; and Bearing 2 and Bearing 3. The plots show other correlations which appear significant due to the large size of N but not because of any relationship between the component bearings.

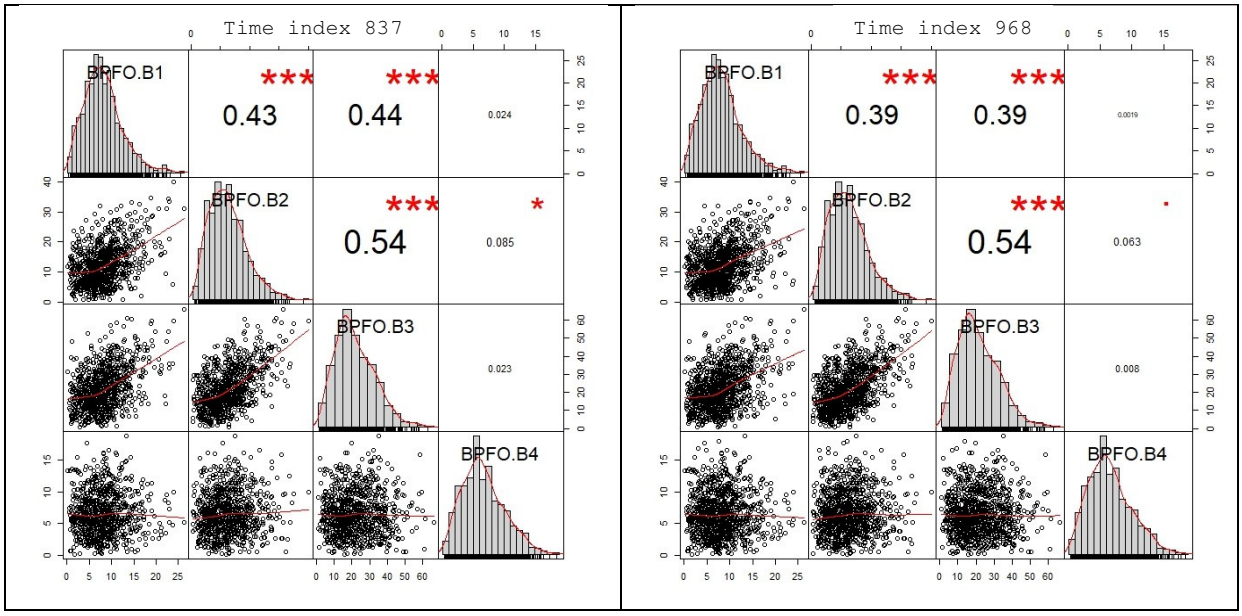


Figure 9.3a: Plots of correlation between bearings for interactions up to time indices of the risk detected in Training Dataset.

The observed significant correlation between Bearing 1 and the other bearings could be due to their contribution to the risk suffered by the bearing. These contributions are usually made through some communication between the component bearings within the process system (i.e. systems exhibiting system complexity) interacting up to the time index of the risk event. As a result, the study proceeds to investigate the significant correlations (i.e. the suspected contribution from the other component bearings to the risk suffered by Bearing 1) using decision tree modelling.

9.3.2. Applying Decision Tree Model

As explained in Section 9.3.1., the decision tree modelling will be used to determine any interactions between the component bearings with the bearing which suffered from the risk up to the time index of the risk event. It will also be used to determine the predictor and moderators for the regression models. Applying the decision tree model to the data, the outcome obtained (Figure 9.3b) reveals that for the interactions up to time index 837 and 968 of the risk events, Bearing 3 appears to be the main component whose vibrations affects Bearing 1, with the vibrations of Bearing 2 being important at both high and low vibrations of Bearing 3 but Bearing 4 shows no effect. Thus, there is evidence of some contribution from Bearing 2 and Bearing 3 to the risk suffered by Bearing 1 which require further investigation.

To achieve this the study employs regression modelling using the most dominant feature from the decision tree model (Bearing 3) as the predictor and the least dominant feature (Bearing 2) as the moderator.

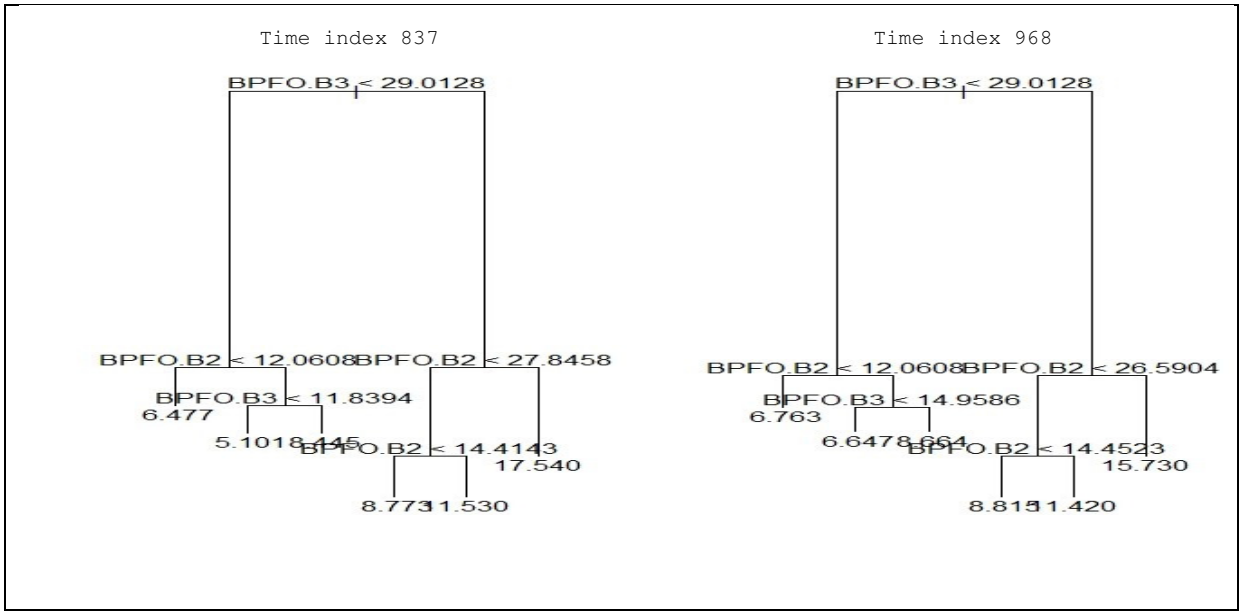


Figure 9.3c: Decision tree for regression models for interactions up to time index of risks detected in Training Dataset.

9.3.3. Applying Regression Model

The applies regression model to investigate the interaction effect up to the point of the risk detected, using the predictors and moderators determined from the decision tree model. The summary output of the regression models obtained from applying package *stargazer* (Table 9.3a) reveals that:

- When the other bearings are not in operation, the average vibration of Bearing 1 up to the time index 837 is 3.48 Hz which increase by 0.174 Hz or 0.112 Hz per unit increase in the vibration of Bearing 2 or Bearing 3 and these are statistically significant (model 1). For the interaction model (model 2), the vibration of Bearing 1 is 5.61 Hz when the other bearings are not in operation. This increases by 0.002 Hz per unit increase in the vibrations of Bearing 2 due to its moderation in the effect of the vibrations of Bearing 3 on that of Bearing 1, which is not significant. However, there is a significant Bearing 2-Bearing 3 interaction effect which causes the vibration of Bearing 1 to increase by 0.006 Hz per unit increase in Bearing2-Bearing 3 interaction vibration.
- When the other bearings are not operation, the vibration of Bearing 1 up to the time index 968 is 4.00 Hz which increase significantly by 0.165 Hz or 0.099 Hz per unit increase in the vibrations of Bearing 2 or Bearing 3 (model 3). For the

interaction model (model 4), the vibration of Bearing 1 is 6.12 Hz when the other bearings are not in operation. This decreases per unit increase in the moderation by the vibrations of Bearing 2 which is not significant. However, there is a significant Bearing 2- Bearing 3 interaction effect which causes the vibration of Bearing 1 to increase by 0.006 Hz per unit increase in the Bearing 2 – Bearing 3 interaction vibration.

Table 9.3a: Output of regression for the of risks detected in Training Dataset

	Dependent variable:			
	BPFO.B1			
	model.1	model.2	model.3	model.4
	(1)	(2)	(3)	(4)
Constant	3.475***	5.611***	4.004***	6.118***
	(0.314)	(0.569)	(0.309)	(0.559)
BPFO.B3	0.112***	0.021	0.099***	0.008
	(0.013)	(0.024)	(0.013)	(0.024)
BPFO.B2	0.174***	0.002	0.165***	-0.008
	(0.023)	(0.045)	(0.023)	(0.045)
BPFO.B3:BPFO.B2		0.006***		0.006***
		(0.001)		(0.001)
Observations	837	837	968	968
R2	0.245	0.263	0.192	0.209
Adjusted R2	0.243	0.260	0.191	0.207
Residual Std. Error	3.807 (df = 834)	3.764 (df = 833)	4.059 (df = 965)	4.018 (df = 964)
F Statistic	135.164*** (df = 2; 834)	98.846*** (df = 3; 833)	114.926*** (df = 2; 965)	84.953*** (df = 3; 964)
Note:				*p<0.1; **p<0.05; ***p<0.01

The significant interactions are confirmed by the ANOVA type II test (Table 9.3b) which reveals that the bearing components exhibits system complexity by interactions up to the time indes of the risks detected. Thus, there is a communication between the three bearings and therefore evidence of contributions from Bearing 2 and Bearing 3 to the risk suffered by Bearing 1. The main contributor to the risk suffered by Bearing 1 is Bearing 3. However, the contributions of Bearing 3 to the risk suffered by Bearing 1 is somehow aided by some contributions from Bearing 2 through a Bearing 2-Bearing 3 interaction effect. The study therefore proceeds with further investigation of the contribution from the Bearing 2 and Bearing 3 to the risk suffered by Bearing 1 using effect plots, Johnson-Neyman plots and Simple-slope analysis.

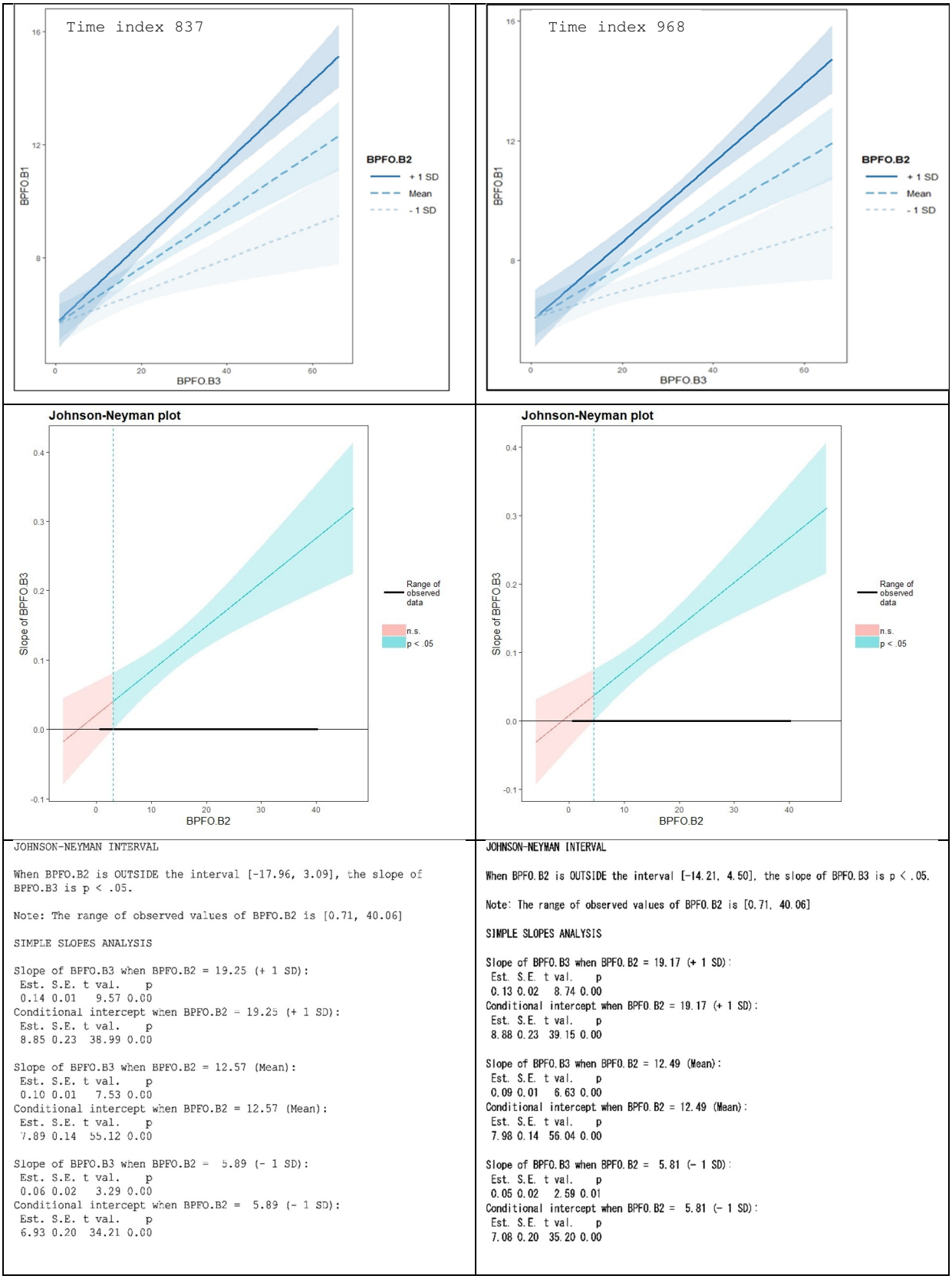
Table 9.3b: ANOVA table for the interactions up to time index of the risks detected in Training Dataset

```
Analysis of Variance Table

Model 2: BPFO.B1 ~ BPFO.B3 * BPFO.B2
Model 4: BPFO.B1 ~ BPFO.B3 * BPFO.B2
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
2    833 11803  1    283.91 20.038 8.65e-06 ***
4    964 15567  1    329.25 20.39 7.098e-06 ***
---
Signif. codes:  0 ・**・0.001 ・*・0.01 ・・0.05 ・・0.1 ・・1
```

9.3.4. Further Investigation into Significant Interactions

Figure 9.3d represents the effect plots, J-N plots and simple-slope analysis for the interactions up to the time index of the risks detected which shows that the models fit better at the 95% confidence interval when the moderation of the vibration of Bearing 2 approaches the one standard deviation above the mean than the model of the mean and one standard deviation below the mean.




```
JOHNSON-NEYMAN INTERVAL

When BPFO.B2 is OUTSIDE the interval [-17.96, 3.09], the slope of
BPFO.B3 is p < .05.

Note: The range of observed values of BPFO.B2 is [0.71, 40.06]

SIMPLE SLOPES ANALYSIS

Slope of BPFO.B3 when BPFO.B2 = 19.25 (+ 1 SD):
 Est. S.E. t val.    p
 0.14 0.01   9.57 0.00
Conditional intercept when BPFO.B2 = 19.25 (+ 1 SD):
 Est. S.E. t val.    p
 8.85 0.23  38.99 0.00

Slope of BPFO.B3 when BPFO.B2 = 12.57 (Mean):
 Est. S.E. t val.    p
 0.10 0.01   7.53 0.00
Conditional intercept when BPFO.B2 = 12.57 (Mean):
 Est. S.E. t val.    p
 7.89 0.14  55.12 0.00

Slope of BPFO.B3 when BPFO.B2 =  5.89 (- 1 SD):
 Est. S.E. t val.    p
 0.06 0.02   3.29 0.00
Conditional intercept when BPFO.B2 =  5.89 (- 1 SD):
 Est. S.E. t val.    p
 6.93 0.20  34.21 0.00
```

```
JOHNSON-NEYMAN INTERVAL

When BPFO.B2 is OUTSIDE the interval [-14.21, 4.50], the slope of BPFO.B3 is p < .05.

Note: The range of observed values of BPFO.B2 is [0.71, 40.06]

SIMPLE SLOPES ANALYSIS

Slope of BPFO.B3 when BPFO.B2 = 19.17 (+ 1 SD):
 Est. S.E. t val.    p
 0.13 0.02   8.74 0.00
Conditional intercept when BPFO.B2 = 19.17 (+ 1 SD):
 Est. S.E. t val.    p
 8.88 0.23  39.15 0.00

Slope of BPFO.B3 when BPFO.B2 = 12.49 (Mean):
 Est. S.E. t val.    p
 0.09 0.01   6.63 0.00
Conditional intercept when BPFO.B2 = 12.49 (Mean):
 Est. S.E. t val.    p
 7.98 0.14  56.04 0.00

Slope of BPFO.B3 when BPFO.B2 =  5.81 (- 1 SD):
 Est. S.E. t val.    p
 0.05 0.02   2.59 0.01
Conditional intercept when BPFO.B2 =  5.81 (- 1 SD):
 Est. S.E. t val.    p
 7.08 0.20  35.20 0.00
```

Figure 9.3d: Effect plots, J-N plots and simple-slope analysis data for interactions up to time index of the risks detected in Training Dataset.

The J-N plots reveals that the slope of Bearing 3 is statistically significant when the range values of vibration of Bearing 2 is greater than 3.09 Hz for the interaction up to time index

837 or 4.50 Hz for the interactions up to time index 968. Thus the vibrations of Bearing 2 has no effect on the contributions from the vibrations of Bearing 3 to the vibrations of the risk associated with risk BPFO suffered by Bearing 1 except when the vibrations of Bearing 2 is higher 3.09 Hz or 4.50 Hz at the time index 837 or 968 respectively.

The simple slope analysis reveals that the changes in the conditional slope of the data from the operations of Bearing 3 increases by 0.14 Hz per unit increase in the vibrations of Bearing 2. The conditional intercept also reveals that while the slope associated with the data of the vibrations of Bearing 3 when the vibrations of Bearing 2 increase, the conditional intercept also increases when the vibrations of Bearing 2 increases. This suggests that increases in the vibrations of Bearing 3 for high vibrations of Bearing 2 will turn towards being equal to the vibrations of Bearing 1. Hence, there is evidence of the components exhibiting system complexity for the interactions up to time index risks detected.

The study therefore concludes that there is evidence of system interactions between the component bearings for the interactions up to the time index of the risks detected. Thus, the risks detected are not just due to the independent operations of the individual bearing which suffered the risk but also the contribution of other bearing components within the process operation system. The study proceeds to investigate any association between the bearing features up to the time index of the risks.

9.4. Investigating Association between Feature Frequencies

When the decision tree was applied to investigate any interactions between the frequencies associated with the three type of risks up to the time index of the risks detected, single node trees were obtained. It was therefore concluded that there is no association between the feature through interactions up to the time index of the two risk events. As a result, no further investigation was conducted by the study.

9.5. The Big Data QRA Method

In contrast to the existing QRA methods, the propose big data procedure in this study uses an entirely big data approach and process operation data for a reliable QRA in HHPIs. As per the definition of the components of risk, both the frequency and consequences associated with a selected accident scenario must be are considered using the most advanced techniques (Kim, Sohn & Paik, 2017). It may be recall that the accident scenario, the associated consequences and frequency has already been established from the final

incident investigation report of the dust fire and explosion incident used as case history under Chapter 4- Real-life Case Histories of Part 2 of this thesis.

Figure 9.5 represents the suggested procedure for QRA for the management of catastrophic events from risks at the HHPIs. Generally, a catastrophic event can occur when a component of the process fails or there is an issue of abnormality within parts of the process system. The failure or abnormality could be caused independently or by a remote event including communication of the various components through interactions in the form of a loop. Thus, type of incident and consequences must be obtained from historical data (incident records and reports) on the process plant under investigation or a similar plant in special situation (e.g. plant is new), prior to the QRA.

The propose procedure only adopts available data and can be divided into 11 steps defined as follows:

a. Selection of the type of process operation.
b. Obtain real-time or historical data of the process operation.
c. Investigate data validity.
d. Explore data.
e. Apply feature extraction to condense the data if necessary.
f. Detect risk using change-point analysis packages *changepoint* and *strucchange*.
g. Investigate relationship using correlation plots and Pearson's correlation test.
h. Fit a regression models to investigate interaction effect.
i. Investigate significant interactions with effect plots, Johnson-Neyman plots and simple slop analysis.
j. Risk calculation for known incident frequency and consequences.
k. Decision making.

In this research, the incident frequency and associated consequences were obtained from the incident investigation reports. The risks and their corresponding time index are detected by the procedure using the process operation dataset. Decision making which is usually conducted based on the low threshold of risk (detected by package *strucchange*) and the main risk (detected by package *changepoint*) or using acceptance criteria based on 'as low as reasonably practicable (ALARP).

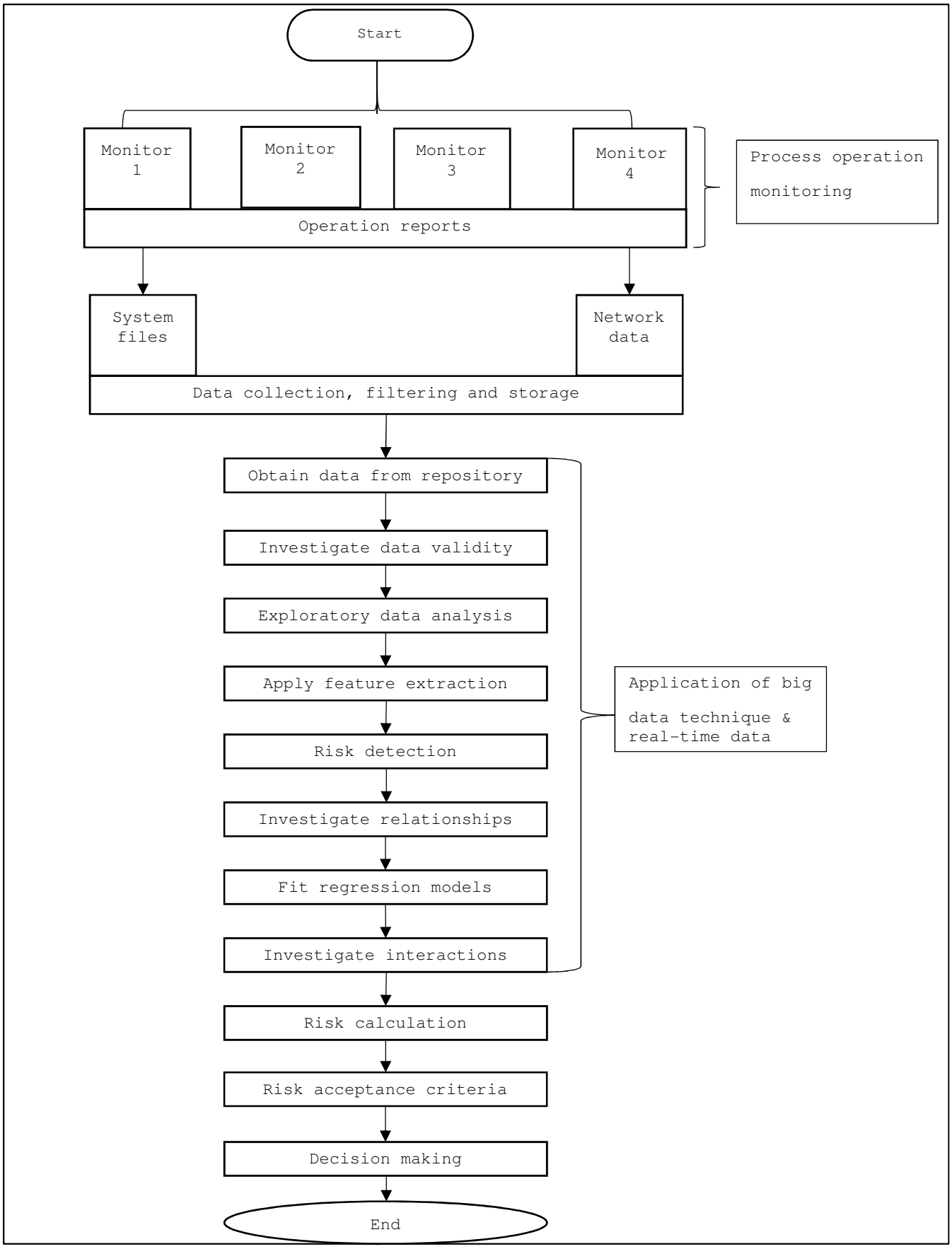


Figure 9.5: Procedure for QRA using big data techniques and real-time data

9.6. Detail Step-by-step Procedure of the Big Data QRA Method

9.6.1. Obtain information about the process for which the QRA is required.

9.6.2. Obtain the datasets relating to the operation of the process.

9.6.3. Investigate data quality and attributes by checking if total number of datafiles in the datasets corresponds to the number of files detailed in the available information of the data.

9.6.4. Investigate other attribute such as sampling rate and sampling times using no less than 10% of the files in the datasets selected at random.

9.6.5. Obtain and review summary and descriptive statistics of each of the selected data files to check if there are missing observations, nulls and N/As.

9.6.7. Where missing observations and N/As are observed apply appropriate techniques (e.g. N/A by approximation, N/A by mean) to fill out these missing observations as detail in the Step 9.6.8 below few steps.

9.6.8. Use the flowing approach to fill out the missing observations:

- Fill the missing observation with the last observation and check with a plot.
- Apply linear interpolation to fill in missing observation then check plot.
- Apply polynomial interpolation to fill the missing observation and then check the plot.
- Compare the plots for best fit and use it to fill the missing observation.

9.6.9. Investigate the magnitude of the mean, max and min values of the observations.

9.6.10. Explore the distribution of the data in the data files with plots such as histogram, box plots, and quantile plots to help establish the underlying distribution of the data.

9.6.11. Use sequence and lag plots to investigate randomness of the observations, time effect of the observations, and any potential bias in the operation of the components.

9.6.12. Use statistical test e.g. Anderson-Darling multivariate normality test to further investigate normality of the data.

9.6.13. Use correlation plots and Pearson correlation test to investigate any relationships which may exist between the process components.

9.6.14. Use data condensing technique (e.g. feature extraction technique such as fast Fourier transform (FFT)) to reduce the size of the data files where necessary to obtain key statistical and time-domain features to create data files which can be handle by PC power and PC software memory when combined.

9.6.15. Because the file reduction techniques depend on features on the component, use appropriate technique (e.g. calculation methods) to obtain any feature whose information may be missing. For instance, the speed and contact angles of the component bearings investigated in this study were described in 'rpm' and 'degrees', which must be converted into Hz and radians using appropriate equations.

9.6.16. After obtaining the key statistical and time-domain features, format the data in a time sequence of the source files to create component-specific or component-channel-specific data files.

9.6.17. Timestamp the files according to the time on each test file in the dataset from which the key features were obtained.

9.6.18. Combine the component-specific or component-channel-specific data files into a new data frame.

9.6.19. Write the combined data files in appropriate format (e.g. 'csv') as a new data or data tables using appropriate function or approach (e.g. R language function '*write.table*').

9.6.20. Using descriptive statistics, inspect the new data file (combine data file) to ensure that there are no missing variables and N/A's, then saved for further use.

9.6.21. Use the approach provided in Step 9.6.8 to fill out any missing observation in the new data file.

9.6.22. Check the health state of the component using plots of time series of the data, and plot of statistical features like skewness or kurtosis. For instance, in this study which involves component bearings, high kurtosis values would indicate unhealthy bearing condition or issues with how the bearing is secured within the process system. Also, a very noisy observation in the time series plot from the onset of its

operation may indicate issues with the bearing including bearing suffering from defect or not securely installed in the process.

9.6.23. Perform change-point analysis using appropriate software packages (e.g. *strucchange* and *changepoint* on the R Language platform as in this study) to detect the lower threshold and upper threshold of the risks in the operation of the component.

9.6.24. Compare the time index of the risks detected by the two packages.

9.6.25. If the time index of the risks detected by the two packages appear similar and the time series plot shows signs of disturbance as describe in Step 9.6.22, conclude that the bearing has suffered some issues (defective bearing or bearing not properly secured).

9.6.26. Plot the RMS data of the entire lifecycle of the process component under investigation.

9.6.27. Compare the time index of the risks detected by the two packages in Step 9.6.23 to any changes in the trend of the RMS.

9.6.28. If remaining useful life (RUL) information is available, use that also to confirm the presence of risk within the system. Please note that RUL on its own cannot provide approximation of the time index of the risks detected by the two packages the times are approximately similar.

9.6.29. With the aid of correlation plots and Pearson correlation test investigate any relationships which may exist between any of the other components with the component which suffered the risk through interaction up to the time index of the risks detected. This can be achieved by applying package *PerformanceAnalytics*.

9.6.30. If no significant Pearson correlation is observed at Step 9.6.29, conclude that there is no significant contribution from the operation of the other components to the risk in the operation of the component which suffered the risk.

9.6.31. Do not perform any further investigation.

9.6.32. If a significant correlation is detected at Step 9.6.29, check the correlation plots to ascertain whether the significant correlation is due to a relationship between the operations of the component which suffered the risk and the other

components in the process but not because of size of the population of the data (N).

9.6.33. If the significant correlation at Step 9.6.29 is due to the size of N, conclude that there is no significant contribution from the operation of any of the other components to the risk detected in the operations of the component which suffered the risk.

9.6.34. Do not perform any further investigation.

9.6.35. If the significant correlation at Step 9.6.29 is not due to actual relationship existing between the operations of the component which suffered the risk and that of the other components but not the size of N, probe the significant correlation with decision tree models.

9.6.36. If the decision tree modelling results in no node, conclude that the is no significant contribution from the operation of the other components to the risk detected in the operations of the component which suffered the risk.

9.6.37. Do not perform any further investigation.

9.6.38. If the decision three modelling result in a single node tree (no branch tree), conclude that there is no significant contribution from the operation of any of the other components to the risk detected in the operation of the component which suffered the risk.

9.6.39. Do not perform any further investigation.

9.6.40. If the decision tree modelling result in a branch tree, there is a likelihood of significant contribution from the operation of the other components to the risk detected in the operation of the component which suffered the risk.

9.6.41. Investigate the contribution of the operation of the other components to the risk detected in the operation of the component which suffered the risk using regression tree modelling.

9.6.42. Use the most dominant featured component in the decision tree model as a predictor, the next dominant featured component as the first moderator, followed by the next moderator, up to the *n*th moderator for the regression modelling.

9.6.43. Use appropriate package e.g. package *stargazer* to produce a regression output to enable sided-by-side comparison of the regression of data which explains the extent of contribution from the operation the other components to the risk detected in the operation of the component which suffered the risk.

9.6.44. Carefully examine the regression output to ascertain if there are any significant moderation or interaction effect.

9.6.45. If there is no significant moderation or interaction effect in Step 9.6.44, conclude that there is no contribution from the operation of the other components to the risk detected in the operation of the component which suffered the risk.

9.6.46. Do not perform any further investigation.

9.6.47. If there a significant moderation or interaction effect in Step 9.6.44, conclude that there is a suspicion of contribution from the operation of the other component to the risk detected in the operation of the component which suffered the risk.

9.6.48. Investigate the significant moderation or interaction using Type II ANOV test.

9.6.49. If the Type II ANOVA reveal that the moderation or interaction is not significant, confirm that there is no significant contribution from the operation of the other components to the risk detected in the operation of the component which suffered the risk.

9.6.50. Do not perform any further investigation.

9.6.51. If the Type II ANOVA test reveal a significant component moderation effect or component-component interaction effect, conclude that there is a significant contribution from the moderation component or component-component interaction to the contribution from the operations of the predictor component to the risk detected in the operation of the component which suffered from the risk.

9.6.52. Probe the significant component moderation or component-component interaction effect using effect plots, Johnson-Neyman plots and simple-slope analysis to help provide clarity on the extent of contribution of the moderation or interactions.

9.6.53. Perform risk calculation using incident frequency and associated consequences data obtained from sources such as incident investigation reports, historical

records of incidents relating to the process or a process of similar characteristics to the process being asses.

9.6.54. Decisions can then be made using risk acceptance criteria based on the lower threshold of risk (detected by package strucchange) and the upper threshold of risk (detected by package changepoint).

The entire process is illustrated by the flowchart of Figure 9.6.

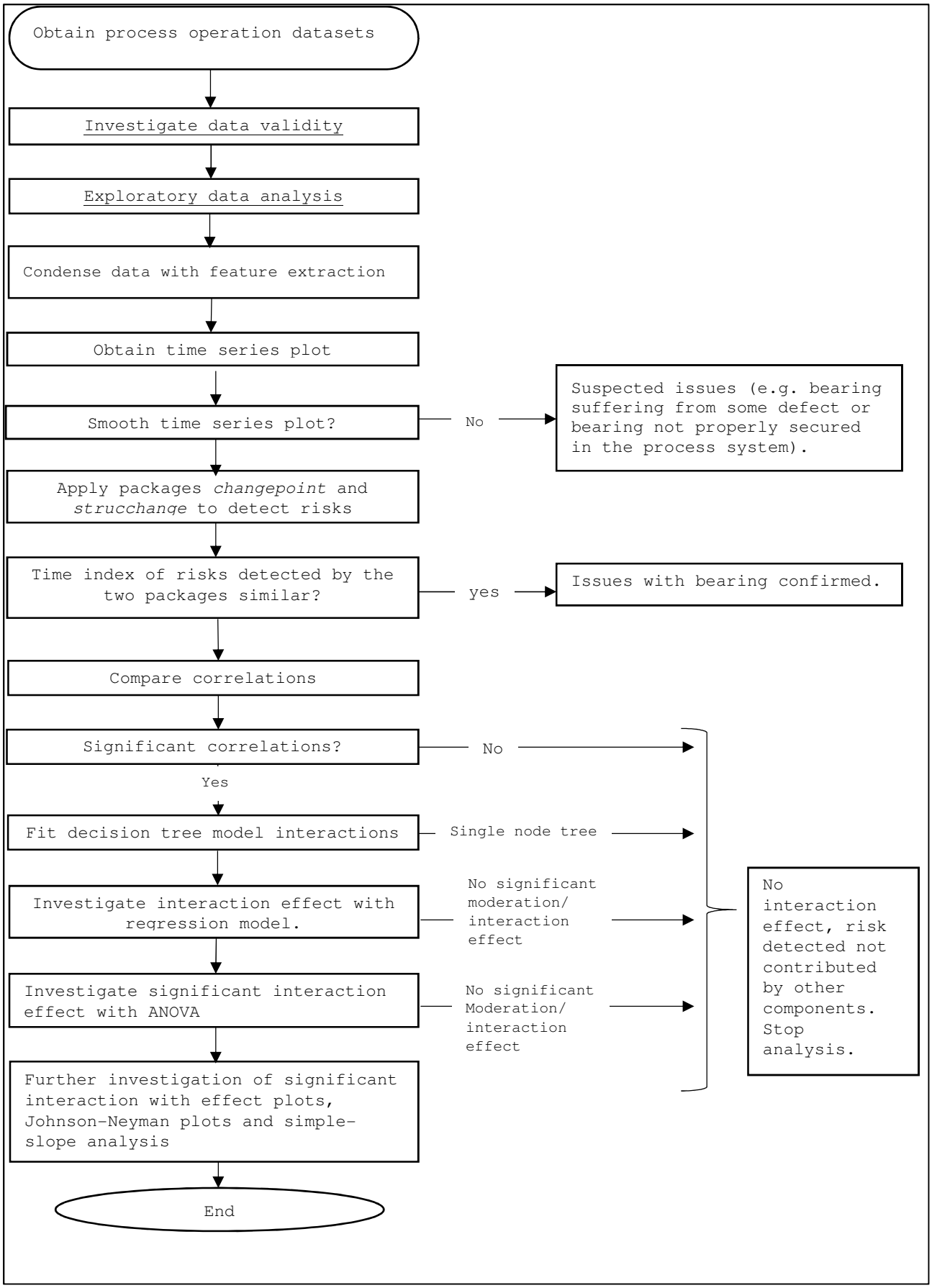


Figure 9.6: Flowchart for the detailed step-by-step big data QRA method procedure

9.7. Conclusion

This chapter has investigated big data techniques and software packages for the method. The study found change-point analysis as a successful big data technique to be applied as a QRA method for the research. The study found two change-point packages on the R platform were successful when applied together for risk detection within the process system. Of the two change-point packages, *strucchange* was found to be successful at detecting low threshold of risk within normal operations. Package *changepoint* using the '*PELT*' algorithm was found to be successful at detecting the high threshold for normal operations.

All signals between the two thresholds are considered as normal operational risk. Between the two thresholds, the process activities like maintenance services may be carried out to ensure that the risk do not get the high threshold. Any signal that gets to the region of the change-point detected by PELT could lead to a catastrophic event. Using the concept of effect of interactions between system components up to the period of the risk event, the study was able to detect the type of system being exhibited by the components at the time index of the risk. Next is Part 4 – Data Analysis, where the method obtained from the investigations in this chapter will be applied to four case datasets available for the research.

Part 4

Data Analysis

Part 4 - Background

In Part 3, the study investigates big data techniques and software packages were investigated after which the following techniques and packages were selected and used to obtain the big data QRA method.

- Fast Fourier transform (FFT) as feature extraction technique to condense the data for handling by PC and PC software memory.
- Change-point analysis using software packages *changepoint* and *strucchange* on the R language platform to detect the risk within the process.
- Correlation plots and Pearson's correlation tests using package *PerformanceAnalytics* to determine relationships between the component parts of the process system interacting up to the time index of the risk events detected.
- Decision tree modelling to investigate the contributions from other components to the component which suffer from the risk event and help establish the component acting as predictor to the risk event and the component acting as the moderator.
- Regression modelling using package *stargazer* to present the regression output for the interactions.
- ANOVA Type II test to investigate significant moderation and interactions observed in the regression modelling.
- Effect plots, Johnson-Neyman plots and Simple slope analysis to provide insight into the significant interactions in pictorial format.

The study establishes that a combination of two change-point packages, *strucchange* and *changepoint*, was successful at detecting of risk in process systems. Package *strucchange* was found to be successful at detecting the lower threshold of risk. Package *changepoint* was found to be successful at detecting the highest threshold of risk. After the risk detection, the study found interaction effect to be successful at determining the type of communication between the system components interacting up to the time index of the risk. The study then provides a detailed step-by-step procedure for the big data QRA method obtained to enable users and safety practitioners to successfully apply the method for risk analysis in the HHPI.

The study now applies the method to case study datasets as Part 4, using the step-by-step procedure detailed in Section 9.6. First, the method will be validated using two case study data sets to test its robustness and validity for QRA. The method will then be tested using two other case study datasets as applied examples of the methods reliability and a proof of the method being fit application for QRA within the HHPI.

Chapter 10 - Case Study Datasets

10.*0. Introduction*

In this chapter, the study introduces the four case study datasets which would be applied for the rest of the research. The study will apply the data exploration and feature extraction technique applied to the training dataset to reduce the datasets to sizes which can be handle by PC software memories as applied in Part 3, prior to analysis using the method. The overall framework from data analysis to arriving at the results is illustrated by the flowchart of Figure 10 below.

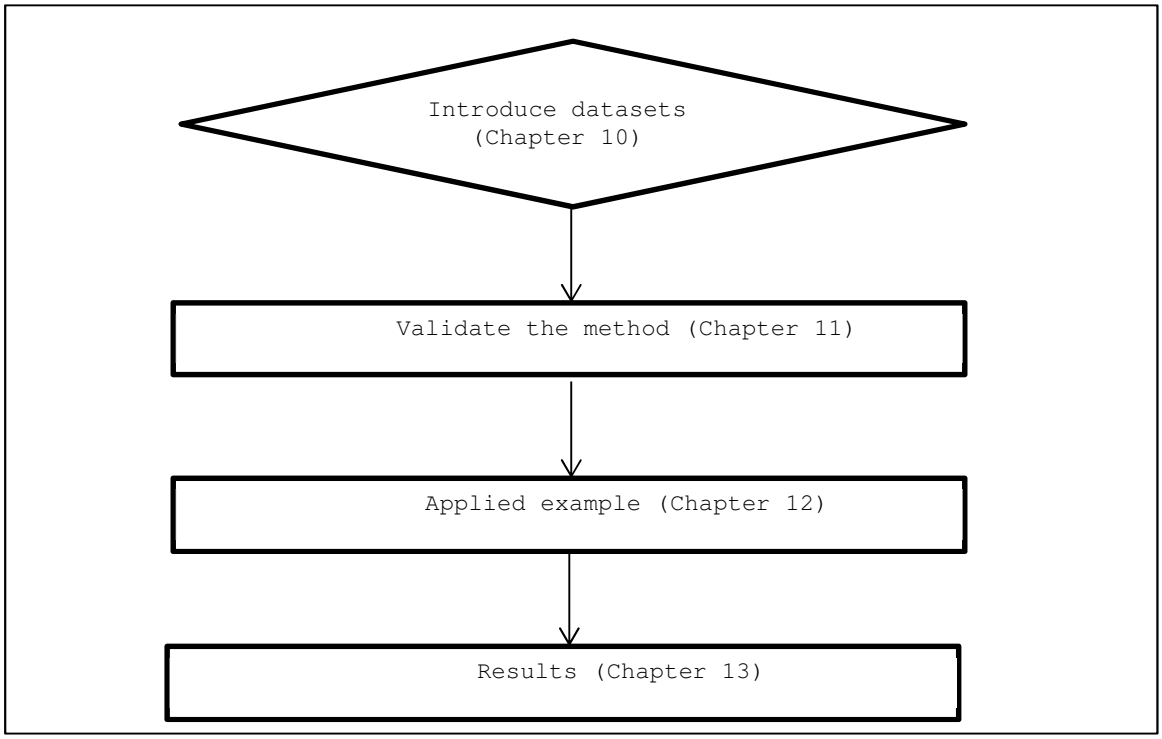


Figure 10: Flowchart showing the overall approach for data analysis

The first step of the overall approach is this chapter, which introduces the case study datasets and the reason for the order by which they are selected for analysis using the method. This will be followed by Chapter 11 – Method Validation, where the method will be validated to test its applicability prior to its use. This will be followed by Chapter 12 – Applied Examples, where the method will be tested with two datasets to show its reliability for use as a QRA method in the HHPI. The study will then present its findings from the method validation and applied example as Chapter 13 - Results.

10.1. Case Studies Dataset 1

The study selects NASA Bearing Data Test Set 3 as Case Study Dataset 1 because it has a similar number of variables as the Training Dataset but more observations. Available information about the data also indicate that risk associated with ball pass frequency outer (BPFO) occurred within the process operation, again like the risk which occurred in the Training Dataset. However, the component bearing which suffered the risk within the process operation was Bearing 3 which made this dataset different from the training dataset but good for use as the first case study dataset to validate the method.

10.2. Case studies Dataset 2

The study selects NASA Bearing Data Test 1 as Case Study Dataset 2 because three types of risk occurred in the process system, (a) Bearing 4 suffered from risks associated with BPFO and ball spin frequency (BSF), and (b) Bearing 3 suffered from risk associated with ball pass frequency inner (BPFI). Another comparative difference of this dataset from the Training Dataset was the number of channels used to monitor each of the component bearing in the process system. According to the available information, the operations of each component bearing was monitored by two channels which are odd and even numbered. This makes the dataset very suitable to be applied as the second case study dataset to validate the method because it provides an opportunity to test the ability of the method at detecting different risks suffered by different components within a process operation.

10.3. Case Studies Dataset 3

The study selects PHMM 2012 Bearing Data Sets Bearing 2_3 Full Test as Case Study Dataset 3 because it differs from the NASA Datasets by having only one bearing component with available information suggesting that the dataset also includes a temperature data. The operation of the component bearing was monitored with two accelerometers placed in a vertical and horizontal orientation. Available information did not provide any specifics about the type of risk suffered by component bearing. Since there study has no available information about any risk suffered by the component bearing within the process, the dataset was deemed more suitable to be use as the third dataset to test the method as one applied example. Additionally, the method has not been tested with temperature data hence this offer the opportunity to test the ability of the method to detect an entirely different kind of risk within a process operation.

10.4. Case studies Dataset 4

Finally, the study selects the PHMM 2012 Challenge Dataset Bearing 1_4 as Case Study Dataset 4 because like Case Study Dataset 3, available information suggests that the it contains two sets of data, i.e. temperature data and vibration data. Again, information about the risk or type of risks suffered by the process was available which offer the opportunity to use the dataset to test the method performance on detecting other kind of risks in a process operation data.

10.5. Conclusion

In this Chapter, the four case study datasets were introduced together with their description and the reasons for the order by which the study will apply them for testing the method was also provided. Next is Chapter 11- Method Validation, where a procedure for will be validated using Case Study Datasets 1 and 2 to test its suitability for a QRA.

Chapter 11 – Method Validation

11.*0. Introduction*

In the previous chapter, the study introduces four case study datasets which would be applied for the rest of the research and explain the order by which the datasets would be used. The reasons for the selected order of use of the datasets were also provided. This chapter presents how Case Study Datasets 1 and 2 are used to validate the method procedure. The study will apply feature extraction and the data exploration techniques applied to the training dataset to reduce the datasets to sizes. The data will be combined after which to the method validation will proceed by applying the method and associate procedure to Case Study Dataset 1 and 2 to demonstrate its applicability for risk analysis.

11.1. Method Validation using Case Studies Dataset 1

The study applied feature extraction to the test files within the dataset to obtain the key frequencies and statistical features of the vibration. The reduced files were combined and then timestamped to obtain four bearing-specific datasets, each having 6324 observations and 24 variables. As an example of the bearing-specific data files, the study presents the first 6 rows of the bearing-specific data obtained for Bearing 1 as Figure 11.1a.

```
Bearing-specific Data Frame for Bearing 1
1           timestamp  Min.x Qu.1.x Median.x Qu.3.x Max.x             Mean.x
SD.x              Skew.x
2 2004-03-04 09:27:46 -0.569 -0.056   -0.005  0.044 0.547 -0.004714599609375
0.0796307732817884  0.0308747370123272
3 2004-03-04 09:32:46 -0.461 -0.054   -0.005  0.044 0.491 -0.005101513671875
0.0785144090283644 -0.0138055503229216
4 2004-03-04 09:42:46 -0.454 -0.054   -0.002  0.049 0.515 -0.002241943359375
0.0797854881288294  0.0334173539439308
5 2004-03-04 09:52:46 -0.552 -0.054   -0.002  0.049  0.52  -0.00255224609375
0.0810442491823903 -0.0270380783172584
6 2004-03-04 10:02:46 -0.381 -0.054   -0.002  0.049 0.369  -0.00253408203125
0.0791263931542156 -0.0177739906252934
                V10                V11              V12              V13
V14              V15
1            Kurt.x              RMS.x            FTF.x           BPFI.x
BPFO.x            BSF.x
2  1.29230113393316 0.0797682761557461 8.68730494503513 11.0199302426349
0.969787540809107 3.88944735775457
3  1.15663108890614 0.0786780583468757 13.1644099792591 2.57102548280298
1.38635462553812 9.53197694611537
4 0.889743432130319 0.0798150336691614 4.54117532661761 10.6661279479568
8.24887103464855 10.0621644611705
5  1.18145301472534 0.0810824492366798  12.172417846639 10.3833806967826
6.6487381765722 11.3728937302604
6 0.585448011498533 0.0791650298624652 21.9490095124815 13.5176986118219
4.10805975777029 8.66973346867923
               V16              V17              V18              V19
V20              V21              V22
1             F1.x             F2.x             F3.x             F4.x
F5.x        VHF.pow.x         HF.pow.x
2 984.471139759742 985.447797636488 992.284402773708 977.634534622522
49.8095517140346 15751.5215744009 40130.4382415189
3 985.447797636488 986.424455513234 993.261060650454                0
49.8095517140346 17377.9680953657 40638.3980953874
4 985.447797636488 986.424455513234 978.611192499268 49.8095517140346
984.471139759742 17410.2663437864 40766.3384848014
5 985.447797636488 986.424455513234 984.471139759742 49.8095517140346
993.261060650454 16651.5813604622 41272.5508591928
6 986.424455513234 985.447797636488 49.8095517140346 993.261060650454
978.611192499268 16236.6256910209 40214.8737853593
               V23              V24
1         MF.pow.x         LF.pow.x
2 12209.6289977439 16733.6341518413
3  11845.643630412 15127.2500782438
4  11832.907781187 14779.9864416853
5 11770.8801021354 15490.7343289516
6 11294.6669549969 14797.7899421258
```

Figure 11.1a: First 6 rows of bearing-specific data for Bearing 1 of Case Study Dataset 1

Further inspection of the bearing-specific datasets reveals no missing values in any of the four datasets. Again, the study presents the structure of the data for Bearing 1 as Figure 11.1b. The four bearing specific data were combined into a new data frame which it was applied to validate the method using the procedure detailed in Section 9.6 of Part 3 of this study.

```
Structure of Bearing-specific Data Frame for Bearing 1
'data.frame':      6324 obs. of  24 variables:
 $ timestamp: Factor w/ 6324 levels "2004-03-04 09:27:46",..: 1 2 3 4 5 6 7 8 9
10 ...
 $ Min.x    : num  -0.569 -0.461 -0.454 -0.552 -0.381 -0.432 -0.422 -0.361 -
0.415 -0.439 ...
 $ Qu.1.x   : num  -0.056 -0.054 -0.054 -0.054 -0.054 -0.054 -0.054 -0.051 -
0.054 -0.054 ...
 $ Median.x : num  -0.005 -0.005 -0.002 -0.002 -0.002 -0.002 -0.002 -0.002 -
0.002 -0.002 ...
 $ Qu.3.x   : num  0.044 0.044 0.049 0.049 0.049 0.049 0.049 0.049 0.049 0.046
...
 $ Max.x    : num  0.547 0.491 0.515 0.52 0.369 0.483 0.461 0.498 0.527 0.396
...
 $ Mean.x   : num  -0.00471 -0.0051 -0.00224 -0.00255 -0.00253 ...
 $ SD.x     : num  0.0796 0.0785 0.0798 0.081 0.0791 ...
 $ Skew.x   : num  0.0309 -0.0138 0.0334 -0.027 -0.0178 ...
 $ Kurt.x   : num  1.292 1.157 0.89 1.181 0.585 ...
 $ RMS.x    : num  0.0798 0.0787 0.0798 0.0811 0.0792 ...
 $ FTF.x    : num  8.69 13.16 4.54 12.17 21.95 ...
 $ BPFI.x   : num  11.02 2.57 10.67 10.38 13.52 ...
 $ BPFO.x   : num  0.97 1.39 8.25 6.65 4.11 ...
 $ BSF.x    : num  3.89 9.53 10.06 11.37 8.67 ...
 $ F1.x     : num  984 985 985 985 986 ...
 $ F2.x     : num  985 986 986 986 985 ...
 $ F3.x     : num  992.3 993.3 978.6 984.5 49.8 ...
 $ F4.x     : num  977.6 0 49.8 49.8 993.3 ...
 $ F5.x     : num  49.8 49.8 984.5 993.3 978.6 ...
 $ VHF.pow.x: num  15752 17378 17410 16652 16237 ...
 $ HF.pow.x : num  40130 40638 40766 41273 40215 ...
 $ MF.pow.x : num  12210 11846 11833 11771 11295 ...
 $ LF.pow.x : num  16734 15127 14780 15491 14798 ...
```

Figure 11.1b: Structure of bearing-specific data frame for Bearing 1 of Case Study Dataset 1

11.2. Method Validation using Case studies Dataset 2

After applying feature extraction and combining the data files in the dataset with timestamp, the study obtains eight bearing-channel-specific data files each having 2156 observations and 47 variables. As an example, the study presents the first 6 rows of the bearing-channel-specific data frame for Bearing 1 as Figure 11.2a. Further inspection of the data frames reveals no missing observation. As an example, the study presents Figure 11.2b as a sample data frame of Bearing 1 -Channel 1 obtained. The eight data frames were then combined into a new data frame and use for the validation of the method using the procedure detailed in Section 9.6 of Part 3 of this research.

```
Structure of Bearing-channel-specific Data Frame for Bearing 1
1           timestamp  Min.x Qu.1.x Median.x Qu.3.x Max.x
Mean.x               SD.x              Skew.x
2 2003-10-22 12:06:24  -0.72 -0.146   -0.095 -0.042 0.388  -
0.09459287109375  0.081124065405696  -0.029990568091718
3 2003-10-22 12:09:13 -0.654 -0.146   -0.095 -0.044 0.388  -
0.09490263671875 0.0795173596870952 -0.0700697936495487
4 2003-10-22 12:14:13 -0.623 -0.149   -0.095 -0.044 0.317  -
0.09618681640625 0.0802188505259673 -0.0416429001339993
5 2003-10-22 12:19:13 -0.598 -0.149   -0.095 -0.042 0.457 -
0.095612744140625  0.080826519774341 0.00516121586920696
6 2003-10-22 12:24:13 -0.623 -0.149   -0.095 -0.042 0.388 -
0.095133056640625 0.0820362559091449  -0.060191220305041
1           Kurt.x             RMS.x            FTF.x           BPFI.x
BPFO.x            BSF.x F1.x
2  1.0687653858144 0.124613819082195 10.4745500337373 6.95225697667996
6.44887046012333 25.3127051397859    0
3 1.16114521646676 0.123811195806483 4.11215241506166 9.24309177429913
3.84348854189629 24.4489149063329    0
4  0.9858962249947 0.125246370877438 3.79377024745674 14.1530355486279
7.5658586445751 23.5946916445979    0
5 1.03389954518093 0.125197460638747 3.68711804266859 5.37066288962372
10.3380547119044 16.9702351316869    0
6 1.10976228835854  0.12561814016563  4.8019626933038 11.8339323762831
8.11705876962034 22.5047314448244    0
1             F2.x             F3.x             F4.x             F5.x
VHF.pow.x         HF.pow.x         MF.pow.x
2 986.424455513234 993.261060650454 493.212227756617 979.587850376013
17044.5531342753 38902.4580907636 9117.49700056245
3 986.424455513234 493.212227756617  987.40111338998 1003.02763941791
17102.8952354312 38024.4693318155 9167.22013192801
4 986.424455513234 493.212227756617   994.2377185272 969.821271608556
17008.8342348964 37976.3702047909 9135.92043858103
5 986.424455513234   994.2377185272 493.212227756617  987.40111338998
17449.5771653669 38702.7934974517  9343.3919305597
6 986.424455513234   994.2377185272 993.261060650454 979.587850376013
17346.1583834048 39450.3454039473 9400.12648284204
               V24
1         LF.pow.x
2 14041.5449690525
3 14010.1541690897
4 13711.5992548746
5 14396.1345164194
6 14638.8418613109
```

Figure 11.2a: First 6 rows of data frame for Bearing 1 -Chanel 1 of Case Study Dataset 2

```
Structure of Bearing-channel-specific Data Frame for Bearing 1
'data.frame': 2156 obs. of  47 variables:
 $ timestamp: Factor w/ 2156 levels "2003-10-22 12:06:24",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Min.x    : num  -0.72 -0.654 -0.623 -0.598 -0.623 -0.564 -0.667 -0.603 -0.696 -0.625
...
 $ Qu.1.x   : num  -0.146 -0.146 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149
...
 $ Median.x : num  -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.098 -0.095 -0.095
...
 $ Qu.3.x   : num  -0.042 -0.044 -0.044 -0.042 -0.042 -0.042 -0.042 -0.042 -0.044 -0.042
...
 $ Max.x    : num  0.388 0.388 0.317 0.457 0.388 0.432 0.439 0.381 0.413 0.479 ...
 $ Mean.x   : num  -0.0946 -0.0949 -0.0962 -0.0956 -0.0951 ...
 $ SD.x     : num  0.0811 0.0795 0.0802 0.0808 0.082 ...
 $ Skew.x   : num  -0.02999 -0.07007 -0.04164 0.00516 -0.06019 ...
 $ Kurt.x   : num  1.069 1.161 0.986 1.034 1.11 ...
 $ RMS.x    : num  -0.564 -0.491 -0.469 -0.474 -0.537 -0.591 -0.5 -0.505 -0.464 -0.535
...
 $ FTF.x    : num  -0.139 -0.137 -0.139 -0.142 -0.142 -0.142 -0.142 -0.142 -0.142 -0.139
...
 $ BPFI.x   : num  -0.093 -0.093 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095
...
 $ BPFO.x   : num  -0.049 -0.051 -0.054 -0.051 -0.051 -0.049 -0.051 -0.051 -0.051 -0.049
...
 $ BSF.x    : num  0.701 0.581 0.549 0.535 0.386 0.703 0.381 0.437 0.579 0.459 ...
 $ F1.x     : num  -0.0939 -0.0939 -0.0959 -0.0953 -0.0955 ...
 $ F2.x     : num  0.0706 0.0695 0.0695 0.0713 0.0722 ...
 $ F3.x     : num  0.2201 0.1265 0.151 0.0995 0.0959 ...
 $ F4.x     : num  3.07 2 1.97 1.74 1.18 ...
 $ F5.x     : num  -0.72 -0.654 -0.623 -0.598 -0.623 -0.564 -0.667 -0.603 -0.696 -0.625
...
 $ VHF.pow.x: num  -0.146 -0.146 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149
...
 $ HF.pow.x : num  -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.098 -0.095 -0.095
...
 $ MF.pow.x : num  -0.042 -0.044 -0.044 -0.042 -0.042 -0.042 -0.042 -0.042 -0.044 -0.042
...
 $ LF.pow.x : num  0.388 0.388 0.317 0.457 0.388 0.432 0.439 0.381 0.413 0.479 ...
 $ Min.y    : num  -0.0946 -0.0949 -0.0962 -0.0956 -0.0951 ...
 $ Qu.1.y   : num  0.0811 0.0795 0.0802 0.0808 0.082 ...
 $ Median.y : num  -0.02999 -0.07007 -0.04164 0.00516 -0.06019 ...
 $ Qu.3.y   : num  1.069 1.161 0.986 1.034 1.11 ...
 $ Max.y    : num  -0.564 -0.491 -0.469 -0.474 -0.537 -0.591 -0.5 -0.505 -0.464 -0.535
...
 $ Mean.y   : num  -0.139 -0.137 -0.139 -0.142 -0.142 -0.142 -0.142 -0.142 -0.142 -0.139
...
 $ SD.y     : num  -0.093 -0.093 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095
...
 $ Skew.y   : num  -0.049 -0.051 -0.054 -0.051 -0.051 -0.049 -0.051 -0.051 -0.051 -0.049
...
 $ Kurt.y   : num  0.701 0.581 0.549 0.535 0.386 0.703 0.381 0.437 0.579 0.459 ...
 $ RMS.y    : num  -0.0939 -0.0939 -0.0959 -0.0953 -0.0955 ...
 $ FTF.y    : num  0.0706 0.0695 0.0695 0.0713 0.0722 ...
 $ BPFI.y   : num  0.2201 0.1265 0.151 0.0995 0.0959 ...
 $ BPFO.y   : num  3.07 2 1.97 1.74 1.18 ...
 $ BSF.y    : num  -0.72 -0.654 -0.623 -0.598 -0.623 -0.564 -0.667 -0.603 -0.696 -0.625
...
 $ F1.y     : num  -0.146 -0.146 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149 -0.149
...
 $ F2.y     : num  -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.095 -0.098 -0.095 -0.095
...
 $ F3.y     : num  -0.042 -0.044 -0.044 -0.042 -0.042 -0.042 -0.042 -0.042 -0.044 -0.042
...
 $ F4.y     : num  0.388 0.388 0.317 0.457 0.388 0.432 0.439 0.381 0.413 0.479 ...
 $ F5.y     : num  -0.0946 -0.0949 -0.0962 -0.0956 -0.0951 ...
 $ VHF.pow.y: num  0.0811 0.0795 0.0802 0.0808 0.082 ...
 $ HF.pow.y : num  -0.02999 -0.07007 -0.04164 0.00516 -0.06019 ...
 $ MF.pow.y : num  1.069 1.161 0.986 1.034 1.11 ...
 $ LF.pow.y : num  -0.564 -0.491 -0.469 -0.474 -0.537 -0.591 -0.5 -0.505 -0.464 -0.535
...
```

Figure 11.2b: Structure of data frame for Bearing 1-Channel 1 of Case Study Dataset 2

11.3. Conclusion

In this Chapter, the two case study datasets selected to validate the method procedure were processed. The data compression with feature extraction gave four bearing-specific datafiles for Case Study Dataset 1 and eight bearing-channel-specific datafiles for Case Study Dataset 2. The bearing-specific and bearing-channel-specific of Case Study Dataset 1 and Case Study Dataset 2 were combined and used to validate the method. The result of the validation of the method and the applied examples will be presented in Chapter 13 – Results and discuss in Chapter 14- Discussion. Next is Chapter 12- Applied Examples, where the method will be applied to Case Study Datasets 3 and 4 to demonstrate its performance.

Chapter 12 – Applied Examples

12.*0. Introduction*

In Chapter 11 the QRA method which relies entirely on big data techniques and real-time datasets obtained was validated to demonstrate its applicability as a procedure for QRA. In this chapter, the study will apply the method to two datasets - Case Study Datasets 3 and 4, to demonstrate its applicability and performance as a QRA method.

12.1. Applying Method to Case Study Dataset 3

Base on available information the dataset was expected to have temperature data files. The temperature of the component bearings was monitored and sampled at a frequency of 10 Hz per second during the process operation. However, on inspection of the dataset the study did not find any temperature data.

The information also states that the vibration data of the bearings were sampled at a rate of 25.6 kHz every 10 seconds, so the study expect to find 2560 observations in each data file. The total duration for the process operation was calculated from the estimated times as 19970 sec (i.e. 2h 12 m 40 sec (train) + 3h 20min 10 sec (test)). However, the study found 1955 data files in the dataset which means 650 files are missing. This led to the suspicion that the duration of the process operation as stated could be an approximation.

Fifteen data files were selected at random and inspected. As an example of the data, the study presents the first six rows of the first file in the dataset as Figure 12.1a after renaming the variables as "Hour", "Minute", "Second", "u-second", "Horiz", "Vert" to reflect their description in the notes. Further investigation using descriptive statistics reveals there are no missing variables as in the example presented as Figure 12.1b. From the minimum and maximum observation on the channels (i.e. -1.504 and 1.438 for the horizontal channel; -9.78 and 9.88 for the vertical channel), the study suspects that any risk suffered by the component bearing may be defined by the vertical channel.

	Hour	Minute	Second	u_second	Horiz	Vert
1	8	39	57	571910	0.176	-0.133
2	8	39	57	571950	0.126	0.064
3	8	39	57	571990	-0.178	0.396
4	8	39	57	572030	-0.341	0.126
5	8	39	57	572070	-0.052	-0.243
6	8	39	57	572110	-0.056	0.166

Figure 12.1a: Sample of the Case Study Dataset 3

```
Descriptive statistics
              Hour Minute Second     u_second         Horiz          Vert
nbr.val       2560   2560   2560 2.560000e+03  2.560000e+03  2.560000e+03
nbr.null         0      0      0 0.000000e+00  1.000000e+00  6.000000e+00
nbr.na           0      0      0 0.000000e+00  0.000000e+00  0.000000e+00
min              8     39     57 5.719100e+05 -1.504000e+00 -9.790000e-01
max              8     39     57 6.718700e+05  1.438000e+00  9.880000e-01
range            0      0      0 9.996000e+04  2.942000e+00  1.967000e+00
sum          20480  99840 145920 1.592048e+09 -8.951000e+00 -1.890000e-01
median           8     39     57 6.218900e+05  2.600000e-02 -2.500000e-03
mean             8     39     57 6.218939e+05 -3.496484e-03 -7.382812e-05
SE.mean          0      0      0 5.706559e+02  8.511301e-03  4.962935e-03
CI.mean.0.95     0      0      0 1.118994e+03  1.668974e-02  9.731776e-03
var              0      0      0 8.336593e+08  1.854522e-01  6.305464e-02
std.dev          0      0      0 2.887316e+04  4.306416e-01  2.511068e-01
coef.var         0      0      0 4.642779e-02 -1.231642e+02 -3.401235e+03
```

Figure 12.1b: Descriptive statistics for sample of Case Study Dataset 3

The skewness (Figure 12.1c) suggests that the distribution on the horizontal channel is negatively skewed while that of the vertical channel is positively skewed which is confirmed by the density plot of the distribution (Figure 12.1d). The RMS on both channels are comparatively less than the standard deviations which suggests that the data is not from a non-zero-mean random stationary signal.

```
> skewness(data$Horiz)
[1] -0.2009063
> skewness(data$Vert)
[1] 0.09529851
> kurtosis(data$Horiz)
[1] 0.3700911
> kurtosis(data$Vert)
[1] 0.1639995
> sqrt(mean(data$Horiz**2))
[1] 0.4305716
> sqrt(mean(data$Vert**2))
[1] 0.2510578
```

Figure 12.1c: Other statistical parameters for sample data of Case Study Dataset 3

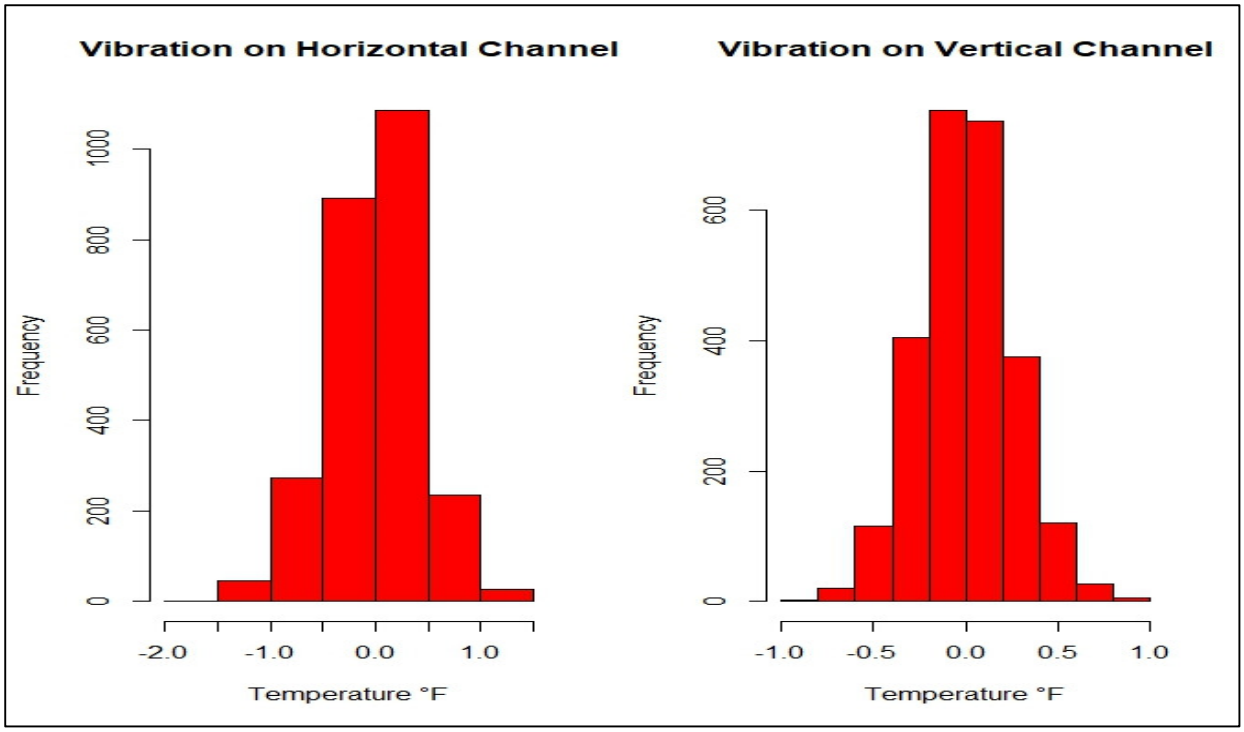


Figure 12.1d: Distribution of sample data of Case Study Dataset 3

The box plot (figure 12.1e) reveals that there more outliers in the data for vertical channel than that of the horizontal channel. Plot of the data (Figure 12.1f) show some healthy vibration which regularly spaced spikes in all four bearings. The plots are of a similar pattern with extreme values and a means closer to 0.

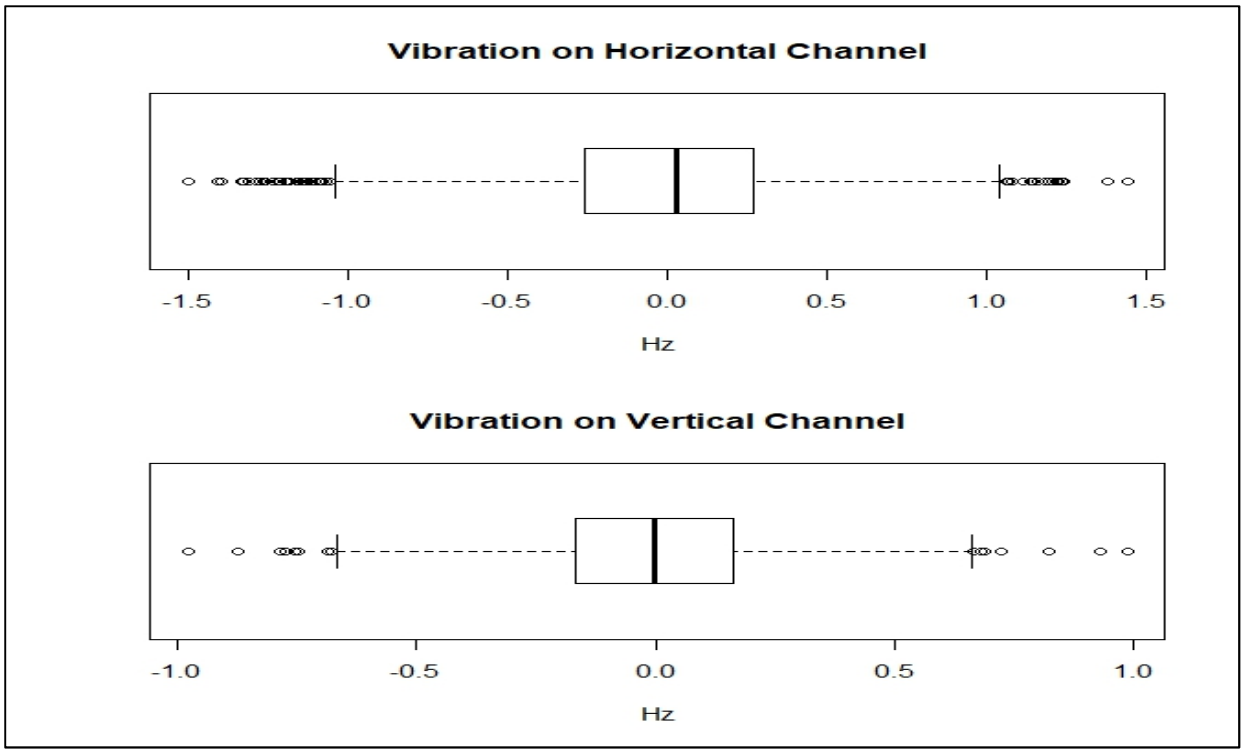


Figure 12.1e: Distribution of sample data of Case Study Dataset 3

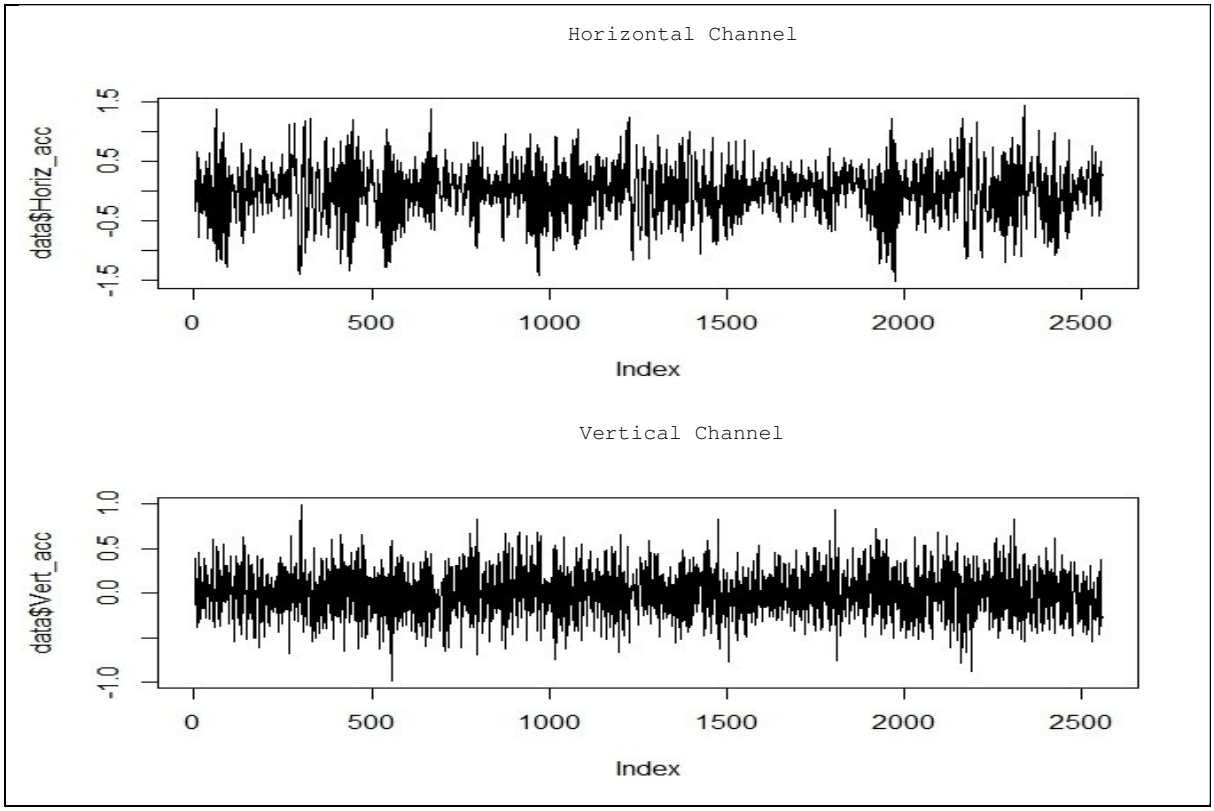


Figure 12.1f: Plots of vibration in sample data of Case Study Dataset 3

Prior to applying feature extraction to reduce the files, the study calculated the shaft rotational frequency (Rf) and contact angle using data from the description of the component bearing. The detail description of the bearings was provided (Qiu et al., 2006) as follows:

Ball diameter (Bd): 3.5 mm = 0.137795 in inches
Number of rolling elements (Nb): 13
Rotational speed (Rs): 2000 rpm = 33.3Hz
Pitch diameter (Pd): 25.6 mm = 1.008 inc
Contact angle (a): 0° = 0 radians

The Rf was therefore obtained from the rotational speed as 33.3 Hz (i.e. 2000/60). As a sample the study presents a zoomed FFT profile of the first file in the dataset showing feature BSF as green, BPFO as blue, BPFI as Red and FTF as brown, as Figure 12.1g.

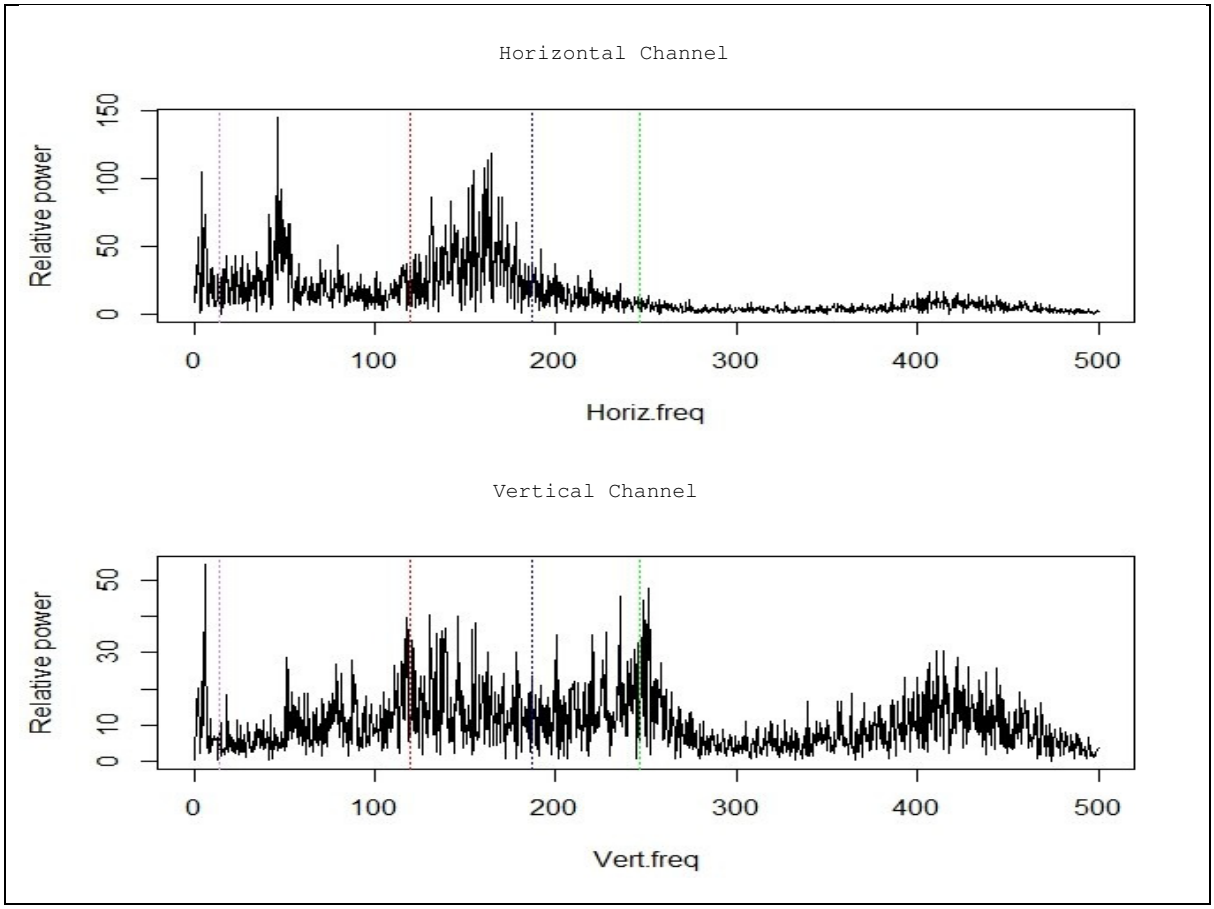


Figure 12.1g: FFT profile of Case Study Dataset 3

Two bearing-channel-specific data frames representing observations recorded by the horizontal and vertical channels obtained after the feature extraction process were combined, timestamped, then saved as csv file. However due to the suspicion that any risk suffered within the process could be defined by the vertical channel, the study proceeds with testing the method with the data for the vertical channel and presents the findings in Chapter 13-Results.

12.2. Applying Method to Case Studies Dataset 4

Available information suggests that this dataset appears also have temperature and vibration data files which was investigated and found to be valid. Again, fifteen vibration data files were selected at random and investigated. The study presents the first six rows of one of the datafile as Figure 12.2a. The descriptive statistics (Figure 12.2b) which reveal no missing observations. Feature extraction was performed as applied to Case Study Dataset 3 from which two bearing channel-specific datasets were obtained. Because the vertical channel was established to befine the risk event as explained in Section 12.1, the study applied the procedure of the method to the data recorded by the vertical channel and presents the result in Chapter 13.

Bearing vibration data Case Study Dataset 4

	Hour	Minute	Second	u_second	Horiz	Vert
1	8	8	0	425040	0.065	-0.058
2	8	8	0	425080	0.438	0.179
3	8	8	0	425120	-0.079	0.646
4	8	8	0	425160	-0.523	-0.411
5	8	8	0	425200	-0.146	-0.387
6	8	8	0	425230	0.292	0.208

Figure 12.2a: First six rows of vibration data of a sample data file for Case study Dataset 4

Descriptive statistics vibration data for Case Study Dataset 4

	Hour	Minute	Second	u_second	Horiz	Vert
nbr.val	2560	2560	2560	2.560000e+03	2.560000e+03	2.560000e+03
nbr.null	0	0	2560	0.000000e+00	5.000000e+00	2.000000e+00
nbr.na	0	0	0	0.000000e+00	0.000000e+00	0.000000e+00
min	8	8	0	4.250400e+05	-1.511000e+00	-2.045000e+00
max	8	8	0	5.250000e+05	1.373000e+00	1.658000e+00
range	0	0	0	9.996000e+04	2.884000e+00	3.703000e+00
sum	20480	20480	0	1.216048e+09	1.634700e+01	4.218000e+00
median	8	8	0	4.750200e+05	3.000000e-03	1.400000e-02
mean	8	8	0	4.750189e+05	6.385547e-03	1.647656e-03
SE.mean	0	0	0	5.706556e+02	7.970820e-03	8.991410e-03
CI.mean.0.95	0	0	0	1.118994e+03	1.562991e-02	1.763118e-02
var	0	0	0	8.336583e+08	1.626470e-01	2.069644e-01
std.dev	0	0	0	2.887314e+04	4.032951e-01	4.549334e-01
coef.var	0	0	NaN	6.078314e-02	6.315749e+01	2.761094e+02

Figure 12.2b: Descriptive statistics of a sample data file for Case Study Dataset 4

Available information states that the temperature data of the component bearing was sampled at a frequency stated as 10 Hz with 600 samples recorded per minute. As a result, the study investigates the temperature data by combining all of the temperature data files then inspect using descriptive statistics. Figures 12.2c and 12.2d represent the first six rows and the descriptive statistics of the data frame.

Sample of Temperature Data

	Hour	Minute	Second	Oxsecond	Temp
1:	8	9	47	9	70.036
2:	8	9	48	0	70.036
3:	8	9	48	1	70.058
4:	8	9	48	2	70.058
5:	8	9	48	3	70.081
6:	8	9	48	4	70.081

Figure 12.2c: First six row of temperature data of Case Study Dataset 4

The descriptive statistics shows reveals that there are 141627 observation in the data frame, no missing variables or zero's, and the minimum and maximum temperatures recorded are 70℉ and 164.6℉ respectively. The mean temperature was greater than the median which suggests that the distribution is negatively skewed. The RMS is comparatively greater than the standard deviation which suggests that the data is from a non-zero-mean random stationary signal.

Descriptive Statistics of Temperature data

	Hour	Minute	Second	Oxsecond	Temp
nbr.val	1.416270e+05	1.416270e+05	1.416270e+05	1.416270e+05	1.416270e+05
nbr.null	0.000000e+00	2.400000e+03	2.360000e+03	1.416300e+04	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	8.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7.003600e+01
max	1.200000e+01	5.900000e+01	5.900000e+01	9.000000e+00	1.646400e+02
range	4.000000e+00	5.900000e+01	5.900000e+01	9.000000e+00	9.460400e+01
sum	1.363040e+06	4.230619e+06	4.178517e+06	6.373140e+05	1.756125e+07
median	1.000000e+01	3.000000e+01	3.000000e+01	4.000000e+00	1.216300e+02
mean	9.624154e+00	2.987156e+01	2.950368e+01	4.499947e+00	1.239965e+02
SE.mean	3.049277e-03	4.575844e-02	4.601917e-02	7.632314e-03	3.775071e-02
CI.mean.0.95	5.976524e-03	8.968566e-02	9.019669e-02	1.495919e-02	7.399066e-02
var	1.316861e+00	2.965435e+02	2.999326e+02	8.250086e+00	2.018349e+02
std.dev	1.147545e+00	1.722044e+01	1.731856e+01	2.872296e+00	1.420686e+01
coef.var	1.192360e-01	5.764828e-01	5.869968e-01	6.382956e-01	1.145747e-01

Figure 12.2d: Descriptive statistics of temperature data of Case Study Dataset 4

The plot of the distribution (Figure 12.2e) shows that the data widely spread (boxplot) and the approaches a unimodal (histogram). The plot of the time series (Figure 12.2f) reveal a sharp rise in temperature to about 95℉ which the plateau just above 120℉ from about 500 – 1300 minutes. The temperature then rises to about 130℉ and levelled from about 1500 min to about 1900 min, after which a rise to about 170℉ was observed up to at about 2500 min.

Comparing the number of vibration data files to the temperature data files in the dataset, and the plot obtained from time series of the temperature data, it appears that the temperature data has no relevance to any risk suffered by the component bearing. As a result, the study therefore applies the procedure of the method only to the vibration data.

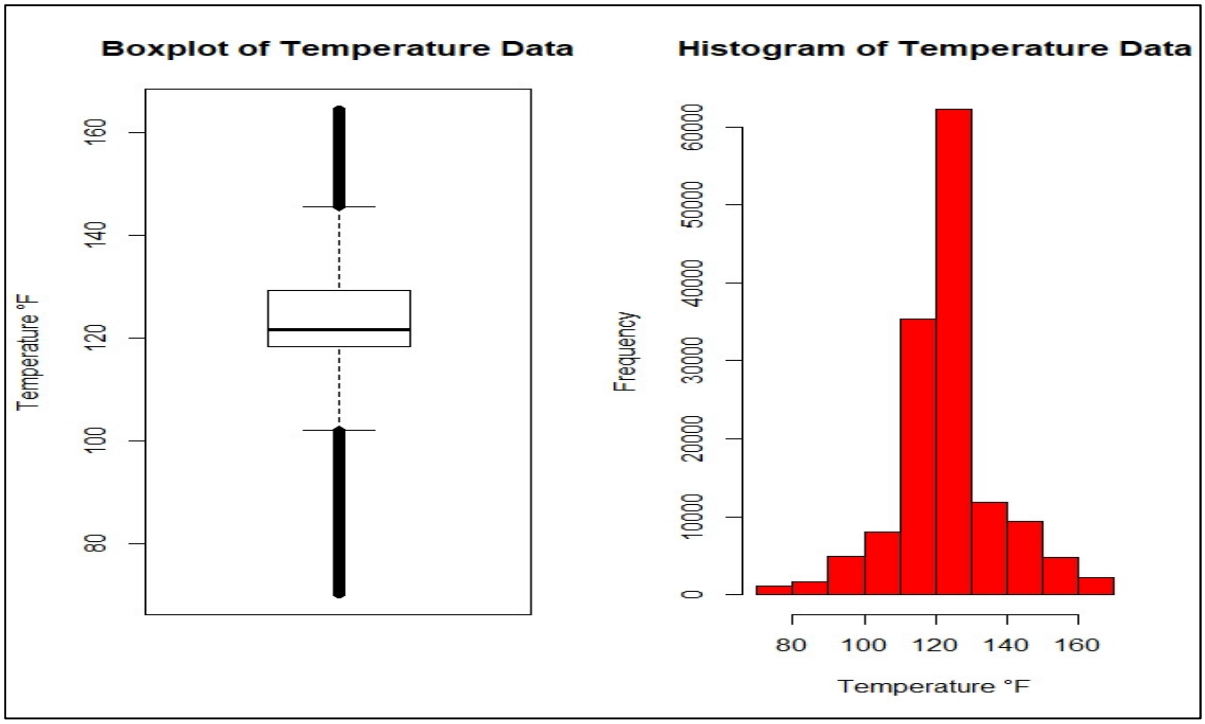


Figure 12.2e: Boxplot and histogram of temperature data of Case Study Dataset 4

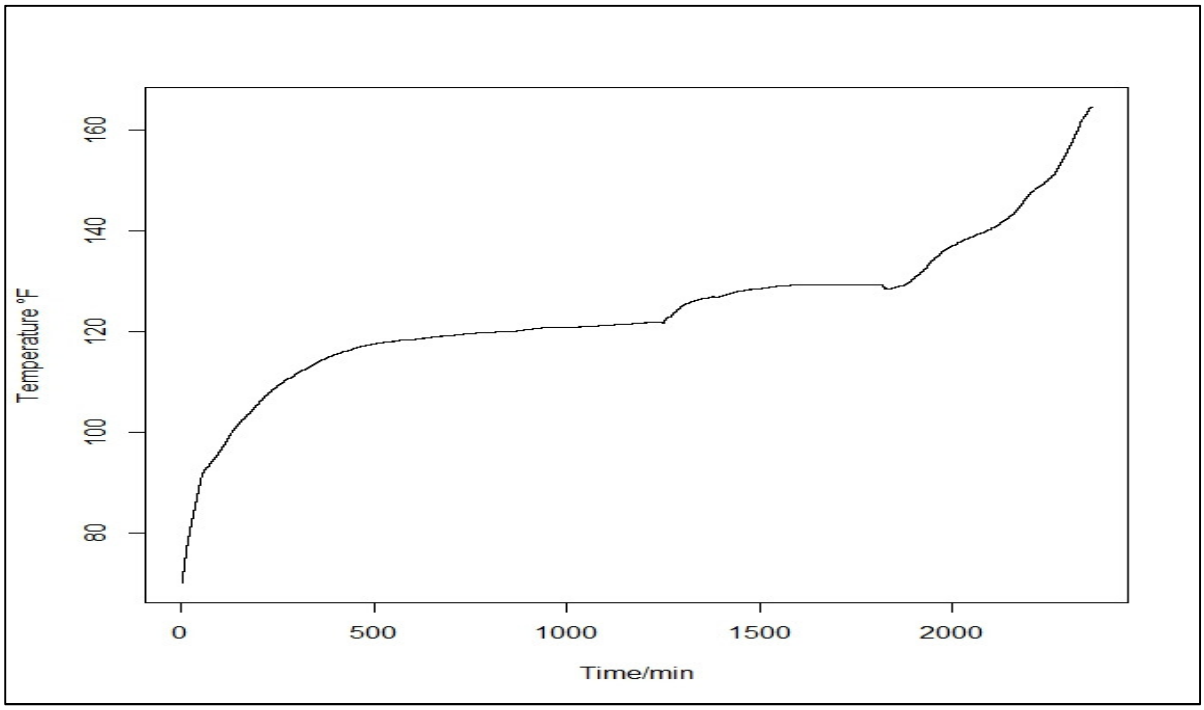


Figure 12.2f: Time series plot of temperature data of Case Study Dataset 4

12.3. Conclusion

This chapter has described the datasets, the data reduction process and findings from exploratory analysis of the data after which the method procedure was applied. Next is Chapter 13 – Results, where the findings from the method validation and application of the method as applied examples are presented.

Chapter 13 - Results

13.0. Introduction

In Chapter 11, the study validated the big data QRA method using two case study datasets. The study then proceeds to test the method with two other datasets in Chapter 12 as applied examples to prove that the method can be applied for QRA in the HHPI. The study now presents the findings of the validation of the method and the testing of the method as applied examples.

To recall from Chapter 10, the typ of risk and the component bearing which suffered from the risk event in the process form which Case Study Dataset 1 and 2 were obtained are known. As a result, that information will be used to ascertain whether the outcome of the method validation proves the validity of the method as suitable for QRA in the HHPI. Again, a plot of root mean square (RMS) of the lifecycle of the component bearing which suffered from the risk event will be used to investigate the time index of the risk detected by the method as explained in Part 3.

For the performance of the method when tested with the Case Study Datasets 3 and 4 in Chapter 12 as applied examples, the type of risk suffered by the component bearings are unknown. As a result, the method was tested for its ability to detect the unknown risk type. Again, a plot of the RMS of the lifecycle of the bearing will be used to confirm the time index of the risk detected. Any other information available e.g. RUL will be used to confirm the time index of the risk if necessary.

13.1. Method Validation using Case Studies Dataset 1

Figure 13.1a shows the plots obtained by applying package *changepoint* (using the PELT algorithm) which reveals the following:

- One risk event was detected at time index 5447 in the data for Bearing 1.
- One risk event was detected at time index 3245 in the data for Bearing 2
- Two risk events were detected at time index 3270 and 6294 in the data for Bearing 3.
- No risk event was detected in the data for Bearing 4.

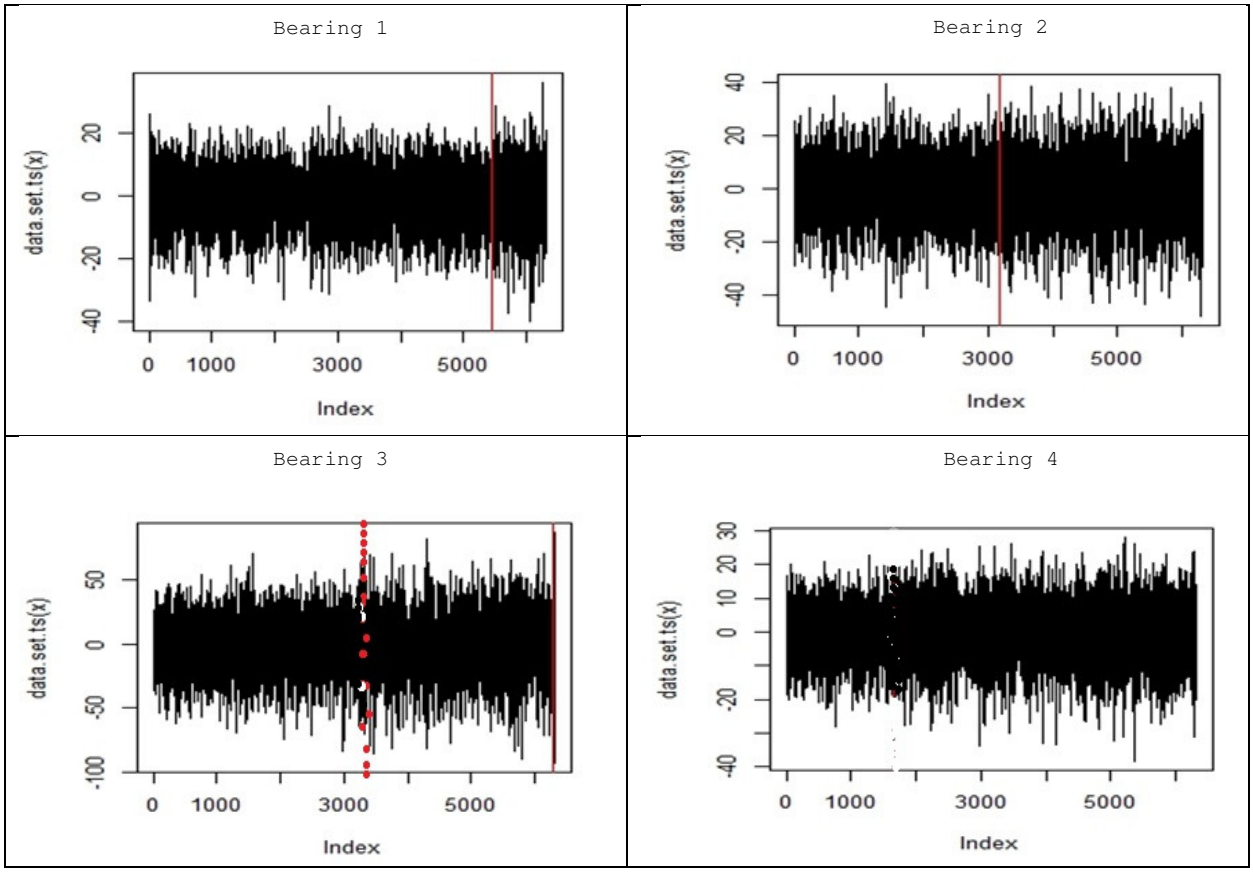


Figure 13.1a: Risks detected by package changepoint in the data of Case Study Dataset 1.

The plots obtained when package *strucchange* was applied (Figure 13.1b) to the data reveals that:

- one risk event was detected at time index 4060 in the data for Bearing 1.
- two risk events were detected at time indices 2787 and 4239 in the data for Bearing 2.
- two risk events were detected at time indices 2778 and 5374 in the data for Bearing 3.
- one risk event was detected at time index 1877 in the data for Bearing 4.

Since available information specified that the component bearing which suffer from the risk event was Bearing 3, the study suspects that the risks detected in the data of the other component bearings could either be due to the data being very noisy or because bearings were beginning to suffer from some form of risks at the time indices registered. However, because the information states that the risk was observed only on Bearing 3 it was decided that the focus must be on Bearing 3. As a result, no further investigation was performed to investigate the risk events detected in the data obtained from the operation of the other bearings apart from that of Bearing 3.

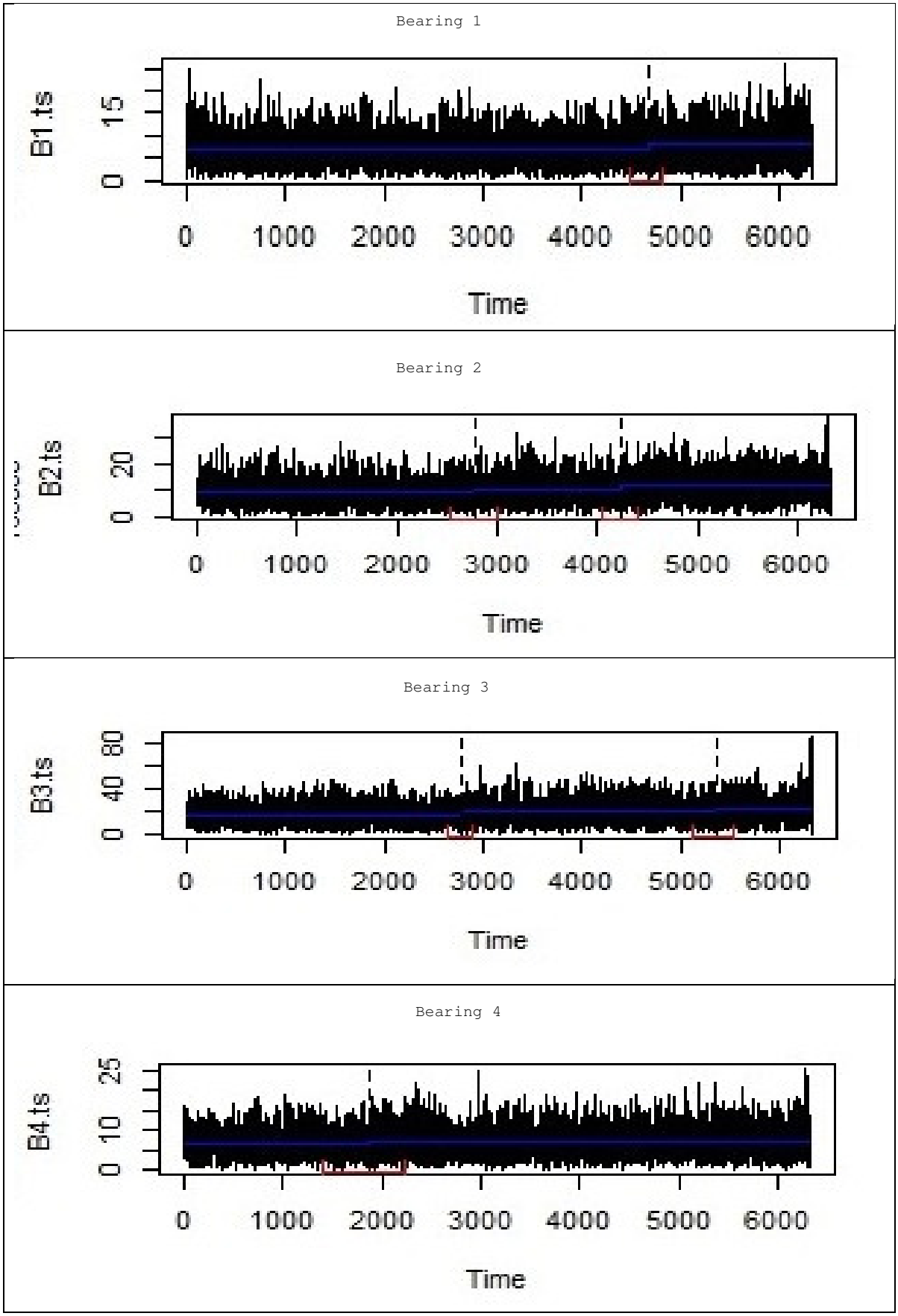


Figure 13.1b: Plot for risks detected by package *strucchange* for Case Study Dataset 1

The plot of the RMS in the data for the lifecycle of Bearing 3 (Figure 13.1c) have the following characteristics:

- The smoothness of the plot changes from around time index 3000 which falls between the time index 2778 of the risk detected by package *strucchange*, and the time index 3270 of the risk detected by package *changepoint*.
- There is a study rise in the trend from time index 6000 which is close to the time index 6294 of the risk detected by package *changepoint*.

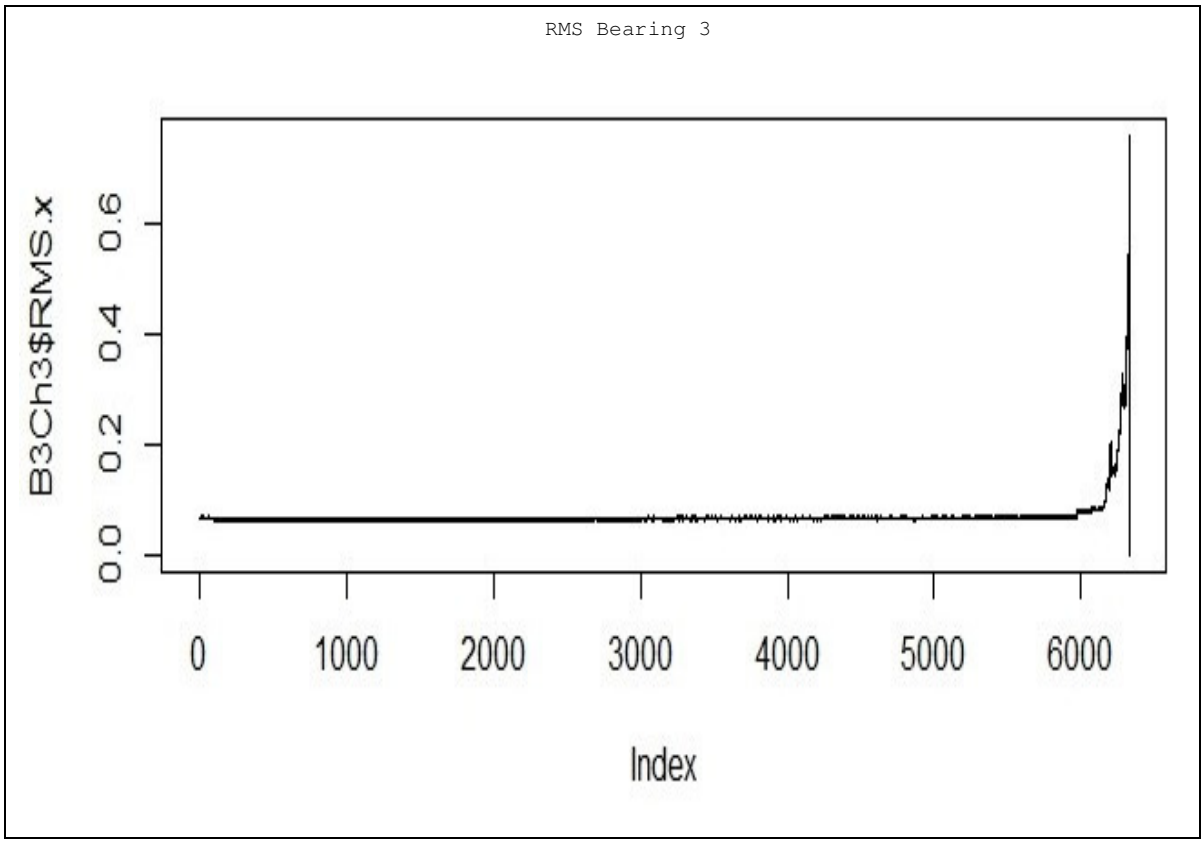


Figure 13.1c: Plot of RMS of the lifecycle of data for Bearing 3 for Case Study Dataset 1.

Plots for the relationships between component bearings (Figure 13.1d) for the interactions up to the time index of the risks detected shows low Pearson correlations except the relationship between Bearing 2 and Bearing 3. There are other correlations with small p-values which are due to the large size of N. Hence, apart from the relationship between Bearing 2 and Bearing 3, the bearings do not appear to have a relationship and therefore any interaction effect. The correlation between Bearing 2 and Bearing 3 at the time index of the risk detected by package changepoint is greater than that of the time index of the risk detected by package strucchange. This led to the suspicion that the risk suffered by Bearing 3 may have been influence by the vibrations from the operations of only Bearing 2. This was therefore investigated with the decision tree modelling.

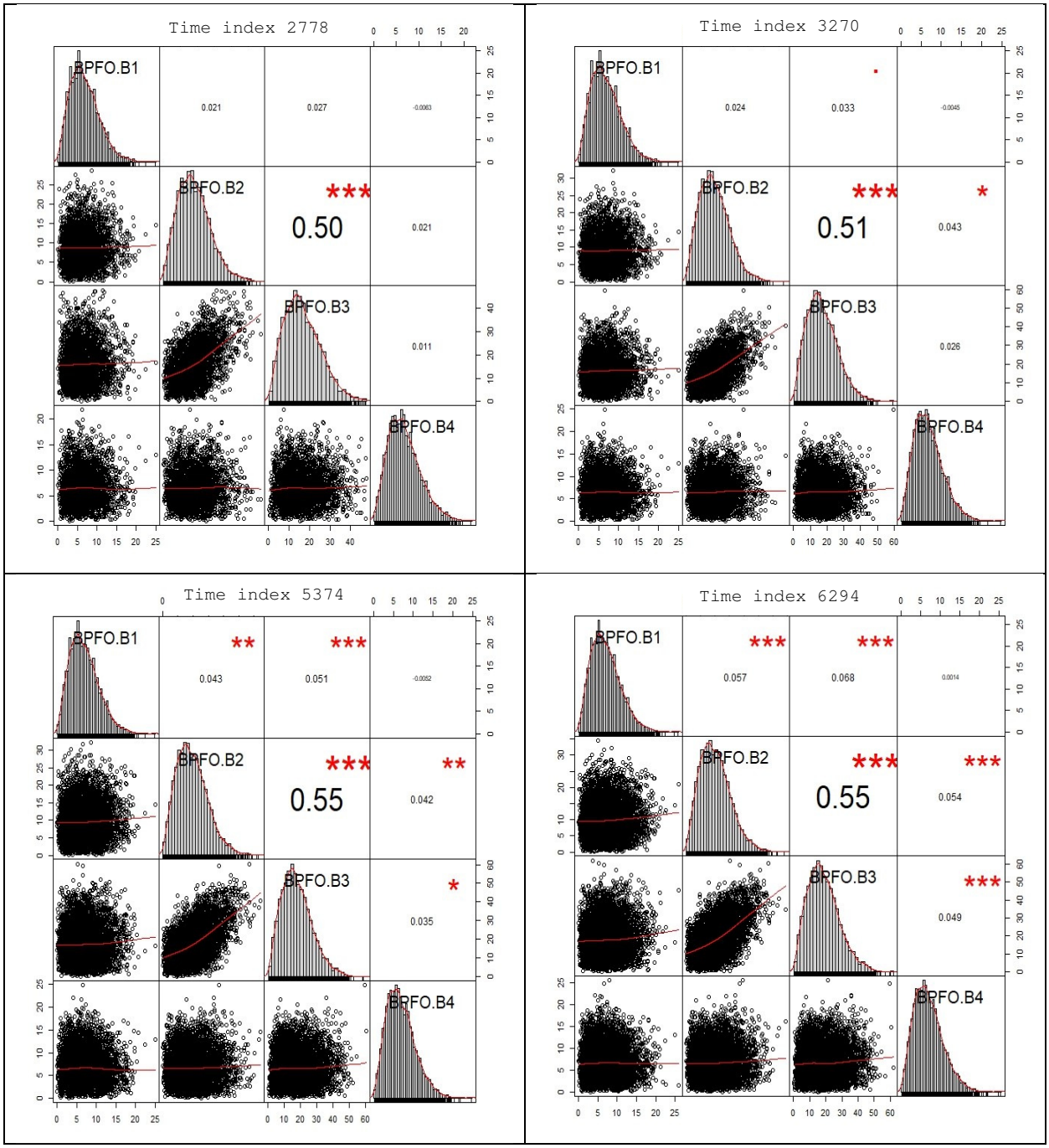


Figure 13.1d: Correlation for interactions up to time index of the risks detected in Case Study Dataset 1.

The decision tree plot (Figure 13.1e) for the interactions up to the time index of the risks detected also shows Bearing 2 as the only component whose operations affects the operations of Bearing 3. Thus, there are no moderation or interaction effects on the contribution of the vibrations of Bearing 2 on the vibrations of Bearing 3 at all four indices of the time of the risks detected. The study therefore concludes the contribution from Bearing 2 to the risk suffered by Bearing 3 is not influence by moderation from any of the other component bearings or

bearing-bearing interaction effect. The study proceeds to review the output of the regression model.

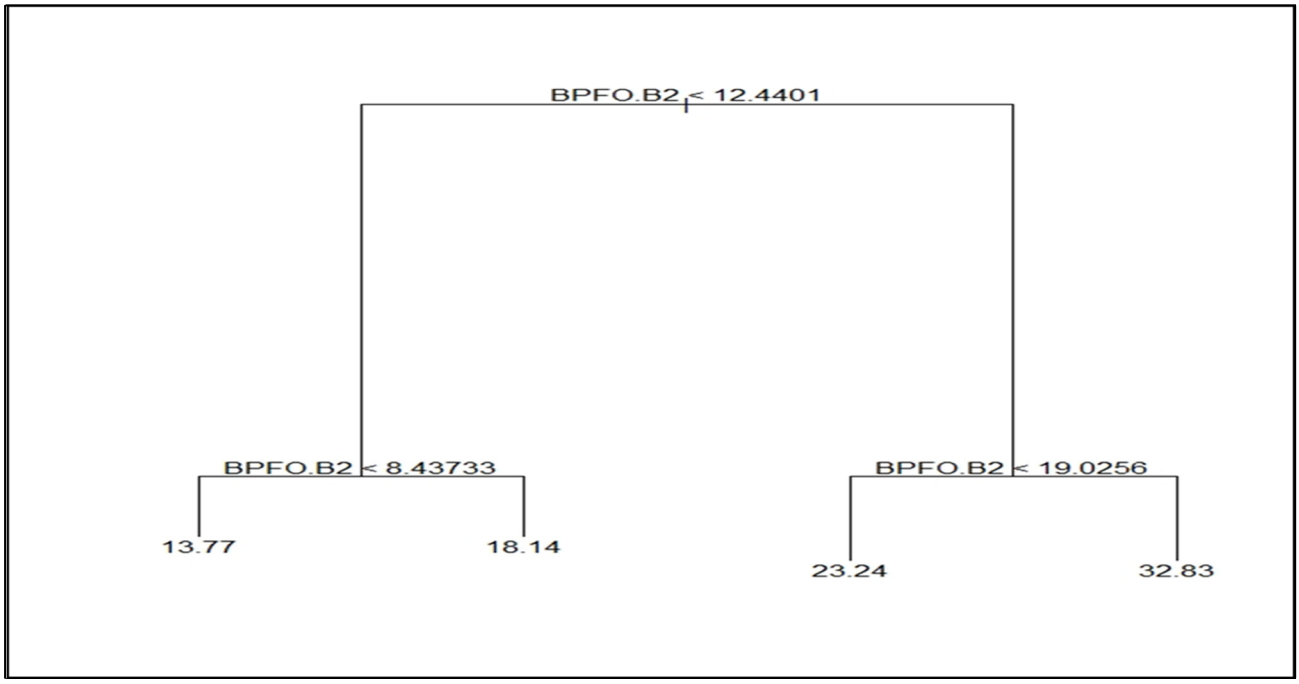


Figure 13.1e: Tree plots for interactions up to time index of the risks detected in Case Study Dataset 1

The summary output of the regression model investigating the interactions up to the time index of the risks detected in the data (Table 13.1a) reveals that:

- When Bearing 2 is not in operation, the average vibration of Bearing 3 up to the time index 2778 of the risk detected is approximately 8.21 Hz which increase significantly by 0.919 Hz per unit increase in the operations of Bearing 2 (model 1).
- When Bearing 2 is not in operation, the vibration of Bearing 3 up to time index 3270 is 8.293 Hz which increases significantly by 0.934 Hz per unit increase in the vibration of Bearing 2 (model 2).
- When Bearing 2 is not in operation, the vibration of Bearing 3 up to the time index 5374 is 8.021 Hz which increases significantly by 1.011 Hz per unit vibration of Bearing 2.
- When Bearing 2 is not in operation, the vibration of Bearing 3 up to the time index 6294 is 8.163 Hz which increases significantly by 1.015 Hz per unit vibration of Bearing 2.

Because the decision tree model provides no evidence of moderation from any of the other bearings or bearing-bearing interaction effect on the contribution from the operations of Bearing 2 to the risk suffered by Bearing 3, no further investigation was performed by the study. The study therefore concludes that there is evidence of system complexity being exhibited up to the time index of the risks detected.

Table 13.1a: Regression output for interactions up to time index of risk detected in Case Study Dataset 1

	Dependent variable:			
	BPFO.B3			
	model.1	model.2	model.3	model.4
	(1)	(2)	(3)	(4)
Constant	8.213***	8.293***	8.021***	8.163***
	(0.309)	(0.290)	(0.233)	(0.221)
BPFO.B2	0.919***	0.934***	1.011***	1.015***
	(0.030)	(0.028)	(0.021)	(0.019)
Observations	2,778	3,270	5,374	6,294
R2	0.252	0.259	0.302	0.305
Adjusted R2	0.252	0.259	0.302	0.305
Residual Std. Error	7.520 (df = 2776)	7.645 (df = 3268)	7.866 (df = 5372)	8.022 (df = 6292)
F Statistic	934.898*** (df = 1; 2776)	1,143.574*** (df = 1; 3268)	2,320.879*** (df = 1; 5372)	2,757.454*** (df = 1; 6292)
Note:				*p<0.1; **p<0.05; ***p<0.01

The decision tree to determine for interaction of the risk features up to the time index of the risks detected produce single node trees. This suggests that the features of the risks are independent and not influenced by one another. As a result, no further investigation was conducted by the study.

13.2. *Method Validating with Case Studies Dataset 2*

Figure 13.2a represent the time series plots of data showing the risks detected by package *changepoint*. The plots reveal that risk BPFI was detected in the data of channels 5 and 6 on Bearing 3 at the time index 2127 and 2119 respectively. Risk BPFO was detected in the data of channels 7 and 8 on Bearing 4 at the time index 1654 and 1710. Finally, risk BSF was detected in the data of channels 7 and 8 on Bearing 4 at time index 1735 and 1722 respectively.

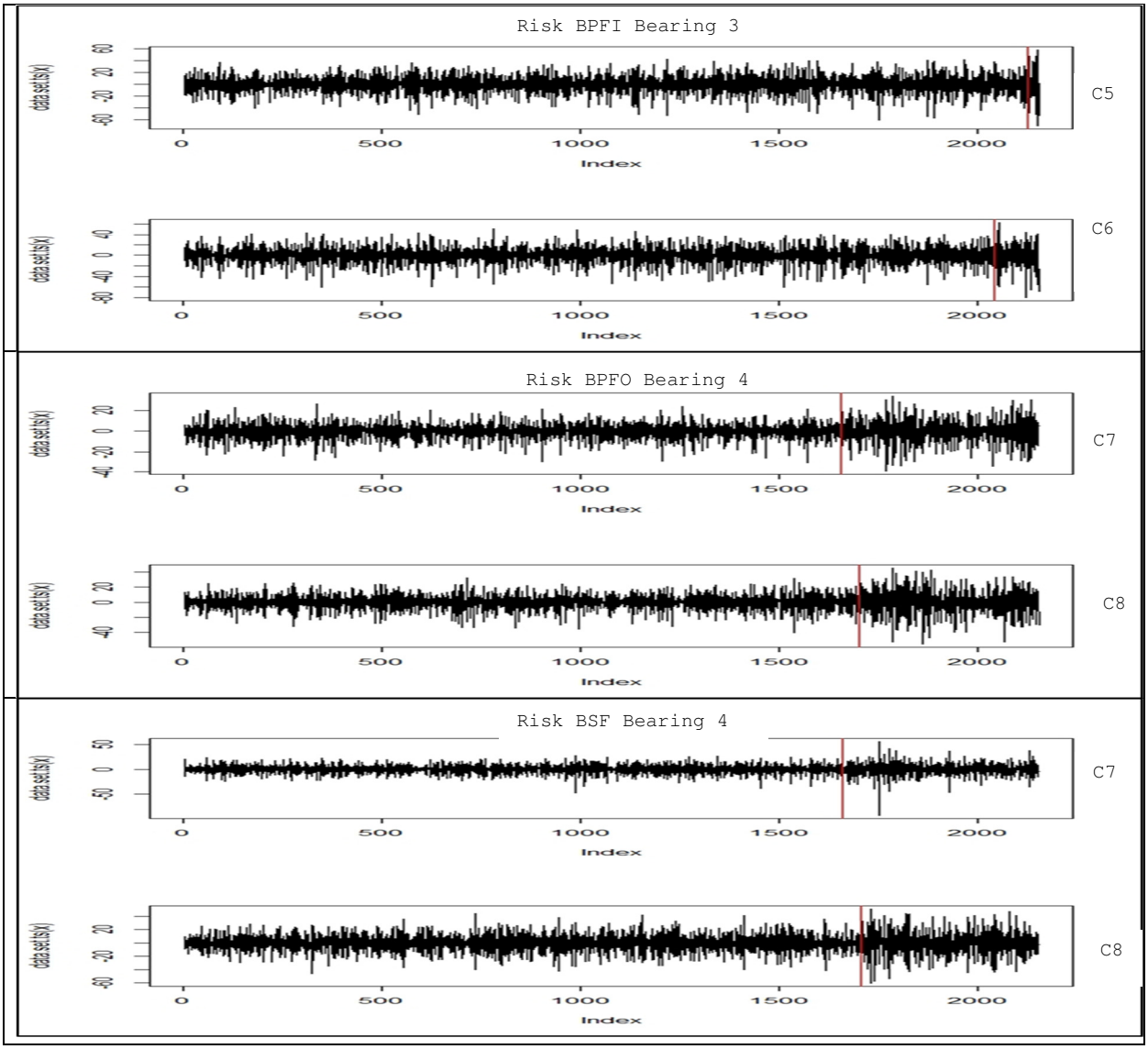


Figure 13.2a: Plots for risks detected by package *changepoint* in Case Study Dataset 2

Figure 13.2b represents the plots of data showing the risks detected by package *strucchange*. The plots reveal that risk BPFI was detected in the data of channels 5 and 6 on Bearing 3 at time index 1730 and 1833 respectively. Risk BPFO was detected in the data of channels 7 and 8 on Bearing 4 at time index 1648 and 1682 respectively. Risk BSF was detected in the data of channels 7 and 8 on Bearing 4 at time index 1699 and 1693 respectively. The package also detected some events in the data which are suspected to have been caused by the presence of noise in the data. Some of these noisy events were detected at time index 417 in the data of channel 6 on Bearing 3 and 652 in the data of channel 8 on Bearing 4 for risk BPFO. However, the magnitude of the noise which cause their detection led to the suspicion that the even numbered channels defines the risk.

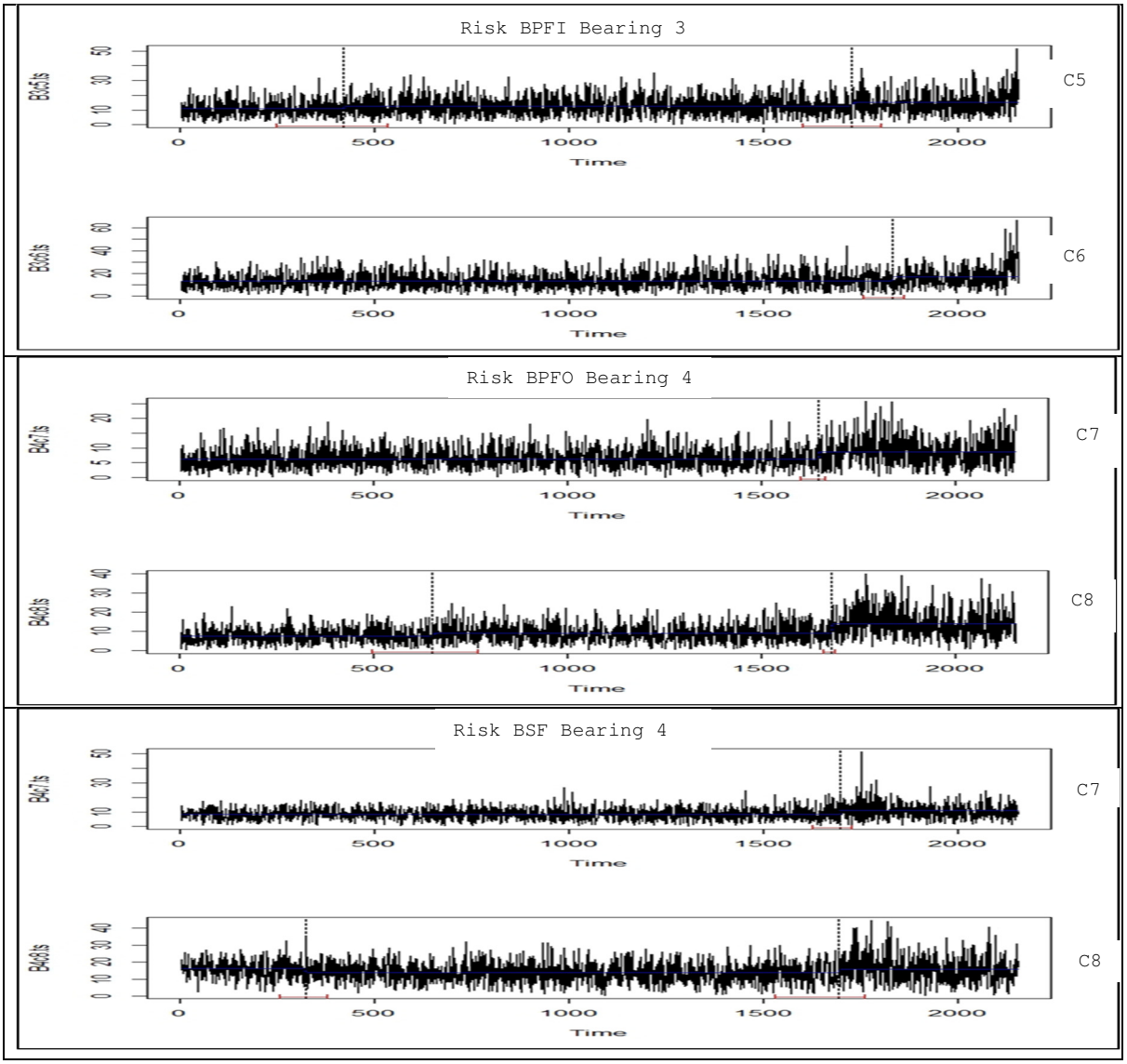


Figure 13.2b: Plots for risks detected by package *strucchange* for Case Study Dataset 2

Since there is no available information to suggest that the data was recorded on the two channels at different rates, it was deemed that the recording on the two channels were obtained at a similar rate. This could explain the close proximity of the time index of the risk events detected in the data from both channels. As a result, the study selects the data recorded on the even channel for onward analysis. The study then plots the RMS of the lifecycle of all four bearings (RMS of Bearing 1 and Bearing 2 are presented as Figure 13.2c, Appendix page 267 -268).

The plots of the RMS of the lifecycle of Bearing 3 and Bearing 4 (Figure 13.2c) shows that the trend of RMS of Bearing 3 rises from around time index 2100 which appear close to the time index of risks detected by package *changepoint* (2119). The trend of the RMS of the lifecycle of Bearing 4 also rises between the time index 1648 and 1795 which appears close to the time

indices of the risks detected in the data for Bearing 4 by the two packages. Additionally, there are the trend of the RMS for Bearing 4 appear spiky between time indices 200 to 1400. The study proceeds with the method validation using the data recorded by channel 6 on Bearing 3 and channel 8 of Bearing 4 for the risks suffered by the two bearings respectively.

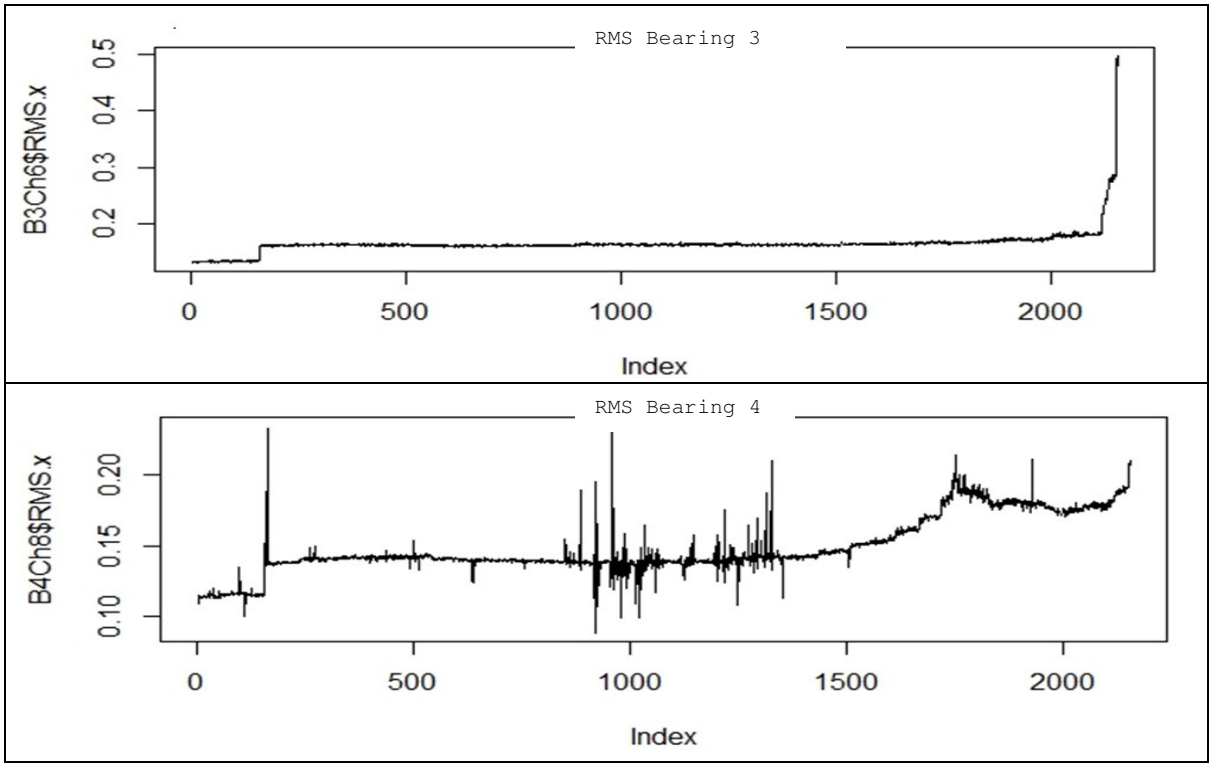


Figure 13.2d: RMS of the lifecycle of the Bearing 3 Bearing 4 of Case Study Dataset 2.

Investigating relationships for the interactions up to the time index of the risks detected, the study found that the correlation for all interaction up to the index 1800 appear similar and those for time indices greater than 2000 were also similar. This was suspected to have been cause by the proximity of the time indices of the risks detected. As an example, the study presents the correlation plots for the interactions up to time index 1710 for the risk event associated with BPFI detected in the data for Bearing 3,time index 2000 and 2119 for the risk events associated with BSF and BPFO detected in the data for Bearing 4 as Figure 13.2e.

The plots show that for risk BPFI, Bearing 3 has a weak but statistically significant Pearson correlation with Bearing 2 at time index 1710 and 2119. For risk BPFO, there are no significant correlations between Bearing 4 and any of the other bearings at time index 1710 and 2119. For risk BSF, Bearing 4 has a weak but significant correlation with all three bearings with small p-values due to the large size of N. The study therefore concludes there is no proper relationship between Bearing 4 which suffered from the risk and the other bearings.

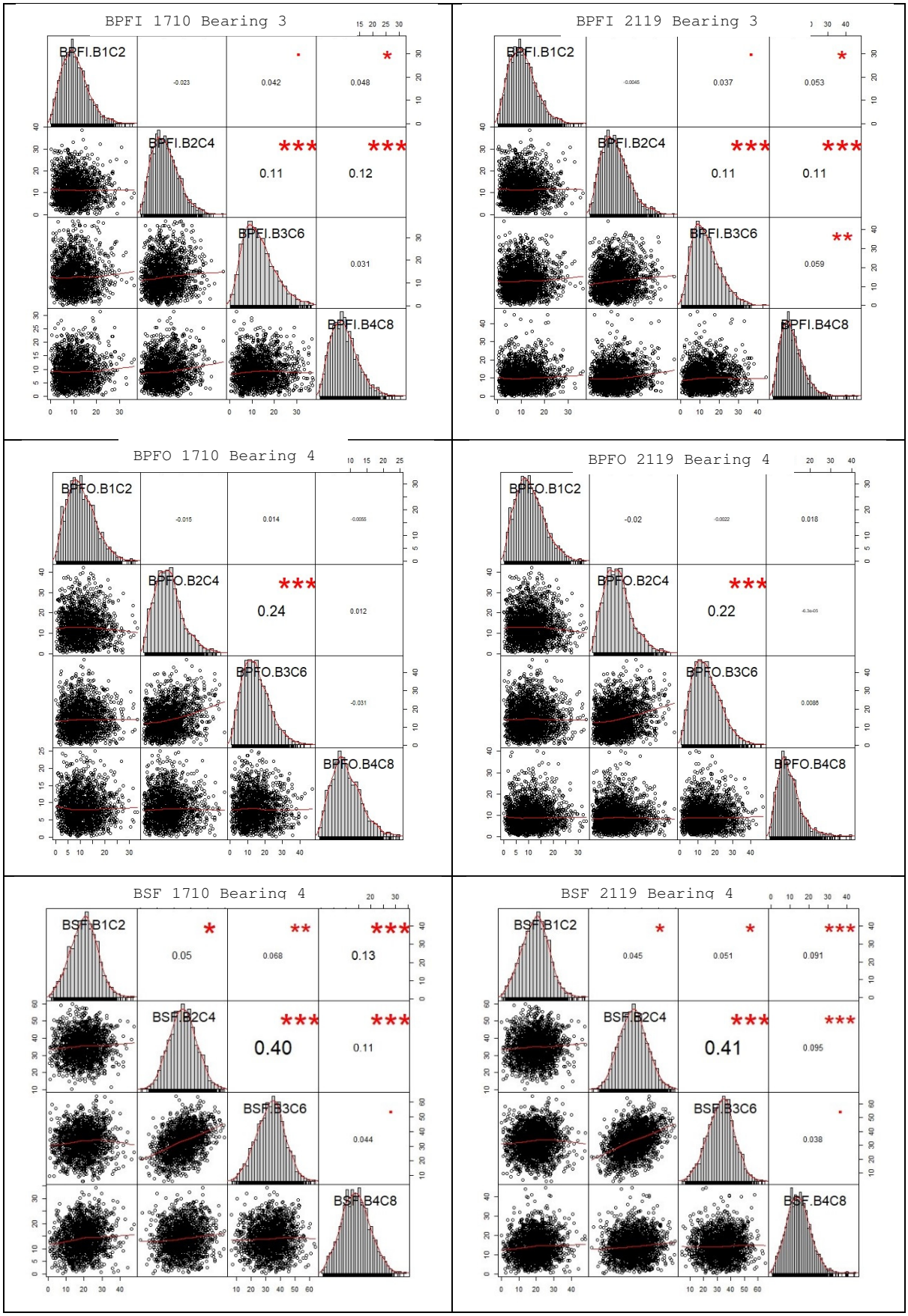


Figure 13.2e: Correlations for interactions up to time index 1710 and 2119 in Case Study Dataset 2.

This was confirmed by the decision tree model which produce single tree nodes. Owing to this, the study concludes that there none of the other component bearings in the process system made a contribution to the risks associated with BPFO and BSF suffered by Bearing 4. This suggests that risks BPFO and BSF suffered by Bearing 4 are not influenced by moderation or interaction of any of the other component bearings in the process system. Hence, there was system simplicity being exhibited through interaction of the component bearings up to the time index of the two risk events detected in the data of Bearing 4. As a result, no further investigation was performed for the risks detected in the data of the bearing.

The decision tree obtained for the interactions up to the risk even associated with BPFI suffered by Bearing 3 show Bearing 2 as the only component whose vibrations affects the vibrations of Bearing 3 up to the time index of the risks detected which. This also suggests that the operations of Bearing 3 are affected by the operations of Bearing 2 without any influence from the other component bearings through moderation or interaction effect.

Table 13.2: Output interactions up to time index of risk BPFI in Case Study Dataset 2.

	Dependent variable:	
	BPFI.B3	
	model.1	model.2
	(1)	(2)
Constant	11.624***	11.946***
	(0.367)	(0.338)
BPFI.B2	0.128***	0.133***
	(0.028)	(0.025)
Observations	1,710	2,119
R2	0.012	0.013
Adjusted R2	0.012	0.012
Residual Std. Error	6.994 (df = 1708)	7.188 (df = 2117)
F Statistic	21.112*** (df = 1; 1708)	27.495*** (df = 1; 2117)
Note:		*p<0.1; **p<0.05; ***p<0.01

The output of the regression for the reveals that when Bearing 2 is not in operation the average vibration of Bearing 3 up to the time index 1710 is 11.62 Hz which increase significantly by 0.128 Hz per unit increase in the vibrations of Bearing 2. However, at time index 2119 the average vibration of Bearing 3 is 11.95 Hz when Bearing 2 is not in operation. This increase significantly by 0.133 Hz per unit increase in the vibrations of Bearing 2. Since the decision tree provides no evidence of moderation or interaction effect on the contributions from Bearing 2 to the vibrations of Bearing 3 by any of the component bearings, no further investigation was conducted. The study therefore concludes that there is evidence of system complexity being exhibited through communication between Bearing 2 and Bearing 3 up to the time index 170 and 2119 of the risk events detected in the data for Bearing 3.

13.3. Applied Example using Case Study Dataset 3

The time series plots showing the time index of the risks detected in Case Study Data Set 3 by package *changepoint* is provided as Figure 13.3a. The plots show non-uniformity in the time series which reveal that there was some disturbance from the beginning of the operation of the bearing up to about time index 700. There appear to be more noise in the time series plots of the data or risk BPFI followed by BSF and finally BPFI. This led to the suspicion that there were issues with the bearings from the onset of the process operations. The suspected issues include (a) the bearing not being properly secured within the process system or (b) bearing already suffering some defect prior to its installation. Because of unavailability of information about the type of risk suffered by the bearing, the study investigates all three risk events associated with bearings.

The detail outcome from the investigation reveals that all three type of risks were detected at time index 1945. The risk event shown in the time series plot at time index 715 for BPFI is suspected to be due to the noise in the data caused by a disturbance in the vibration. As discussed in Chapter 11, there was suspicion that the risk could be defined by the vertical channel. Besides there is no available information on the rates at which the data was recorded. As a result, the study applies the procedure of the method to the data recorded by the vertical channel.

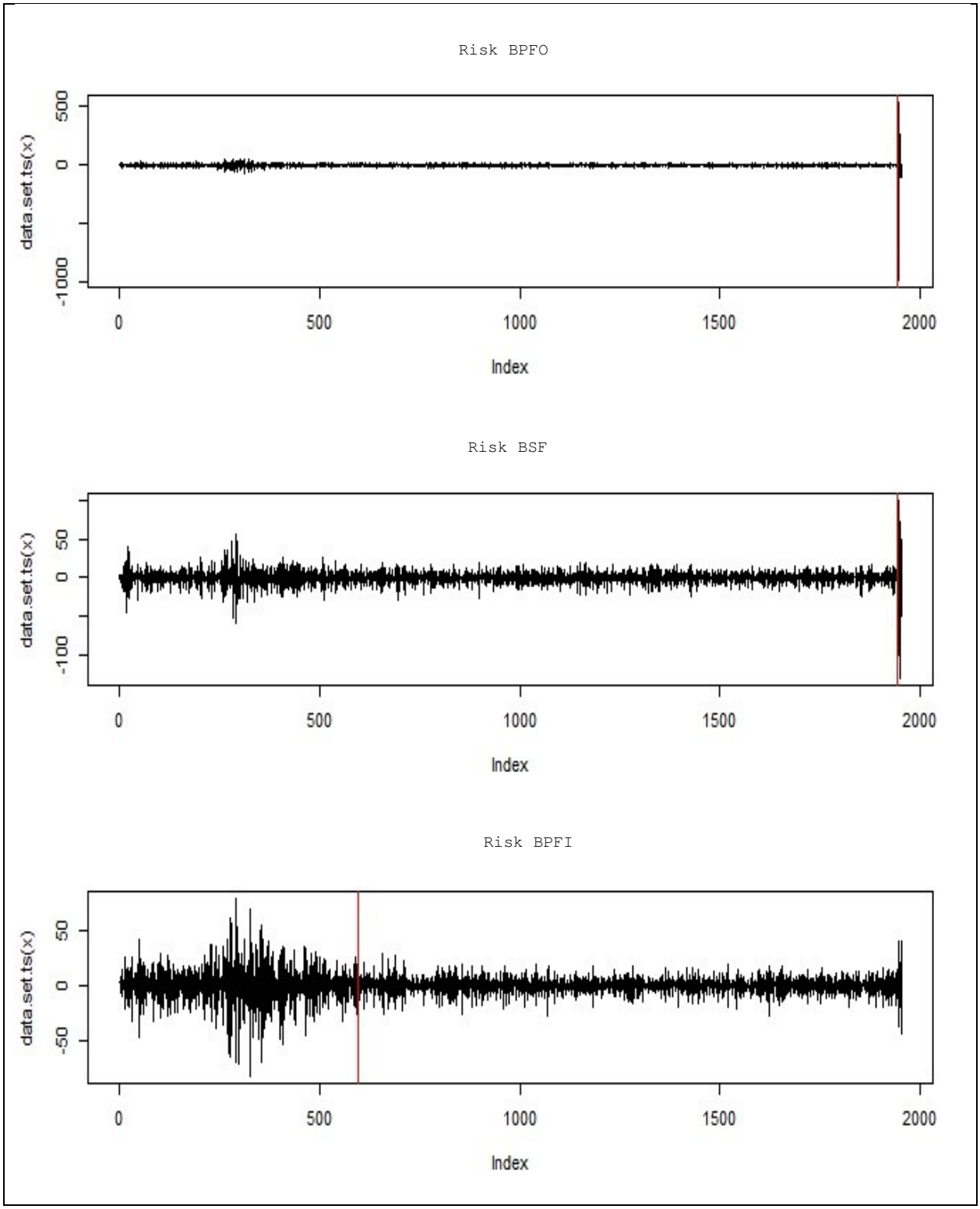


Figure 13.3a: Risks detected by package *changepoint* in Case Study Dataset 3

The plots for the risks detected by package *strucchange* (Figure 13.3b) also reveals the disturbance in the vibration from the beginning of the operation of the bearing. The risks were detected around time index 1656. Again, the package detected some of the noisy vibrations such as the detection at time indices 454 and 907 for risk BSF and BPFI respectively.

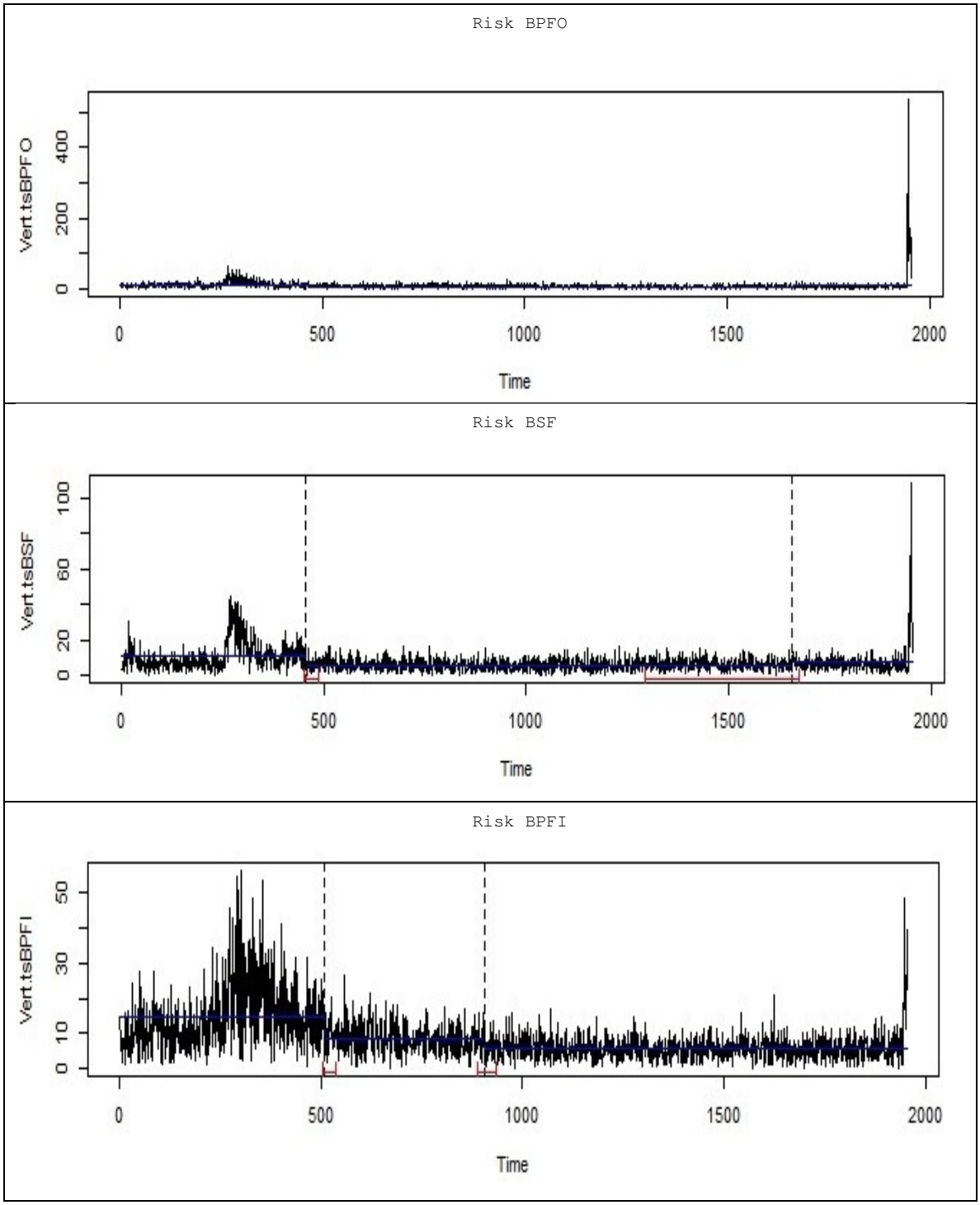


Figure 13.3b: Risks detected by package *strucchange* in Case Study Dataset 3

The plot of RMS of the bearing lifecycle (Figure 13.3c) reveals a sharp rise with spikes in the trend of both variables from time index around 250 - 500 and between time index 650 – 1945, which seem to confirm the issues of the bearing highlight suspected issues of the bearing and provides explanations to why the time indices for the risks detected appear almost similar. As a result, the study investigated relationships and interaction effect for each defect at change-point 1656 (for the lower end) and 1945 (for the upper end).

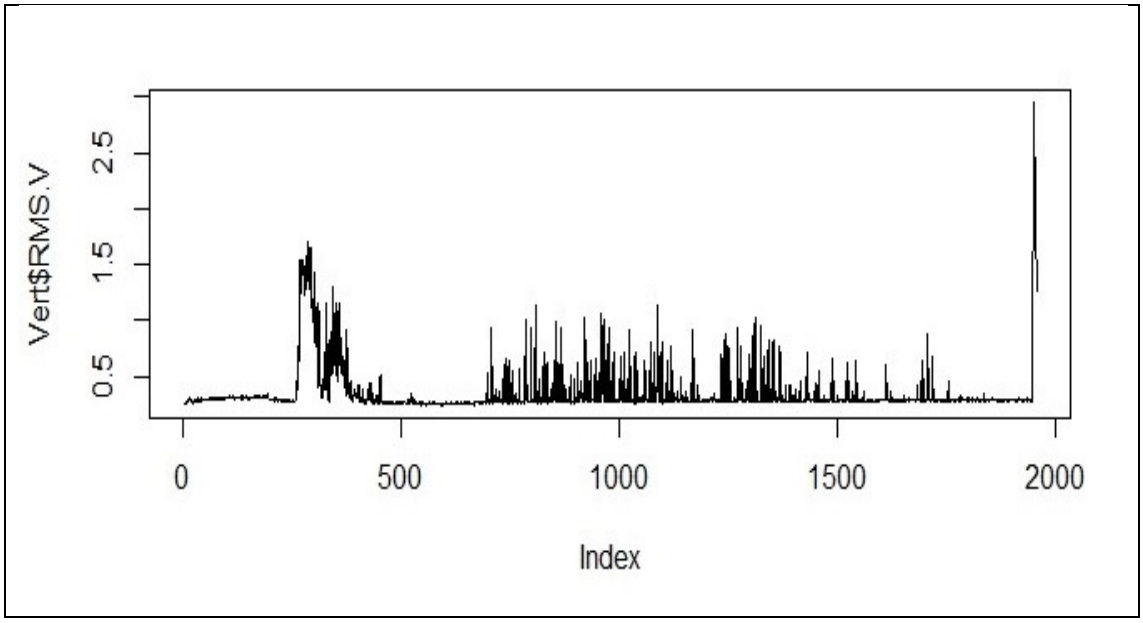


Figure 13.3c: Plot of RMS of data for bearing lifecycle for Case Study Dataset 3

The study also verifies the time indices of the risks detected using information from a publication about the data reveals that the actual predicted RUL (β) was 410 sec (Wang, 2012). From the notes of the data, the expected number of test files in the dataset is 19970. However, 19550 test files were determined in the dataset by the study. The study therefore calculated the predicted RUL as:

$$Predicted\ time\ (sec) = \quad \frac{19970 - 410}{19970} \times 19550$$

This gave a predicted time index of for the start of the RUL as 19149 sec which is within 1.5% of the time index 19450 of the risks detected by package *changepoint* as the main risk of failure. Again, the risks detected by package *strucchange* at time index around 16550 are early signal or the low threshold of risk of failure. Wang also reveals that the predictions were ranked after conversion into percent errors which suggests that relative uncertainties of the predictions may be too wide compared to the expected practical results.

The process being analyse has one bearing component in operation. As a result, the study investigates the interaction between the risk feature to ascertain whether there is any association effect. The plots of the correlations for the interactions up to time indices of the risks detected (Figure 13.3d) shows low but significant correlations between the three type of risks under investigation. The correlation between BPFO and BSF appear strongest followed by the correlation between BPFI and BPFO, then the relationship between BPFI and BSF.

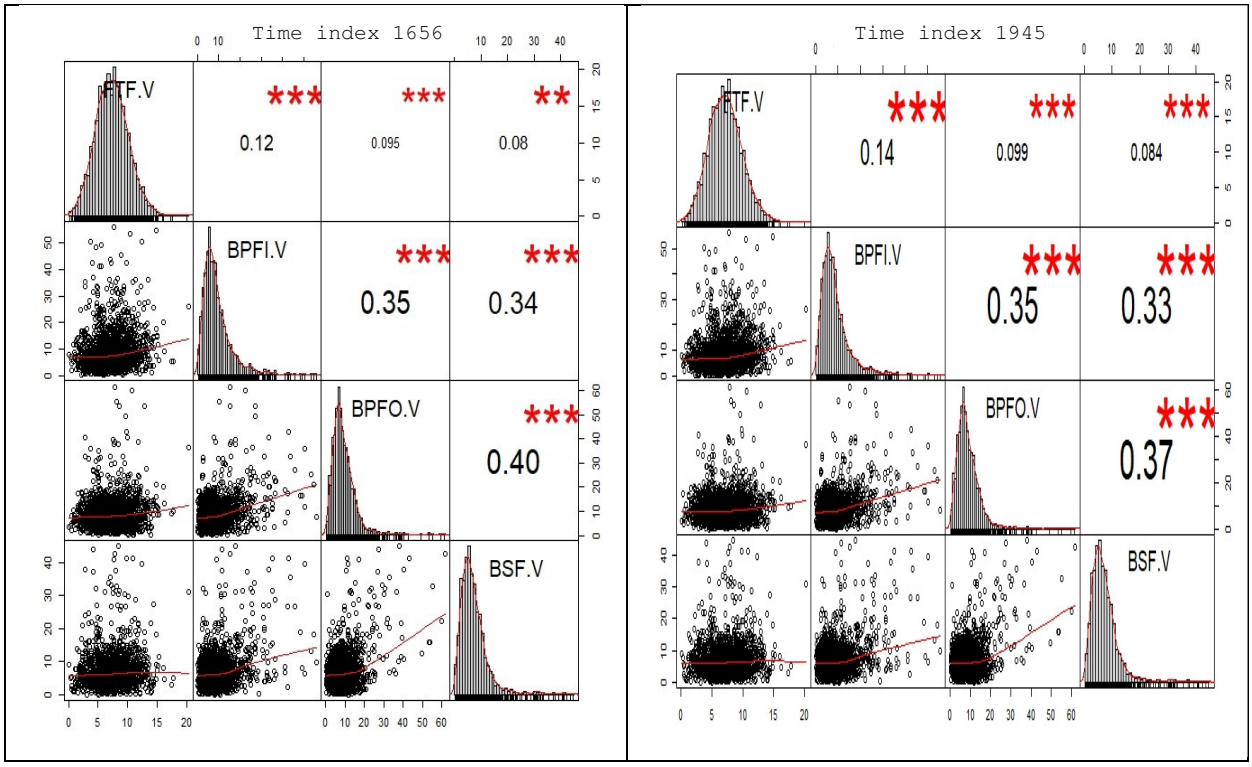


Figure 13.3d: Correlation for interactions up to time index of the risks detected in Case Study Dataset 3.

When the correlations were investigated with the decision tree model, a similar information about the dominant feature were obtained. As an example, the study provides Figure 13.3e for the interactions up to tow time indices which reveals that:

- For risk associated with BPFO, BSF is the most dominant feature followed by BPFI at lower levels of the BSF.
- For risk associated with BSF, BPFO is the most dominant feature followed by BPFI at lower levels of the BPFO.
- For risk associated with BPFI, BPFO is the most dominant feature followed by BSF at lower levels of the BPFO.

Owing to these findings, the study applied BSF as the predictor and BPFI as the moderator in the regression model to further probe the contribution of the other features to risk associated with BPFO through their interactions up to the time index of the risk detected. For the contributions to risk associated with BSF, the study applied BPFO as the predictor and BPFI as the moderator. Finally, the study applied BPFO as predictor and BSF as the moderator for the contribution from the other features to risk BPFI.

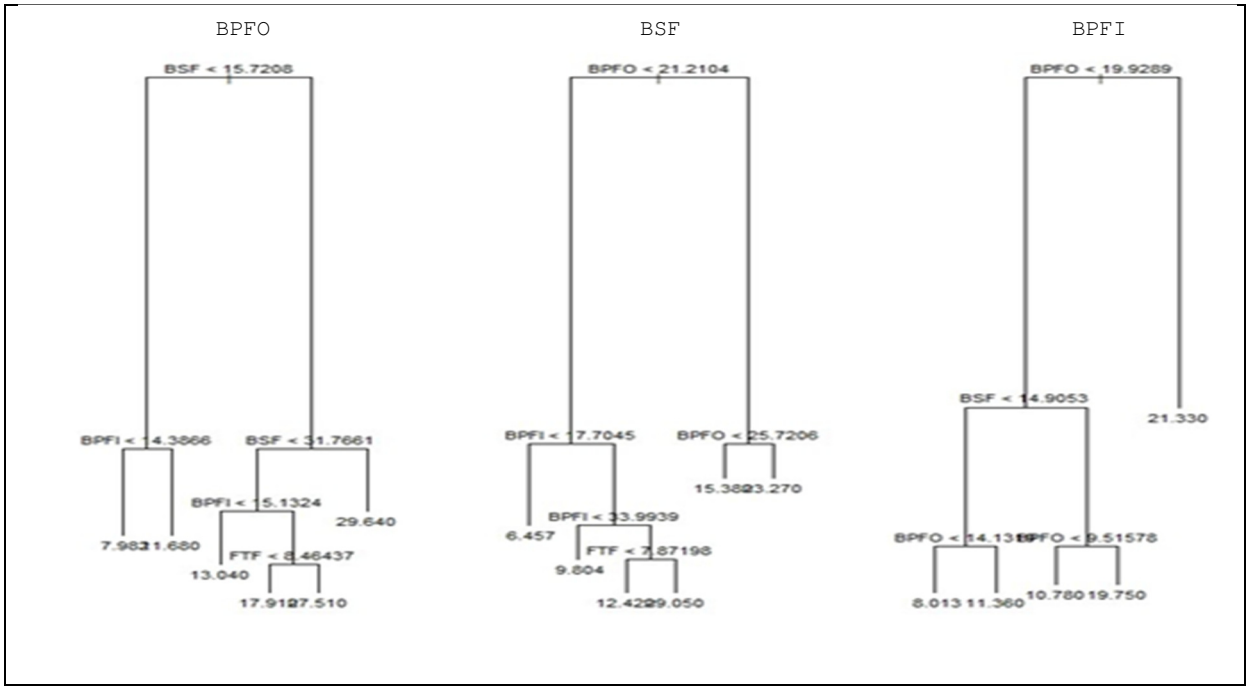


Figure 13.3e: Tree plot for the interactions up to time index of risks detected in Case Study Dataset 3.

The regression output for the interactions up to time index 1656 (Table 13.3a) reveals that:

- In the absence of the other risk features, the average vibration for risk BPFO of the bearing is 4.497 Hz which increases significantly by 0.365 Hz or 0.211 Hz per unit increase in the vibrations of BSF or BPFI respectively (model 1). For the interaction model (model 2), the average vibration for risk BPFO is 4.922 Hz which increases significantly by 0.315 Hz per unit increase in the vibrations for BSF. However, there is a significant moderation from the vibrations for BPFI which causes the vibrations for BPFO to increase in the 0.176 Hz per unit increase in moderation of the BPFI. There is no noticeable significant BSF-BPFI interaction effect.
- In the absence of the other risk features, the average vibrations for risk associated with BSF of the bearing is 3.150 Hz which increases by 0.276 Hz or 0.176 Hz per unit increase in the vibrations of BPFO or BPFI (model 3). The interaction model (model 4) reveals that in the absence of the other risk features, the average vibrations for risk associated with BSF is 4.333 Hz which increases significantly by 0.162 Hz per unit increase in the vibrations associated with BPFO. However, there is evidence of moderation from the vibrations associated with BPOFI which causes the vibrations for BSF to increase by 0.073 Hz per unit increase effect of the moderation. There is evidence of a significant two-way BPFO-BPFI interaction effect which cause the BSF to increase by 0.008 Hz per unit increase in interaction.

- In the absence of the other risk features, the average vibrations for risk associated with BPFI is 4.247 Hz which increases significantly by 0.320 Hz or 0.289 Hz per unit increase vibrations associated with BSF or BPFO respectively (model 5). The interaction model (model 6) reveals that in the absence of the other risk features the average vibrations for the risk associated with BPFI is 3.236 Hz which increases significantly by 0.376 Hz per unit increase in the vibrations associated with BPFO. There is evidence of moderation from the vibrations associated with BSF which causes the BPFI to increase by 0.424 Hz per unit in the moderation. Additionally, there is evidence of a significant two-way BSF-BPFO interaction effect which causes the vibrations associated with BPFI to decrease by 0.007 Hz per unit increase in interaction.

Table 13.3a: Output for interactions up to time index 1656 of risks detected in Case Study Dataset 3

	Dependent variable:					
	BPFO		BSF			BPFI
	model.1	model.2	model.3	model.4	model.5	model.6
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	4.497***	4.922***	3.150***	4.333***	4.247***	3.236***
	(0.259)	(0.381)	(0.232)	(0.358)	(0.313)	(0.480)
BSF	0.365***	0.315***			0.320***	0.424***
	(0.027)	(0.042)			0.032)	(0.049)
BPFO:BSF						-0.007***
						(0.002)
BPFO			0.276***	0.162***	0.289***	0.376***
			(0.020)	(0.033)	(0.028)	(0.042)
BPFI	0.211***	0.176***	0.176***	0.073**		
	(0.020)	(0.031)	(0.018)	(0.030)		
BSF:BPFI		0.003				
		(0.002)				
BPFO:BPFI				0.008***		
				(0.002)		
Observations	1,656	1,656	1,656	1,656	1,656	1,656
R2	0.211	0.212	0.207	0.216	0.172	0.176
Adjusted R2	0.210	0.210	0.206	0.214	0.171	0.174
Residual Std. Error	5.606 (df = 1653)	5.604 (df = 1652)	4.873 (df = 1653)	4.847 (df = 1652)	6.564 (df = 1653)	6.551 (df = 1652)
F Statistic	220.613*** (df = 2; 1653)	147.961*** (df = 3; 1652)	215.508*** (df = 2; 1653)	151.419*** (df = 3; 1652)	171.849*** (df = 2; 1653)	117.588*** (df = 3; 1652)
Note:						*p<0.1; **p<0.05; ***p<0.01

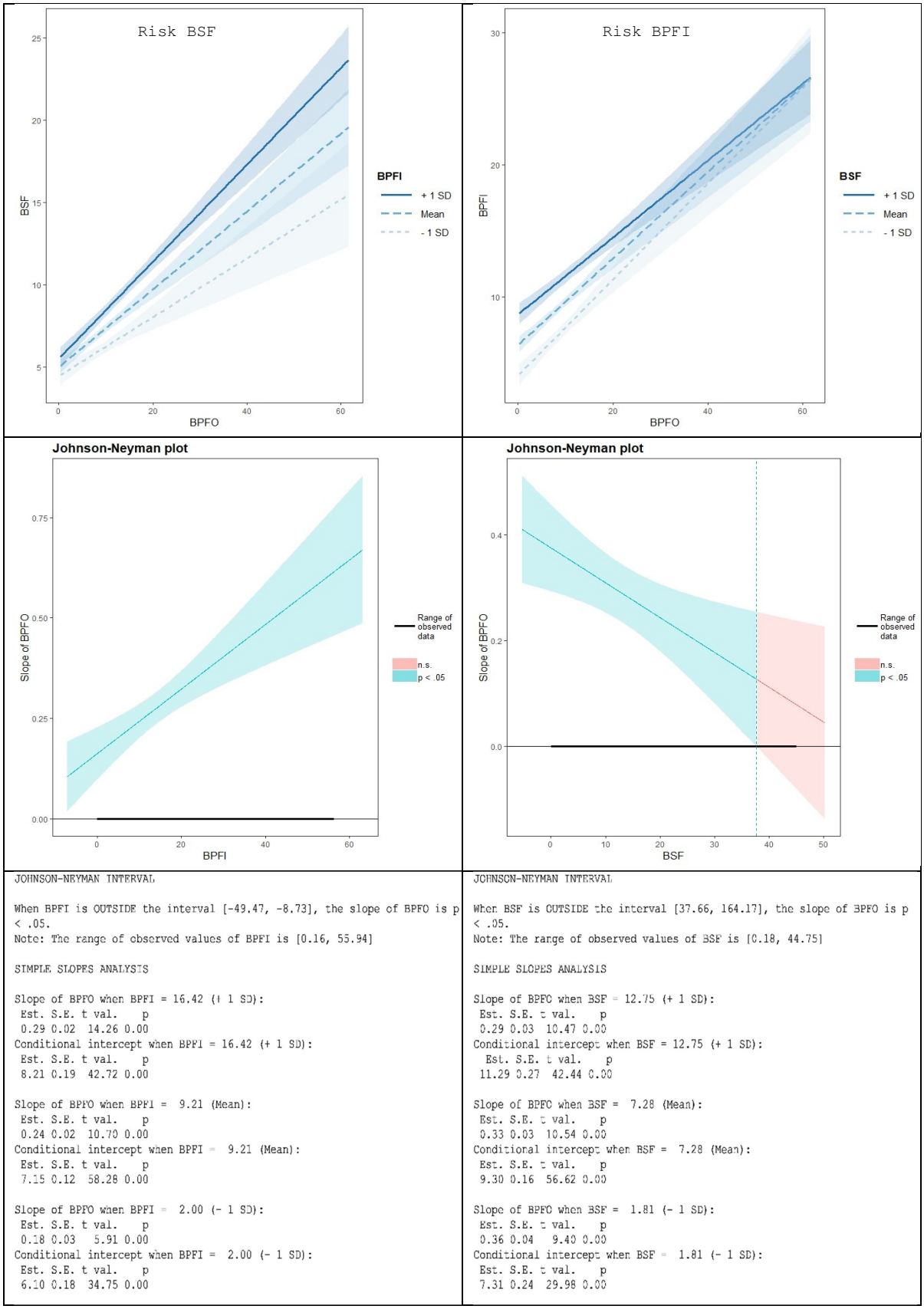
The output of the Type II ANOVA reveals that the models for the investigation of the significant interactions using ANOVA type II test (Table 13.3b) reveals moderation and interactions of model 4 and model 6 up to the time index of 1656 are statistically significant. As a result, the study probes the significant interactions with effect plots, Johnson-Neyman plots and simple-slope analysis.

Table 13.3b: Output of ANOVA Test for Significant Interactions for Case Study Dataset 3

```
Analysis of Variance Table
Model 4: BSF ~ BPFO * BPFI
Model 6: BPFI ~ BPFO * BSF
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
4   1652 38816  1    438.01 18.642 1.671e-05 ***
6   1652 70893  1    329.54 7.679 0.005649 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effect plots, J-N plots and simple-slope analysis for the significant interactions for the risks associated with BSF and BPFI obtained are presented as Figure 13.3f. The effect plots reveal that the model fits better at the 95% confidence level when the values of the frequencies from the vibrations for the risk associated with both BSF and BPFI approaches the one standard deviation above the mean than that of the mean and one standard deviation below the mean.

The Johnson-Neyman plot for model 4 reveals that the conditional slope of the values of the frequencies of the vibrations for the risk associated with BPFO increases when the frequencies of the vibrations for the risk associated with BPFI moderating up to the time index of the risk event detected increases. The print from the simple-slope analysis reveals that this effect of moderation is statistically significant for the range of values of BPFI observed. The trend in the conditional intercept appears to correlate with the conditional slope. The print also reveals that the conditional intercept increases when the BPFI increases. This suggest that any increase in BPFO for a high BPFI observation will lead to an increase in the risk associated with ball spinning frequency (BSF).




```
JOHNSON-NEYMAN INTERVAL

When BPFI is OUTSIDE the interval [-49.47, -8.73], the slope of BPFO is p
< .05.
Note: The range of observed values of BPFI is [0.16, 55.94]

SIMPLE SLOPES ANALYSIS

Slope of BPFO when BPFI = 16.42 (+ 1 SD):
 Est. S.E. t val.    p
 0.29 0.02  14.26 0.00
Conditional intercept when BPFI = 16.42 (+ 1 SD):
 Est. S.E. t val.    p
 8.21 0.19  42.72 0.00

Slope of BPFO when BPFI =  9.21 (Mean):
 Est. S.E. t val.    p
 0.24 0.02  10.70 0.00
Conditional intercept when BPFI =  9.21 (Mean):
 Est. S.E. t val.    p
 7.15 0.12  58.28 0.00

Slope of BPFO when BPFI =  2.00 (- 1 SD):
 Est. S.E. t val.    p
 0.18 0.03   5.91 0.00
Conditional intercept when BPFI =  2.00 (- 1 SD):
 Est. S.E. t val.    p
 6.10 0.18  34.75 0.00
```

```
JOHNSON-NEYMAN INTERVAL

When BSF is OUTSIDE the interval [37.66, 164.17], the slope of BPFO is p
< .05.
Note: The range of observed values of BSF is [0.18, 44.75]

SIMPLE SLOPES ANALYSIS

Slope of BPFO when BSF = 12.75 (+ 1 SD):
 Est. S.E. t val.    p
 0.29 0.03  10.47 0.00
Conditional intercept when BSF = 12.75 (+ 1 SD):
  Est. S.E. t val.    p
 11.29 0.27  42.44 0.00

Slope of BPFO when BSF =  7.28 (Mean):
 Est. S.E. t val.    p
 0.33 0.03  10.54 0.00
Conditional intercept when BSF =  7.28 (Mean):
 Est. S.E. t val.    p
 9.30 0.16  56.62 0.00

Slope of BPFO when BSF =  1.81 (- 1 SD):
 Est. S.E. t val.    p
 0.36 0.04   9.40 0.00
Conditional intercept when BSF =  1.81 (- 1 SD):
 Est. S.E. t val.    p
 7.31 0.24  29.98 0.00
```

Figure 13.3f: Effect plot, J-N plot and simple-slope analysis for interactions up to time index 1656 of risks detected in Case Study Dataset 3.

The Johnson-Neyman plot for model 6 reveals that the conditional slope of the values of the frequencies from the vibration for the risk associated with the BPFO decreases when the frequencies of the vibrations for the risk associated with the moderation of BSF increases. However, the print from the simple slope analysis reveals that this is only significant when the values of BSF are less than or equal to 37.66 Hz. This suggests that the moderation of BSF will have no effect on the contribution from the vibrations associated with BPFO to the vibrations associated with BPFI when its vibration values associated with BPFO exceed 37.66 Hz.

The print from the simple-slope analysis also reveals that the conditional intercept increases as BSF increases. This suggest that while BPFO decreases when values of the vibrations associated with BSF increases, the conditional intercept increases with increasing in values of the vibrations associated with BSF. Thus, any increase in the vibrations associated with BPFO for a low value of the vibrations associated with BSF observation will tend towards being equal on the values of vibrations associated with BPFI.

The information form Johnson-Neyman plots and print from the simple-slope analysis of the interaction models for risks associated with BSF and BPFI (models 4 and 6 respectively) led the study to conclude that there is evidence of exhibition of system complexity in the features interacting up to the time index of 1656. Thus, the risks BSF and BPFI detected in the operations of the bearing were not isolated but a contribution from other risk features. This also led to the suspicion that the bearing in operation may be defective prior to its installation on the process system.

For the interactions up to time index 1945, the regression output (Table 13.3c) also reveals that:

- In the absence of the other features, the vibrations for the risk associated with BPFO is 4.458 Hz which increase significantly by 0.338 Hz or 0.220 Hz per unit increase in the vibration of BSF and BPFI respectively (model 1). The interaction model (model 2) reveals that in the absence of the other features, the average BPFO of the bearing is 5.113 Hz which increases significantly by 0.261 Hz per unit increase in BSF. However, this increase in influenced by moderation from the vibrations associated with BPFI which causes the vibration associated with BPFO to increase by 0.163 Hz per unit increase moderation. Additionally, there is a two-way significant BSF-BPFI interaction effect which causes the average vibrations associated with BPFO to increase by 0.005 Hz per unit increase in the vibrations associated with the interaction.

- In the absence of the other features, the vibrations for the risk BSF of the bearing is 3.406 Hz which increase significantly by 0.253 Hz or 0.175 Hz per unit increase in the vibrations of BPFO or BPFI respectively (model 3). The interaction model (model 4) reveals that in the absence of the other features the average vibrations for the risk associated with BSF is 4.776 Hz which increases significantly by 0.119 Hz per unit increase in the vibrations associated with BPFO. There is evidence of significant moderations from the vibrations associated with BPFI which causes the vibrations associated with BSF to increase by 0.049 Hz per unit increase of in moderation. Also, there is a two-way BPFO-BPFI interaction effect which causes the average vibrations associated with BSF to increase significantly by 0.010 Hz per unit increase in the vibrations associated with the interaction.
- In the absence of the other features, the average vibrations associated with risk BPFI is 3.872 Hz which increase significantly by 0.312 Hz or 0.213 Hz per unit increase in the vibrations of BSF and BPFO respectively (model 5). The interaction model (model 6) also reveals that the average vibrations associated with risk BPFI 3.217 Hz which increases significantly by 0.350 Hz per unit increase in vibrations associated with BPFO. There is evidence of significant moderation from the vibrations associated with BSF which causes the BPFI to increase by 0.380 Hz per unit increase in moderation. Again, there is evidence of a significant two-way BPFO-BSF interaction effect which causes the average BPFI to decrease by 0.004 Hz per unit increase in the vibrations associated with the interaction.

Table 13.3c: Output for interactions up to time index 1945 of the risks detected in Case Study Dataset 3

	Dependent variable:					
	BPFO		BSF		BPFI	
	model.1 (1)	model.2 (2)	model.3 (3)	model.4 (4)	model.5 (5)	model.6 (6)
Constant	4.458*** (0.234)	5.113*** (0.345)	3.406*** (0.206)	4.776*** (0.317)	3.872*** (0.281)	3.217*** (0.432)
BSF	0.338*** (0.025)	0.261*** (0.039)			0.312*** (0.029)	0.380*** (0.045)
BPFO:BSF						-0.004** (0.002)
BPFO			0.253*** (0.019)	0.119*** (0.030)	0.293*** (0.025)	0.350*** (0.038)
BPFI	0.220*** (0.019)	0.163*** (0.029)	0.175*** (0.017)	0.049* (0.028)		
BSF:BPFI		0.005*** (0.002)				
BPFO:BPFI				0.010*** (0.002)		
Observations	1,945	1,945	1,945	1,945	1,945	1,945
R2	0.195	0.198	0.187	0.200	0.168	0.170
Adjusted R2	0.194	0.197	0.186	0.199	0.167	0.168
Residual Std. Error	5.429 (df = 1942)	5.421 (df = 1941)	4.694 (df = 1942)	4.657 (df = 1941)	6.274 (df = 1942)	6.269 (df = 1941)
F Statistic	235.529*** (df = 2; 1942)	159.704*** (df = 3; 1941)	223.081*** (df = 2; 1942)	161.830*** (df = 3; 1941)	196.070*** (df = 2; 1942)	132.251*** (df = 3; 1941)
Note:					*p<0.1; **p<0.05; ***p<0.01	

The output of the Type II ANOVA (Table 13.3d) reveals that all three interaction modeal are statistically significant. As a result, the study probs the significant interaction further using effect plots, Johnson-Neyman plots and simple-slope analysis.

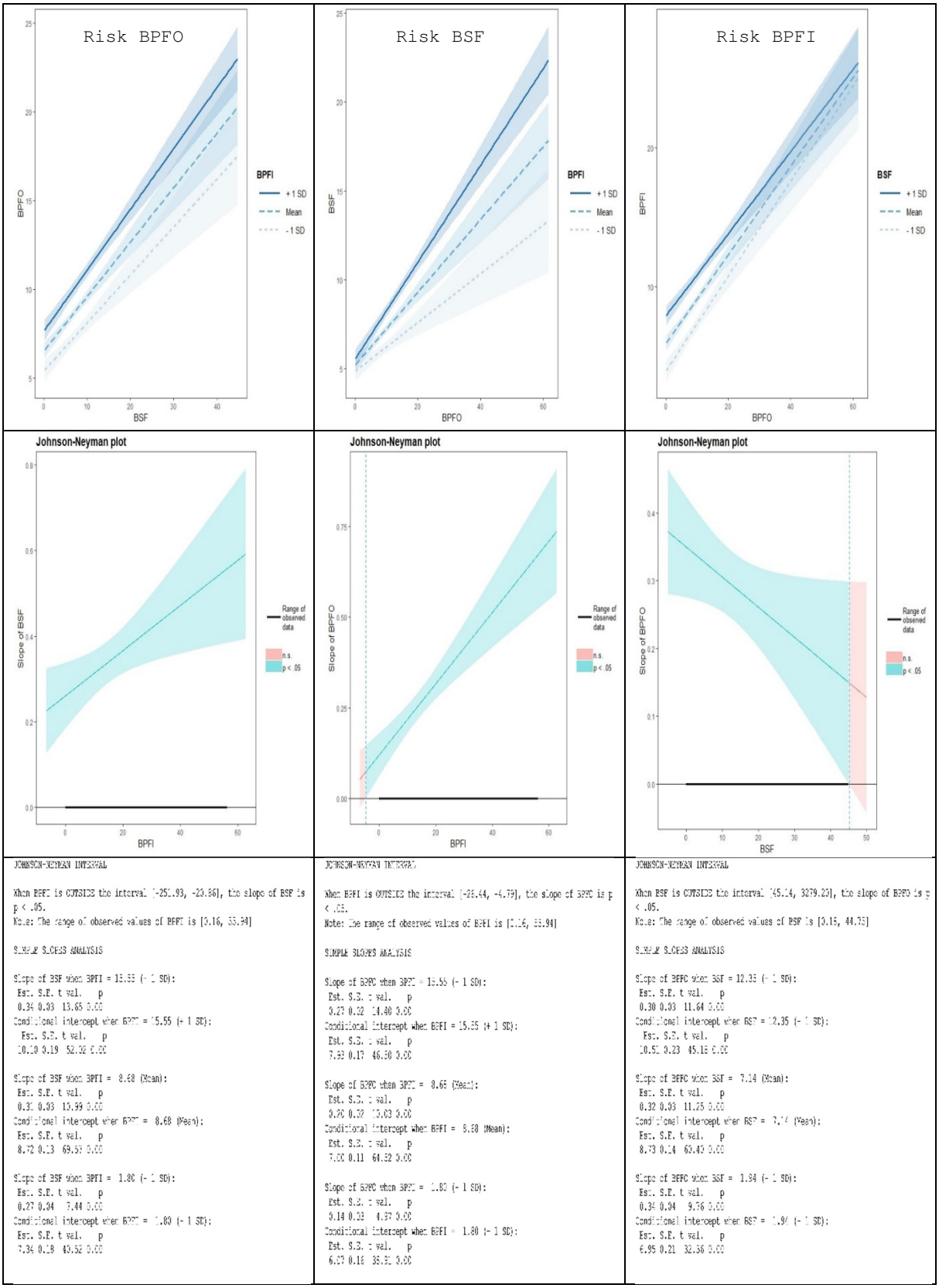
Table 13.3d: Output of ANOVA Test for Significant Interactions

```
Analysis of Variance Table

Model 2: BPFO ~ BSF * BPFI
Model 4: BSF ~ BPFO * BPFI
Model 6: BPFI ~ BPFO * BSF
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
2   1941 57038  1    196.23 6.6778 0.009835 **
4   1941 42089  1    697.54 32.168 1.626e-08 ***
6   1941 76276  1    157.44 4.0064 0.04547 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The outcome of the further probe into the interactions up to time index 1945 is presented as Figure 13.3g. The effect plots reveal that the models of the one standard deviation above the mean fits better than that of the mean and one standard deviation below the mean.

The Johnson-Neyman plot for risk BPFO (model 2) reveals that the conditional slope of the of the values of the frequencies of the vibrations associated with BSF increases as the frequencies of the vibrations associated with BPFI increases. This is significant for the range of the observed values of BPFI. The print from the simple-slope analysis reveals that the conditional intercept also increases as the frequency of the vibrations associated with BPFI increases. This suggests that any increase in the frequency of vibrations associated with BSF caused by an increase in the frequency of the vibrations associated with BPFI through the moderation from the vibrations associated with BPFI or BSF-BPFI interaction effect will cause an increase in the vibrations associated with risk BPFO.



JOHNSON-NEYMAN INTERVAL

When BPFI is OUTSIDE the interval [-251.93, -20.86], the slope of BSF is p < .05.
Note: The range of observed values of BPFI is [0.16, 33.94]

SIMPLE SLOPES ANALYSIS

Slope of BSF when BPFI = 15.55 (- 1 SD):
Est. S.E. t val. p
0.34 0.03 13.65 0.00
Conditional intercept when BPFI = 15.55 (- 1 SD):
Est. S.E. t val. p
10.10 0.19 52.32 0.00

Slope of BSF when BPFI = 8.68 (Mean):
Est. S.E. t val. p
0.31 0.03 13.99 0.00
Conditional intercept when BPFI = 8.68 (Mean):
Est. S.E. t val. p
8.72 0.13 69.57 0.00

Slope of BSF when BPFI = 1.80 (- 1 SD):
Est. S.E. t val. p
0.27 0.04 7.44 0.00
Conditional intercept when BPFI = 1.80 (- 1 SD):
Est. S.E. t val. p
7.34 0.18 40.52 0.00

JOHNSON-NEYMAN INTERVAL

When BPFI is OUTSIDE the interval [-28.44, -4.79], the slope of BPFO is p < .05.
Note: The range of observed values of BPFI is [0.16, 33.94]

SIMPLE SLOPES ANALYSIS

Slope of BPFO when BPFI = 15.55 (- 1 SD):
Est. S.E. t val. p
0.27 0.02 14.40 0.00
Conditional intercept when BPFI = 15.55 (+ 1 SD):
Est. S.E. t val. p
7.93 0.17 46.30 0.00

Slope of BPFO when BPFI = 8.68 (Mean):
Est. S.E. t val. p
0.20 0.02 10.03 0.00
Conditional intercept when BPFI = 8.68 (Mean):
Est. S.E. t val. p
7.00 0.11 64.32 0.00

Slope of BPFO when BPFI = 1.80 (- 1 SD):
Est. S.E. t val. p
0.14 0.03 4.37 0.00
Conditional intercept when BPFI = 1.80 (- 1 SD):
Est. S.E. t val. p
6.07 0.16 36.81 0.00

JOHNSON-NEYMAN INTERVAL

When BSF is OUTSIDE the interval [45.14, 3279.23], the slope of BPFO is p < .05.
Note: The range of observed values of BSF is [0.15, 44.73]

SIMPLE SLOPES ANALYSIS

Slope of BPFO when BSF = 12.35 (- 1 SD):
Est. S.E. t val. p
0.30 0.03 11.64 0.00
Conditional intercept when BSF = 12.35 (- 1 SD):
Est. S.E. t val. p
10.51 0.23 45.15 0.00

Slope of BPFO when BSF = 7.14 (Mean):
Est. S.E. t val. p
0.32 0.03 11.25 0.00
Conditional intercept when BSF = 7.14 (Mean):
Est. S.E. t val. p
8.73 0.14 60.40 0.00

Slope of BPFO when BSF = 1.94 (- 1 SD):
Est. S.E. t val. p
0.34 0.04 9.76 0.00
Conditional intercept when BSF = 1.94 (- 1 SD):
Est. S.E. t val. p
6.95 0.21 32.56 0.00

Figure 13.3g: Effect plots, J-N plots and simple-slope analysis for interactions up to time index 1945 for Case Study Dataset 3

The Johnson-Neyman plot for model 4 reveals that the conditional slope of the values of the frequency of the vibration for the risk associated with the BPFO increases as the frequency of the vibrations associated with risk BPFI increases. However, the print of the simple-slope analysis reveals that this increase is significant for the range of values observed. The conditional intercept also increases as the frequency of the vibrations associated with BPFI increases. This suggests that any increase in the frequency of vibrations associated with risk BPFO caused by an increase in the frequency of the vibrations associated with BPFI through the effect of the moderation of BPFI or the BPFO-BPFI interaction will cause an increase in the risk associated with BSF.

Finally, the Johnson-Neyman plot for model 6 reveals that the conditional slope of the values of the frequency of the vibrations associated with risk BPFO decreases when the frequencies of the vibrations associated with the moderation of BSF increases. This is only significant when the values of the frequencies associated with BSF are less than or equal to 44.73 Hz. This suggests that the moderation of BSF will have no effect on the prediction by BPFO when its vibration values exceed 44.73 Hz. The print from the simple-slope analysis also reveals that the conditional intercept increases as BSF increases. This also suggest that while frequency of the vibrations associated with BPFO decreases when frequencies of the vibrations of associated with BSF increases, the conditional intercept increases with increasing BSF. Thus, any increase in BPFO for a low frequency of the vibration of BSF observation will tend towards being equal on the risk BPFI vibration values.

13.4. Applying the Method Procedure to Case Study Dataset 4

The plots obtained when package *changepoint* was applied to the data (Figure 13.4a) reveals that:

- Risk BPFO was detected at time index 1189.
- Risk BSF was detected at time index 1089.
- Risk BPFI was detected at time index 1090.

There appears to be some disturbance in the vibrations from the onset of the process operations which leads to the suspicion that the bearing might not be secured properly within the process system or may be suffering from some form of defect. Careful consideration of the plots reveals that, the defect may be in the order of BSF followed by BPFI then BPFO.

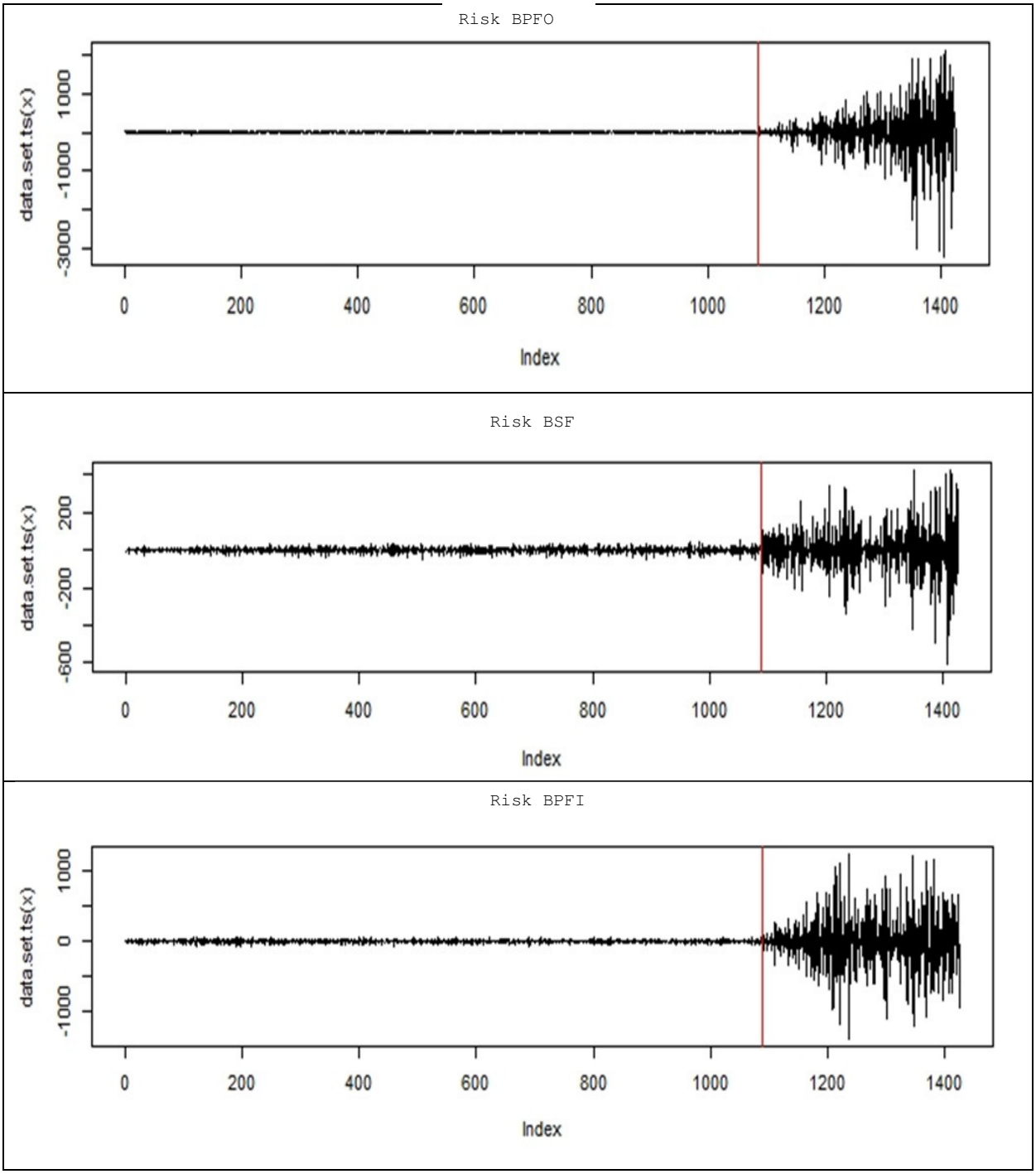


Figure 13.4a: Plots of risks detected by package *changepoint* Case Study Dataset 4

Figure 13.4b is the output of the plots obtained for the risks detected by package *strucchange* which reveals that:

- Risk BPFO was detected at time index 1190.
- Risk BSF was detected at time index1144.
- Risk BPFI was detected at time index 1151.

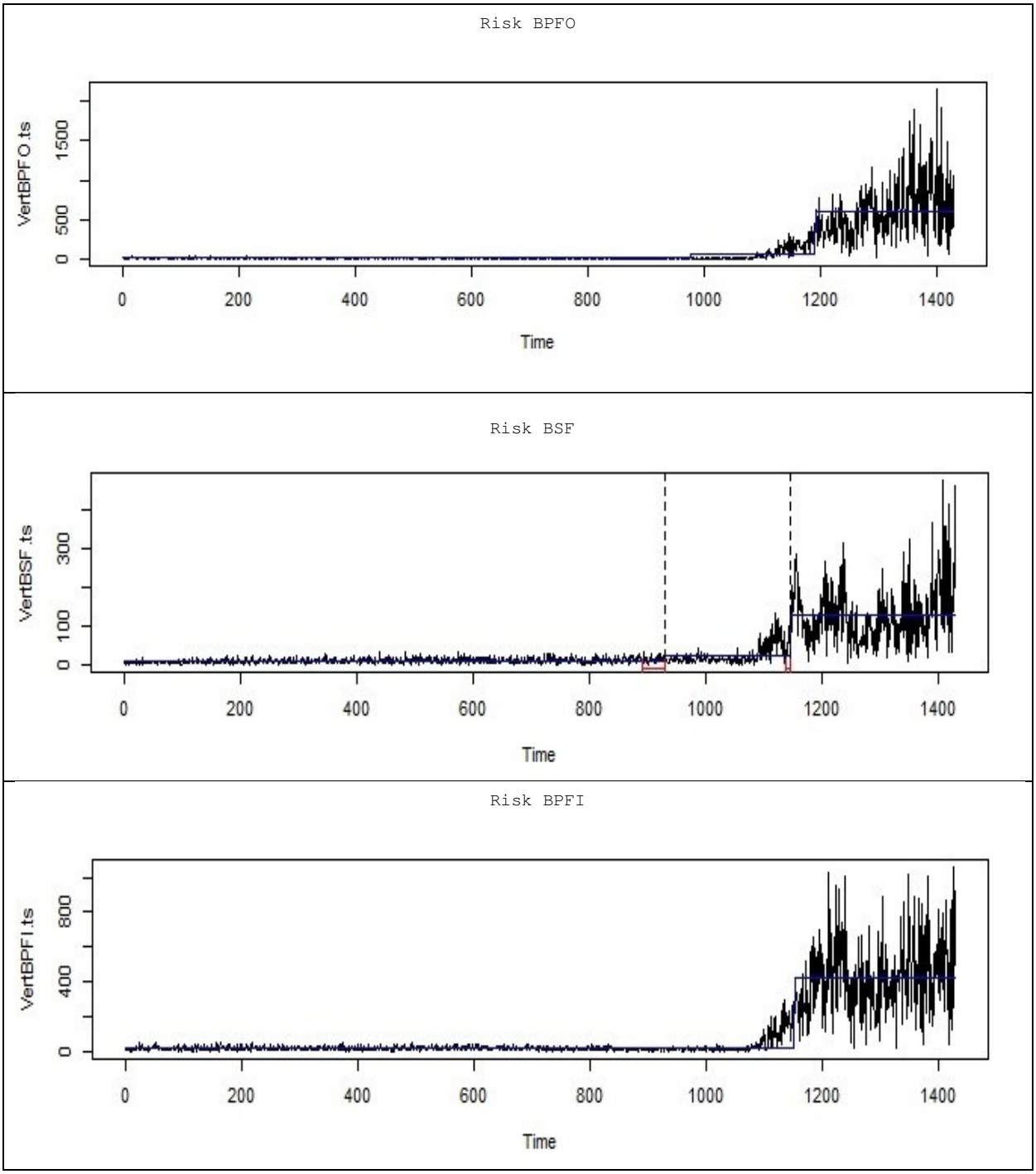


Figure 13.4b: Plots of risks detected by package *strucchange* in Case Study Dataset 4

The close proximity of the times for the risks detected by the two change-point packages confirms that the bearing was suffering from some defect or not properly secured within the process system. The plot of the RMS of the lifecycle of the bearing (Figure 13.4c) shows a study rise from around time index of the risks detected by the two packages (1089 – 1190). Due to the proximity of the time indices of the risks detected, the study selects time index 1089 for investigation of interaction effect.

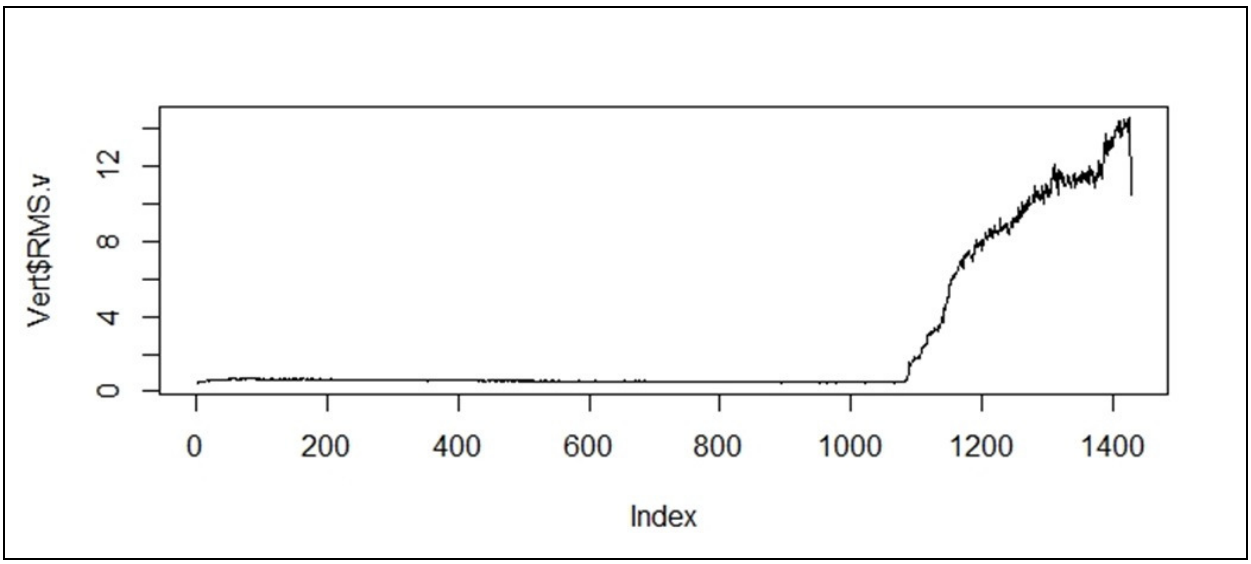


Figure 13.4c: Plots of trend in RMS of lifecycle of the Bearing in Case Study Dataset 4

The correlations plots obtained for the interactions up to change-point 1089 (Figure 10.4d) reveals positively skewed distributions with no significant correlations except correlations which have small p-values because of the size of N. The study therefore concludes that there is no evidence of significant interactions up to the time index of the risks detected. As a result, no further investigation was performed.

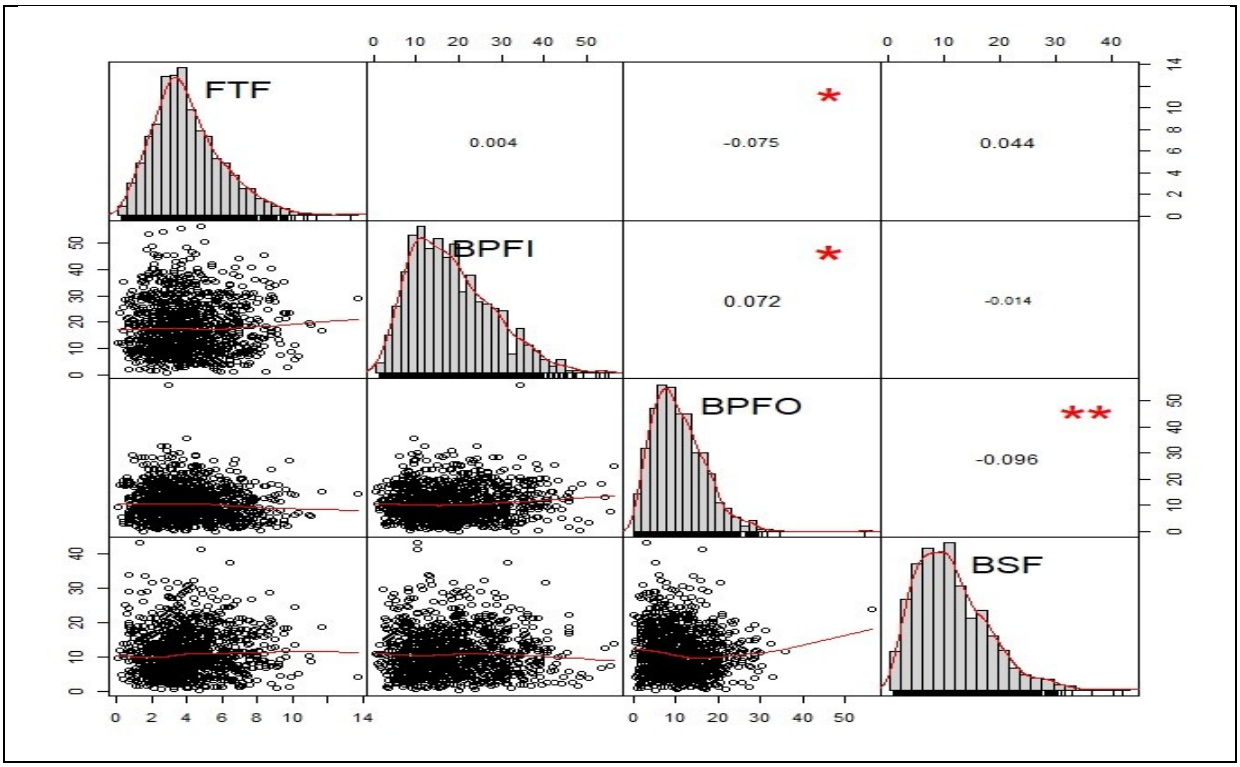


Figure 13.4d: Correlations for interactions up to change-point 1089 of the risks detected in Case Study Dataset 4

13.5. *Conclusion*

This chapter has provided the results of the validation of the procedure of the big data method and the outcome when the method was applied examples. Next is Part 5 – where the study discusses the findings of the research and provides the conclusion and recommendations.

Part 5

Discussion, Conclusion and Recommendation

Part 5 - Background

In Chapter 1, the study introduces the research and laydown the research questions and objectives. This was followed by Part 1- Risk, where the study discusses risk as a concept and theories, followed by the selection of contingency theory as the appropriate theoretical framework for the research. The study then proceeds to Part 2 -Literature Review and Systematic Content-analysis, where it reviews of research publication relating to the focus topic to find gap in knowledge and establish if the research can provide some contribution to science and/or practice. In Part 3 – Methods, the study investigates big data techniques and PC software packages using one of the available datasets to obtain the step-by-step procedure for the big data QRA method. In Part 4 – Data Analysis, the study validated the big data QRA method using two Case Study Datasets and tested the method with two other Case Study Dataset as applied examples to show the validity and applicability of the method. This was followed by a presentation of the findings as the last chapter of Part 4.

Part 5 – consists of Chapter 14 - Discussion, where general discussion of the research findings in relation to the research objectives, research questions and the contributions of the research to science and practice in the field of QRA. This will be followed by Chapter 15 – Conclusion and Recommendation, where the study provides a conclusion of the research, research limitations and provide suggestions on how future research in this area can be continued.

Chapter 14 - Discussion

14.0. Introduction

The last chapter of Part 4 provide the findings of validation of the method using two case study datasets and applied examples using another two datasets. In this chapter, the study provides a discussion of the results obtained from the method validation and the applied examples as part for the final presentation of the thesis. This will be followed by Chapter 15 where the study provides a conclusion and recommendation. The general framework from this and the last chapters is illustrated by the flowchart of Figure 14.

The first step of the overall approach will involve a recall of the research questions and provide that answers obtained from different chapters of the thesis. This will be followed by the thesis summary which briefly summarises the thesis after which the research limitations will be provided. The study will then proceed to the provide suggestion for direction of future research and closing remarks.

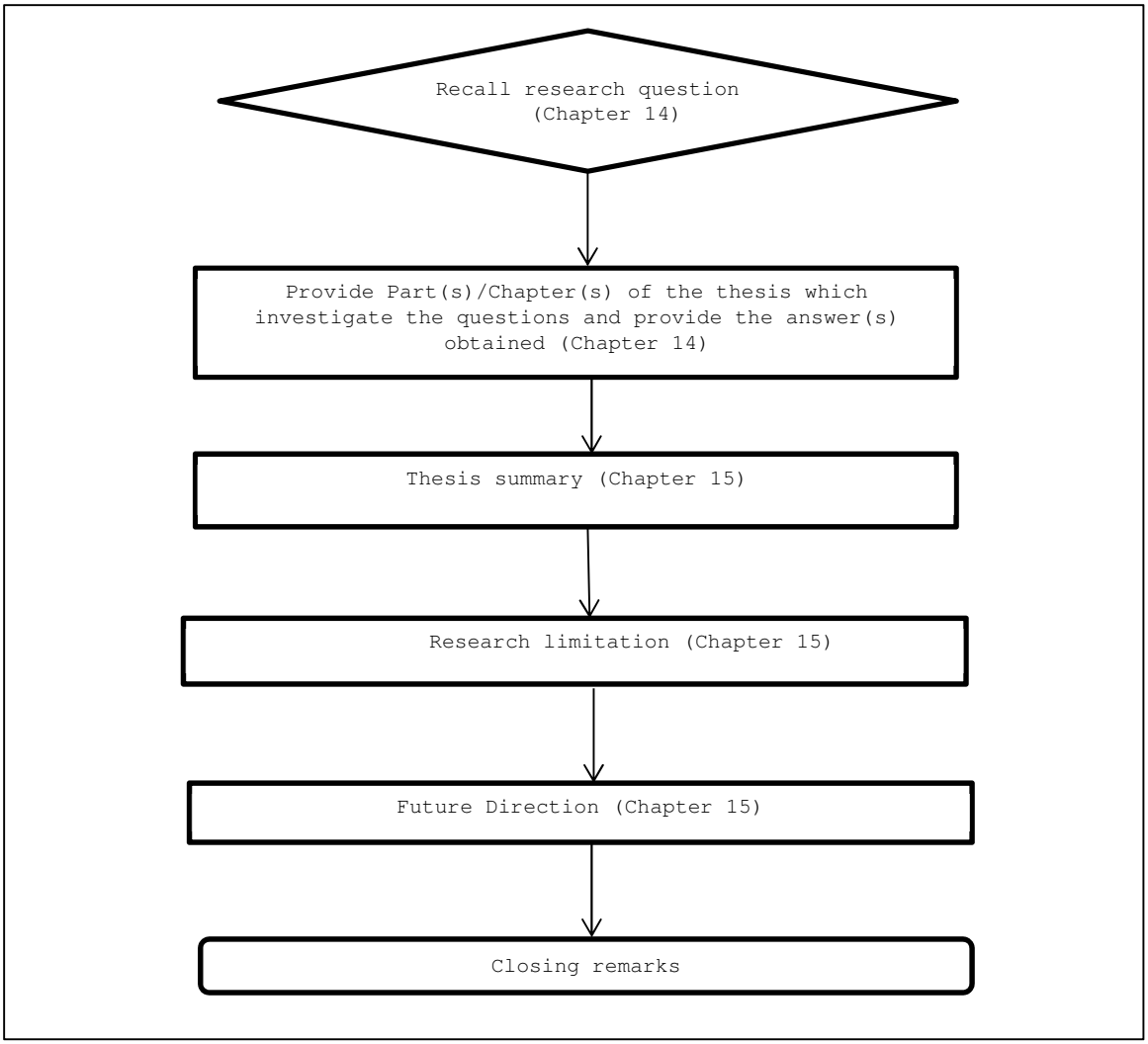


Figure 14: Flowchart illustrating general approach to Part 5

14.1. *Recall of Research Questions*

The study recalls the research questions as below:

14.1.1. The Research Questions

- How is the knowledge about risk in the HHPIs utilized and which of the risk theories can the study adopt for this research?
- Is behavioural safety programs (BSPs) more effective at preventing major accidents in the HHPIs for the focus of risk to be on monitoring human elements instead of monitoring the process itself?
- What are the existing QRA methods applied in the HHPIs and what are some of the challenges encountered with their use?
- Is there any evidence of existing review and systematic content-analysis of published research on the use of big data techniques and real-time data for QRA in the HHPI?
- How are big data techniques and data from the process operation applied for QRA and what are some of the challenges to overcome and practical solutions?
- Can big data techniques and real-time data be applied to obtain an effective QRA method for use in the HHPI?

These research questions were investigated in in different parts of the thesis and the answers are discussed below.

14.1.2. How is the knowledge about risk in the HHPIs utilized and which of the risk theories can the study adopt for this research?

The study considered factors such as the nature of operations of the HHPIs, their geographical locations, the process itself and products, or any combination of these factors to answer this research question. Considering the aforementioned factors, it could be accepted that each risk is unique and therefore must be managed in accordance with its specific characteristics and within a specific period. Since the focus of this study is to obtain a QRA method which relies entirely on the use of big data techniques and process operation data for risk analysis in an industry which has unique characteristics, it could be accepted that there cannot be ‘one best way’ to manage risk from the target industry.

Owing to this, the study evaluated various theoretical and conceptual frameworks from which contingency theory was found to be more appropriate for the for this research. Contingency theory best fits the goal of the research because it has a correlation with the focus of the research in that it does not oppose alternative pathways that might be more appropriate for each specific contingency. Although some experts in safety science and safety professionals may have objections to the

concept of contingency theory in PSM, its suitability for this research is based on risk-based concept which is also the focus of the research.

As a risk management concept, the study found that contingency analysis (CA) has been applied to minimise risk in the HHPI's and industries like nuclear, oil and gas, aviation, healthcare, and in the event of other emergencies (Everdij and Blom 2016). A typical CA procedure involves identifying all potential accidents and elevating adequacies of emergency measures. Since the big data method produced by this research has been proven to (a) be valid and reliable at detecting risk, (b) have the ability to distinguishing between low and high threshold of risk within a typical process system and thereby eliminating the issues of false alarm, and (c) have the ability to detect any communications between the system components and contribution to any risk detected within the process, the study believes that the contingency theory has been fully applied.

14.1.3. Is behavioural safety programs (BSPs) more effective at preventing major accidents in the HHPIs for the focus of risk to be on monitoring human elements instead of monitoring the process itself?

Evaluating the effectiveness of BSP at minimising accidents in the HHPIs, the study found that there have been several reviews of various aspects of process safety management (PSM) and BSP since Control of Major Accident Hazards (COMAH) Regulations 1999 to present. For instance, Bell and Healey (2006) reviewed existing literature on effectiveness of behavioural and relevant control measures in prevention of major hazard incidents, Patel and Sohani (2013) reviewed existing QRA methods used in PSM, and Besserman and Mentzer (2017) reviewed literature on process safety regulations and its enforcement globally. The study found that although the application of BSPs appears to have improve the awareness of safety for frontline personnel within the HHPIs, its effectiveness at reducing incidents does not depend on the human elements alone but also that of technical aspect of the programs.

It was also established that not enough research has been done on BSPs and where research has been conducted, very small amount of data was used. As a result, there is awareness of behavioural change but the overall effectiveness of the programs at minimising the occurrences of accidents has not been well explored. The insufficient evidence of the effectiveness of BSPs raises some concern of bias at its implementation in the HHPIs. It also shows that the programs are not addressing the purpose for which the implementation was done or a combination of these.

It is evident that though future implementations of the BSPs may reduce the risk of incidents, there is the need for further assessment of the effect of each of the BSP parameters using large sample sizes. There is also the need to reduce the over reliance on reports which usually consider the main

effects of the BSP methodologies and its perceived reliability at reducing incidents but fail to measure the validity of the method.

The review of literature on the BSPs therefore reveals that focus on human elements are less effective at minimising risks. The review also found evidence that BSP has both human behavioural elements as well as elements of the process such as application of mechanical and technological elements at the process installations. As a result, the study concludes that the focus of safety must be on the process itself instead of behavioural elements of human activities at the process sites.

14.1.4. What are the existing QRA methods applied in the HHPIs and the challenges encountered with their use?

Upon review of peer reviewed publication on existing QRA methods, the study found two publications (Tixier et. al. 2002; Patel & Sohani 2013) which has extensively reviewed the existing QRA methods. From the two reviews, it was established that the existing QRA methods has been classified into three categories based on the type of '*output data*' as below:

- Deterministic methods - QRA methods which incorporate data from the process, products, and quantification of consequences.
- Probabilistic methods – QRA methods which incorporate data such as probability or frequency of incident, with the focus on failure probability of equipment or equipment components.
- Combined deterministic and probabilistic methods – QRA methods which involves combining the deterministic and probabilistic methods to investigate the entire process site.

The study also found from reviews that, the procedures for the selection of any of the existing QRA methods are prescriptive based on the '*user expected'* outcome and '*available data'*. A flow chart which explains the procedures involve in the use of the methods for a QRA has been adapted by the study and presented as Figure 2.6a of Chapter 2 p. 17. Because the methods are prescriptive, there their application are not flexible like the dynamic method obtained by this research. This is because their application for risk analysis are usually

- Based on expert judgement, several assumptions which are accepted by peers of safety professionals and engineering although these assumptions are not validated as part of the risk analysis process.
- Knowledge based i.e. based on knowledge and contribution from peers of engineering and safety professionals.
- Generally, use single case study validation in cases where the risk analysis method was validated.

- Generally, uses data from external sources instead of real-time data or historical data from the process itself or process with similar characteristics in the case of new process.

For instance, selecting a method to apply in a hypothetical situation involves the 'safety expert' considering the expected outcome based on previous knowledge, experience and certain assumptions as the initial step. This is followed by reading through the result column of the chart to select '*expected output data'*, then the predefined method, before the selection of '*input data'* required from a list of suitable data for the analysis.

A careful consideration of this process provides knowledge of some evidence of expert bias in the application of the existing QRA methods. This is because the validity of most of the multiple assumptions applied cannot be verified. It is generally argued by safety professional and engineers that they use single case studies to validate the methods when applied for risk analysis hence the methods are fit for purpose. However, it has been long been explained by Cavaye (1996) that although the use of single case study validation models provides a satisfactory relationship and the approach to provide generalisable theory about the validity and applicability of a method, the procedure does not necessarily define a priori constructs and the relationships. Thus, for the validation to be deemed successful, multiple case study validation must be applied as in the case of the validation of the big data QRA method obtained by this research.

The study also found from the reviews of literature on the application of the existing QRA methods that most of the data use either generated by '*expert knowledge*' or obtained from literature sources. In most situations only one dataset is applied for the risk analysis. It was therefore inferred from the approach use for selecting data for use by the existing QRA methods and the source of the data used that there can be issues of objectivity in the outcomes of the risk analysis which could make the outcomes somewhat questionable. Even for risk analysis in which the existing methods applies data from small-scaled experimental investigation or process operation data, very small sample (data) size were used. As a result, the robustness of the method and/or the outcome are not well explored.

14.1.5. Is there any evidence of existing review and systematic content-analysis of published research on the use of big data techniques and real-time data for QRA in the HHPI?

The study performed a general search for previous publication on review and content-analysis and found evidence of extensive review of the existing QRA methods but not reviews relating to the focus topic. Some of the search terms, phrases, and operators used in most of the modified search strings were found to produce results which are not published peer reviewed research articles. Several attempts were made but no article relating to the focus topic was found. This could be that most researchers in the area:

- do not perform literature reviews and content-analysis prior to commencing their research.
- perform the literature reviews but without systematic content analysis.
- do not apply big data techniques in their QRA method.
- do not use real-time dataset or data from the process itself.

Surprisingly the search for existing literature review and content-analysis on the focus topic from the chemical industry incidents databases did not produce any result. However, there is a suspicion that this may be due to insufficient reporting of incidents so the QRA methods applied for the risk analysis do not include none captured in the reports. It was also revealed that some of the bibliographies of the incident reports found are marked "confidential" which made it extremely difficult to ascertain whether a literature review and content-analysis were performed before the application of the QRA method.

14.1.6. How are big data techniques and data from the process operation applied for QRA and what are some of the challenges to overcome and practical solutions?

During the literature review and content-analysis stage of the research, the study found from assessing quality of the performance of the existing QRA methods that there are issues with data limitations because of critical data voids. This was suspected to have arisen from

- the daunting tasks of updating data,
- data validity issues,
- issue of uncertainties and assumptions associated with data,
- data analysis and statistical techniques applied by the methods.

These issues make the QRA process very time-consuming because the application of the method usually requires peer involvement and several unvalidated assumptions which leads to expert bias and sometimes reduce the reliability of the methods. It was also revealed that most of the method apply data obtained from various literature sources, event and incident databases, or experimental simulation data, instead of data generated from the process operation itself. Additionally, some of the data use were captured from surveys, case studies, field studies or a combination of any of these. Owing to this, small sample size datasets are used with the application of fewer statistical analysis techniques.

During the investigation of big data techniques as part of the process to obtain the big data QRA method in Part 3 – Methodology, the study found several challenges associated with data including sampling time and sampling rate which accompany almost all five datasets. This means that, quality control of the data and, investigation into the validity of the data prior to its use for QRA using big data techniques is a very important step. These will provide some understanding about the data, so

that any corrective measures required can be undertaken prior to risk analysis. It will also help eliminate assumptions in the use of the data and thereby producing more reliable outcomes.

It was also discovered from the investigation that PC computational power and memory capacity of software packages struggle to handle all the large data generated from the process operation. However, these challenges can be overcome by using other big data techniques to help with the management and handling of the data. Some of these techniques adopted by this research detailed in Part 3 – Methodology includes:

- Applying data compression methods such as principal component analysis (PCA) and factor analysis (FA) to compresses large datasets into a manageable size by extracting few important variables to provide the direction of the most dominant variance and give indications of the unwanted events within the process system. Fortunately, PCA and FA has been applied in previous research and found to be successful (Russell, 2000; Imtiaz, 2007).
- Applying big data tools available on various PC software platforms to help handle big datasets there by alleviating some of the problems. The big data investigated which can be recommended by this study include those on R language platform like packages '*readtable*', '*bigmemory*' *'biganalytics'* and '*big.tabulate'*; Other data handling tools and services includes '*MapReduce'* for creating data-intensive applications and their deployment on clouds; and cloud computing services like '*Microsoft Azure'* and '*Amazon Web'*.
- Applying feature extraction technique (depending on the domain area) to extract the most dominant features from the data without losing important information. For instance, this research found a feature extraction technique using fast fourier transform (FFT) algorithm as more suitable for compressing the datasets and this was use prior to data analysis.

The study therefore concludes that risk analysis using the existing QRA methods rarely applies big data techniques or real-time datasets. And though the use of big data and real-time data for QRA comes with some challenges, these challenges can be overcome, and this has been proven by the study in Part 3- Methodology. Additionally, this research has proven that the use of real-time data and big data techniques would make the QRA methods more robust and reliable. Besides, activities and process operations within the HHPIs generate deluge of data but it appears the industry is not data ready. However, applying big data techniques as part of QRA like the one obtained by this research which uses process operation datasets will help reduce the extra cost of conducting experimental simulations, applying unvalidated assumptions and time required for a QRA process.

14.1.7. Can big data techniques and real-time data be applied to obtain an effective QRA method for use in the HHPI?

As part of the investigations to obtain the big data QRA method, several big data techniques were investigated with the Training Dataset as detailed in Part 3- Methodology of this thesis. From that investigation, the study found that the following big data techniques can be applied to real-time data for a QRA.

- Data compression technique like Fast Fourier Transformation, principal component and factor analysis.
- Time series analysis.
- Change-point analysis.
- Pearson's correlation analysis.
- Decision tree modelling.
- Interaction effect.
- Linear regression modelling.
- ANOVA type II tests
- Effect plots.
- Johnson-Neyman plots.
- Simple slope analysis.

The study found from its investigations that for risk detection by the big data QRA method, two change-point techniques, i.e. (a) change-point by changes in the variance using the PELT algorithm in the package *changepoint* and (b) change-point by break-point in the structure of the profile of the data in package *strucchange*, on the R language platform are required. The two packages help the big data QRA method to detect and help distinguish between acceptable operational risk from main operational risk. To recap, acceptable operational risk refers to lower thresholds of the risk event, which is detected by package *strucchange*, for which normal operations can continue. The main operational risk refers to the highest threshold of the risk event, which is detected by package *changepoint*, for which a continued operation of the process could lead to a catastrophic event. All events between the two thresholds are deemed as acceptable risk.

The study also found that for the big data QRA method to identify and distinguish between the type of system being exhibited by the components of the process through communication by the interaction up to the the point of the risk events, big data techniques like plots of the correlation between paired component data and Pearson's correlation test, decision tree modelling, regression modelling, type II ANOVA test, effect plots, Johnson-Neyman plots and simple slope analysis are

required. To recap from Chapter 1, the two main system component groups in the process are system exhibiting organised simplicity and systems exhibiting organised complexity.

For clarity, the description of the two main system groups highlight above adopted from the distinctions provided by Goerlandt and Reniers (2018):

- Systems exhibiting organised simplicity, i.e. a system in which each component (subsystem, system element) acts independent of one another, hence their operations do not affect one another (no interaction effect).
- Systems exhibiting organized complexity, i.e. a system in which the system operation of the components affect one another through non-linear interactions and feedback loops (interaction effect).

The study finds from the outcomes of the method validation using two case study datasets and that of the applied examples using two other datasets, that big data techniques can be reliably applied to real-time dataset for QRA in the HHPIs. The performance of the big data QRA method shows that big data techniques, can be successfully use for QRA. The full step-by-step approach to use the big data QRA method obtained by this research, which is my major contribution to science and practice has been provided as Section 9.6in Part 3 of this thesis.

14.2. Conclusion

This chapter has provided the answers the research questions and explain that a QRA method which relies entirely on big data techniques and real-time process monitoring dataset can be successfully applied for risk analysis in the HHPI, which is a major contribution of this research to science and practice. Next is Chapter 15 – Conclusion and Recommendations, where the study provides aa summary of the research, research limitations, future directions and the closing remarks.

Chapter 15 - Conclusion and Recommendation

15.0. Introduction

In Chapter 14, the research questions were recalled by the study after which the answer to the question were provided together with the parts of the thesis where the answer to the questions were covered. The study then concludes the chapter that QRA can be performed using a method which rely entirely on big data techniques and real-time process datasets, a major contribution by the research to science and practice. This chapter provides a summary of the thesis, future directions and the final remarks of the thesis.

15.1. Thesis Summary

The aim of the research was to prove that QRA can be performed using a method which rely entirely on big data techniques and real-time process dataset. To achieve this, the study set out objectives which are covered by six research questions which has been discussed under Chapter 14. At the beginning of this study, the appropriate big data techniques which can be applied to the available dataset to obtain the big data QRA method were not readily known. As a result, the study presents risk as a concept and provide some historical example of catastrophic process safety risk events as Part 1 – Risk in this thesis to help provide clarity.

The study then investigated the existing QRA methodologies and their use of big data techniques and real-time datasets through a review and systematic content analysis of peer reviewed literature publication relating to the focus topic from which gaps in science and practice was discovered. The novel systematic review and content analysis has been presented in Part 2 of this thesis.

To successfully achieve the aim of the research, the study investigates big data techniques using one of the available datasets in Part 3 -Methodology, from which the big data QRA method was obtained and a detailed step-by-step procedure for its use was provided as Section 9.6. This process involves a consideration of real-time case histories of process safety incidents, then critiquing the incident investigation reports and its findings, which led to the justification for using the available dataset for the research. After careful consideration followed by investigations of big data techniques to ascertain their practical application for a suitable big data QRA using one of the available datasets, the study selected the techniques which were deemed suitable for use in the method.

A combination of two change-point techniques and their corresponding software packages on the R language platform (*changepoint and strucchange*) were found to be successful for use for the detection of risk events within the process systems from the data. Communication of the process components through interaction up to the time index of the risk event detected were also investigated as part of the method to provide understanding of system behaviour i.e. whether the operation of the process components leading to the detected risk events are independent of one another (i.e.

systems exhibiting organised simplicity) or the risk event is a contribution from the operations of the individual components through non-liner interactions and loops (system exhibiting complexity). This was aimed at characterising the cause of risk which commonly causes catastrophic risk in the HHPIs.

The findings of the validation of the big data QRA method obtained with two case study datasets and testing the method using two other case study datasets as discuss in Part 4, the study finds the method to be effective and reliable for use for QRA. This provides a justification that big data techniques can be applied to real-time process dataset for QRA in the HHPIs. This is because the big data QRA method obtained was able to detect the risk events within the data from process operation and establish the type of system being exhibited through communication between the component parts interacting up to the time index of the risk events detected.

Ultimately, these findings as presented in this thesis provides a fundamental understanding of how big data techniques can be effectively applied as an alternative to the current "prescriptive QRA procedures associated with rules" (Kim et. al. 2016). This indicates that a dynamic method such as the big data QRA method obtained by this research could help overcome some of the challenges relating the existing QRA methods used for risk analysis within the HHPIs, which is a major contribution by this research to science and practice.

15.2. *Primary Conclusions from Preceding Chapters*

Six primary conclusions can be drawn as a summary of the results presented in the preceding chapters.

- The findings presented throughout this study clearly demonstrate the importance of using real-time process operation data for QRA instead of data from other sources such as literature sources, single case studies data, or event and incident databases. In the face of time, cost and labour, it appears convenient to neglect the real-time data from process operation when performing QRA within the HHPIs. However, Chapter 2 shows that due to the source of some of the data used by the existing methods (e.g. surveys, case studies, field studies, experiments or a combination of all four approaches), small sample size data are generally used. With each cycle of the process operations generating a deluge of data, a QRA which applies the real-time data could be more effective at detecting risk events and other important hiding information about the process which is not the case when data from the sources other than the process itself are used. This will help reduce the cost of the QRA process because the extra cost of obtaining data from external sources and the application of unvalidated assumptions can be eliminated, thereby reducing the time required for QRA process. The lack of application of real-time data could sometimes leads to the misinterpretation of QRA results and thereby making it less robust.

- The collective results in study strongly support the use of big data techniques for QRA method because unlike the existing QRA methods the big data QRA method obtained was successful at detecting onset of risk events in five process operation datasets and the main risk event within the process operation. This was clearly demonstrated in Part 3 and Part 4, where the method was investigated, validated and tested to demonstrate its robustness and applicability for QRA. This method comprises of eleven completely distinct big data techniques has demonstrated that risk events within a typical process operation and the type of interaction between the system components up to the period of the risk events detected can be detected big data techniques are applied to real-time data from the process operation. This also clearly demonstrate that expert bias in the selection and application of the existing QRA methods can be eliminated.
- The consideration of appropriate theoretical and conceptual framework help selects contingence theory as appropriate theoretical framework for the research. This is because raw materials used by each process, their products, services, and geographical locations have unique characteristics which makes the risks in the HHPIs unique. As a result, there is no one best way to deal with process safety risk events within the industry. Thus, the contingency theory fits best for this research since it does not oppose alternative and more appropriate pathways for each specific contingency. Besides this research was conducted with a risk-based management concept which makes the contingency analysis, an aspect of contingency theory, more appropriate. The theory was successfully applied in the research because through the activities which led to obtaining the method, the method validation and applied examples, the big data QRA method obtained was proven to

  (a) be effective at detecting risk events within process systems,

  (b) have ability to distinguish between low and high threshold of risk events,

  (c) have ability to help explain how the components parts of the process made contributions to the risk through interacting communications up to the time index of the risk event risk detected.

  By detecting all potential risk events, adequacies of emergency can be elevated so that a list of all potential contingency activities which could be used to minimise contingency violations of any undesirable events at the process site can be provided.
- The results from the effectiveness of BSP for PSM evaluated in Chapter 2 reveals that although the BSP has help to improve safety awareness of frontline personnel, its effectiveness also depends on technical aspect of the process. Besides there is very little evidence that the programs have help to improve safety in the HHPIs. Thus, the focus of risk must be directed more on the process itself than the behaviour of frontline personnel. In

addition, BSP implementation also involves unvalidated assumptions based on experience of experts and peers and applies small datasets which raises come concerns of bias.

- The outcome of the systematic review and content-analysis of published research literature on existing QRA methods in relation to the use of big data techniques and real-time process monitoring datasets reveal that there are gaps in the science and practice of QRA hence alternative QRA methods such as the one obtained by this research are required. This big data QRA method obtained by this research is therefore a major contribution by the research to science and practice.
- Finally, this research has demonstrated that there are issues associated with using big data and real-time datasets for QRA. The list of issues observed by the research includes issues associated with data integrity, data validity, and PC hardware and software capabilities at processing large datasets for QRA. However, these issues can be overcome by
  (a) applying quality analysis to the data as part of data exploration process to ensure all attributes of the data are valid prior to its usage,
  (b) applying any corrective method and eliminate any unvalidated assumptions associated with the data prior to its application of the method,
  (c) applying other big data techniques including PCA and FFT to reduce large datasets to aid data processing and analysis by PC and PC software platforms without losing important information,
  (d) applying big data tools on various PC software language platforms like those on the R platform applied in this study to handle large datasets (e.g. *big.analytics*, *big.tabulate*, *read.table*) and
  (e) applying cloud computing services like Microsoft Azure and Amazon Web.

Although these findings alone do not influence the use of big data techniques and real-time data for QRA, the study hopes that the big data QRA method obtained by this research will be an important tool for future studies of QRA in the HHPIs. Further investigations of the big data QRA method may be required to help address the process safety risk events other than those relating to dust fire and explosion. For instance, the method could be applied to data obtained from other process such as chemical reaction process, petrochemical industry, transport of dangerous goods, water treatment facilities, and other non-ideal small and medium HHPIs which applied dangerous chemicals as part of their processes.

15.3. Research Limitations

This thesis acknowledges some limitations due to reasons including the limited time frame for conducting the research.

15.3.1. Research Design and Duration

The design of this research could have been different in data the study could have been a comparative design or adapted to make the research more comprehensive. Adapting to a comparative design could have to compare the big data QRA method obtained by this study with other QRA methods. It could also have adopted a longitudinal design so that the trend in the performance of the big data QRA method obtained can be evaluated to provide some clarity on how changes in process condition may impact on its performance. However, adopting these designs could prolong the research beyond the specified time frame. Moreover, considering the objectives of the research and the questions this research seeks to answer, it could be inferred that the adopted design appropriately fulfils the research goals. This is because the study investigates big data techniques with one of the process operation datasets to obtain the big data QRA method. After obtaining the method, it was validated by two Case Study Datasets to test its validity for application for risk analysis within the target industry. This was followed by testing the method with two other Case Study Datasets as applied examples to provide examples of the reliability of the method. From the outcome of the processes highlight above, it could be inferred that though the other suggested designs could have provided more insight into the research, they are not best suited for the study because they could have caused a distraction from reaching the present conclusions.

Using the current designs, the research could go further by applying the method to data from other forms of process operations like chemical reaction process data, process facility transport operations data, and electrostatic discharge data to mention a few. This will help to detect risks from various sources within specific but different process systems. And because these processes are different, the outcome of the performance of the method could be similar or different. The similarities/differences of the outcomes could then be compared with the finding from the current investigation to understand the general application of the big data QRA method for risk analysis within the target industry.

For instance, although the components of a chemical reactor include bearings, the process itself measure the amount of heat involved in the reactor. This implies that though the reactor operation provide data on thermal energy accompanying the chemical reaction within the reactor, which is the main objective for the reactor, the process also generates bearing vibration due to the operation of the bearings. A risk in the operation of the bearings or a risk of a chemical run-away reaction within the reactor can lead to a catastrophic event. Thus, the big data QRA method could be applied to the individual process monitoring parameter data i.e. to the reaction temperature or pressure data, the bearing vibration data, or a combination of all three dataset for risk analysis of the reactor. The differences or similarities could provide understanding to how risk in one type of system component affects other type of components of the same process.

15.3.2. Issues Associated with Data

Since this research was focusing on HHPIs, there are trade secrets, competitions, and other confidential issues. As a result, it was extremely difficult to obtain data from the industry. Although the outcome of this research could be beneficial to the HHPIs, requests for process operation data from the industry for this study was turndown. However, from available reports of investigations by the US Chemical Safety Board the study was able to find two sources of open source data which were deemed appropriate for used for this research. Although useful outcome was achieved from using the datasets, data from the industry itself other than open source data could have been effective for further application of the method to specific process operation parameters and systems.

Considering the findings of this research, it was established that although findings about contingency theory and contingency analysis could be generalised, its strategies and guidelines may not be applicable if the method is applied to very different set of process data other than bearing vibration data which was applied. However, the study is of the view that the big data QRA method has the potential of producing similar results when applied for risk analysis with minor adaptation based on the type of dataset being use or may produces different results if no adaptation is applied to the method. Having considered these limitations and limitations relating to different aspects of this study, the study acknowledge that suggestions can be provided for future research to improve on the science and practice in the area process safety domain.

15.4. Future Directions

As with any research of this magnitude, each conclusion leads to more questions. Fortunately, the big data QRA method obtained by this research has been validated and successfully applied to other datasets in this study and therefore provide a framework for answering some of these questions. Although most of the questions cannot be directly address, the study hopes that the following may briefly introduce some future directions for this work.

15.4.1. Big Data QRA and Existing QRA

It is evident that conditions within process plants are generally unknown, but the amount of data generated by their operations, especially in this era of automation, is enormous. Thus, the trove of data-trove of the from activities of the industry may contain hidden treasure of information which could be used for their risk analysis. Although the standards and protocols associated with the existing QRA methods are thoroughly documented and technically accurate, they are limited (Niewoehner 2008) hence they must be adapted for proper application to specific process operation.

This study has established a baseline of advanced big data QRA method by investigating how big data techniques can be applied to real-time process data to detect the onset of risk and the main risk for ideal process operations in the HHPIs. However, both advanced QRA methods such as the big data QRA method obtained by this study, and the existing QRA methods can be improved for risk analysis in more complex processes. For instance, the existing QRA methods must be adapted to specific process environment by using real-time process operation data to help minimise cost, time and the use of unvalidated assumptions and concerns of bias associated with the results. This will help the industry to take advantage of data generated from their process operations. It will also help the industry to be data ready by creating their own data repository. In tandem, application of the existing to real-time or historical data from process operations, just like the big data QRA method obtained by this study could drive the industry towards a deeper understanding of process risks which will then help in effective management of process safety.

15.4.2. Data Handling and Computational Power

Another open issue found by this research is the ability of PC and PC software to handle large data sets. However, this can be overcome using data compressing technique like FFT and PCA to extract the most important features. Although data compressing techniques were found to be successful as found in this study, it sometimes affects the visual presentation of the risk in that risks detected which are in close proximity are sometimes presented as one risk in one approximate time index in the visual plots. Such issue could have been eliminated if computation power and software memory capabilities improve in the future or by investing in to cloud computing platforms like Microsoft Azure and Amazon Web. Although these alternatives may require some additional investments, the overall outcome could help in providing insight into the outcome of the future research of big data application and real-time data for QRA.

15.5. Closing Remarks

At the beginning this study, the existing QRA methods use in the HHPIs were found to be prescriptive and involves the use of numerous unvalidated assumptions. The type of data used by the existing QRA method are obtained from literature, incident databases, and minimal application of data from the process itself. Although some new advanced methods have been investigated and applied for QRA particularly in the petrochemical industry (Kim et al., 2017), they also incorporate data from sources other than the process operations. Taking a step back to review the methods, it became clear that the existing QRA methods and the type of data used use for their application needs reconsideration.

Following years of investigating big data techniques for use in QRA and painstaking analysis, this final document demonstrates that the focus of safety must be on the process itself and not elements

of behaviour of frontline staff. The careful consideration of big data techniques including change-point analysis, machine learning like decision tree modelling, regression analysis and various method applied for analysis of interaction effect, the study obtained a big data QRA method which was validated and tested to demonstrate its applicability for risk analysis within the target industry. The outcome of the method performance in the validation and testing as applied example demonstrate that, in fact, big data can be applied to real-time process operation data for QRA in the HHPI.

The newly obtain big data QRA method, which involves big data techniques and use of real-time data represent a breakthrough in the contribution to science and practice of risk analysis. The study hopes that this big data QRA method will impact numerous real-world applications, including the safety and engineering of industrial process safety. To this, the author hopes that the insight established in this thesis will shed light on the use of abundance of historical and real-time data generated by industries for QRA.

REFERENCES

Abt, E., Rodricks, J. V., Levy, J. I., Zeise, L. and Burke, T. A. (2010) Science and decisions: Advancing risk assessment. *Risk Analysis* 30(7): 1028-1036.

Adom, D., Hussein, E. K. and Agyem, J. A. (2018) Theoretical and Conceptual Framework: Mandatory Ingredients of a Quality Research. *International Journal of scientific Research*, 7(1): 438-441.

Ageyev, D. and Qasim, N. (2015) LTE EPS Network with Self-Similar Traffic Modelling for Performance Analysis. 2015 Second International Scientific-Practical Conference Problems of Infocommunications. *Science and Technology:* 275 – 277.

Ahooyi, T., Arbogast, J., Seider, W., Oktem, U. and Soroush, M. (2016) Model-predictive safety system for proactive detection of operation hazards. *AIChE Journal* 62: 2024-2042.

AIChE (2016) Guidelines for Integrating Management Systems and Metric to Improve Process Safety Performance, Appendix A: 131 -137. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118795262.app1 Access date 20th April 2017.

Allocco, M., Bush, D., Celiktin, M., Kirwan, B., Mana, P., Mickel, J., Slater, K., Smith, B., Strater, O. and Van der Sluis. E (2016) Safety Methods Database Version 1.1. An overview of Techniques, Methods, Databases, or Models that can be used during a Safety Assessment. A database Maintained by NLR Available at http://www.nlr.nl/documents/flyers/SATdb.pdf, Access Date: 17th July 2016.

American Institute of Chemical Engineers (1995) Guidelines for Process Safety Documentation, ISBN:9780470938072

Aminikhanghahi, S. and Cook, D. J. (2017) A Survey of Methods for Time Series Change Point Detection. *Knowl Inf Syst, 51(2):* 339–367.

Amyotte, P. R. and Lupien, C. S. (2017) Elements of Process Safety Management, *Methods in Chemical Process Safety* 1: 87-148

Anderson, M (2005) Behavioural Safety and Major Accident Hazards, Magic Bullet or Shot in the Dark? *Trans IChemE, Part B, Process Safety and Environmental Protection,* **83(B2)**:109-116

Andersson, U., Cuervo-Cazurra, A. and Nielsen, B. B. (2014) From the Editors: Explaining interaction effects within and across levels of analysis. *Journal of International Business Studies* 45(9): 1063–1071.

Arif, S. N. A. M., Mohsin, M. F. M., Bakar, A. A. and Hamdan, A. R. (2017) Change point analysis: A statistical approach to detect potential abrupt change. Jurnal Teknologi (Sciences & Engineering) 79 (5): 147–159.

DOE (2004) DOE Handbook: Chemical Process Hazards Analysis. *US Department of Energy*, DOE-HDBK-1100-2004, http://energy.gov/sites/prod/files/2013/06/f1/DOE-HDBK-1100-2004.pdf, Access Date 14th December 2014.CSoCE (2004) Risk Assessment – Recommended Practices for Municipalities and Industry. *Canadian Society for Chemical Engineering*, ISBN No. 0-920804-92-6: 1-82.

ICH (2005), ICH Harmonised Tripartite Guideline prepared within the Third International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH), Text on Validation of Analytical Procedures, Step 4,

https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf, Access date 23rd August 2017.

CSB (2006) Combustible Dust Hazard Study. *US Chemical Safety Board Investigation Report Number 2006-H-1*:1- 107.

TCC (2011) Cochrane Handbook for Systematic Reviews of Interventions 5.1.0, Editors: Haggins, J.P.T & Green, S. (updated March 2011), *The Cochrane Collaboration*, www.cochrane-handbook.org, Last Access date: 1st November 2014.

HSE (2012) Failure Rate and Event Data for use within Risk Assessments, HSE Publication, http://www.hse.gov.uk/landuseplanning/failure-rates.pdf, 1-96, Download date: 20Th December 2016.

Baghouse (2012) Dust Collector Fire and Explosion Highlights Need for Combustible Dust Considerations In System Designs. Baghouse Editorial Report May 2012. Download source: https://www.baghouse.com/category/industrial-health-and-safety/, Download date: 21st June 2016.

OSHA (2012) Citation and Notification of penalty, New England Wood Pellet LCC. US Department of Labour – OSHA Citation and Corrective Working Sheet -108074: 1-28. Download source: https://www.osha.gov/ooc/citations/New-England-Wood-Pellet-108074-04-17-2012.pdf, Download date: 21st June 2016.

HSE (2014) Risk assessment: A brief guide to controlling risks in the workplace. *HSE Publication INDG163* (rev4): 1 – 5.

IAOGP (2014) SAFETY PERFORMANCE INDICATORS -2013 DATA, Fatal incidents report No. 2013sf, OGP DATA SERIES, *International Association of Oil & Gas Producers*, http://www.ogp.org.uk/pubs/2013sf.pdf, Access date 20th December 2014.

HSE (2015) The Control of Major Accident Hazards (COMAH) Regulations 2015, *HSE Guidance on Regulations L111*, Third Edition: 5-7, ISBN 978 0 7176 6605 8

Ventiv-Aon (2017) Driving the data dividend. Making use of Analytics in Risk Management. *Airmic Technical*: 1 – 13.

NFPA (2017) Standard for the Prevention of Fires and Explosions in Wood Processing and Woodworking Facilities. *National Fire Protection Association (NFPA) Regulation* 66: 1-129.

OSHA (2018a) OSH Answers Fact Sheets. Canadian Centre for Occupational Health and Safety. Source: https://www.ccohs.ca/oshanswers/chemicals/combustible_dust.html, Access date: 30th September 2018.

OSHA (2018b) Combustible Dust, Does your company or firm process any of these products or materials in powdered form? US OSHA poster of some combustible dust. Source: https://www.osha.gov/Publications/combustibledustposter.pdf , Access date: 1st September 2018.

Apostolakis, G. E. (2004) How Useful Is Quantitative Risk Assessment? *Risk Analysis* 24 (3): 515-520.

ARE (2014) Difference Between Data Mining and Machine Learning. May 2014. Access date February 2019. http://allroundexpert.blogspot.com/2014/05/difference-between-data-mining-and.html.

Aryal, G. R. and Tsokos, C. P. (2011) Transmuted Weibull Distribution: A Generalization of the Weibull Probability Distribution. *European Journal of Pure and Applied Mathematics* 4(2): 89-102.

Ashan, S. N and Sakale, R. (2014) Risk management in construction projects. *International Journal of Advances in Engineering and Applied Science* 1(3): 162 – 166.

Barondes, A. (2012) Comments and Observations on "The Science and Superstition of Quantitative Risk Assessment", *Journal of System Safety*, 48(6): 1-4.

Bastos, P., Lopes, I. and Pires, L. (2014) Application of data mining in a maintenance system for failure prediction. Safety, Reliability and Risk Analysis: 1-8.

Bellamy, L. J., Geyer, T. A. W., Astley, J. A. (1989). Evaluation of the human contribution to pipework and inline equipment failure frequencies. *HSE Contract Research Report CRR15*, HSE Books. http://www.hse.gov.uk/research/crr_pdf/1989/crr89015.pdf Access Date; 12 June 2017

Berztiss, A. T. (2004) Knowledge and uncertainty. Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004., Zaragoza, 2004: 476-480.

Besserman, J. and Mentzer, R. A. (2017) Review of global process safety regulations: United States, European Union, United Kingdom, China, *India. Journal of Loss Prevention in the Process Industries* 50: 165 -183.

Boettcher, M. (2011) Contrast and change mining. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery* 1(3): 215–230.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T. and Mehta, P. (2011) OpenMx: An Open Source Extended Structural Equation Modelling Framework. *Psychometrika* 76(2): 306-317.

BS EN 61511 (2017edition 2) Functional safety - Safety instrumented systems for the process industry sector. Guidelines for the application of IEC 61511-1: 2 - 84. http://www.cechina.cn/eletter/standard/safety/iec61511-1.pdf Access Date: May 2018.

Caesarendra, W. and Tjahjowidodo, T. (2017) A Review of Feature Extraction Methods in Vibration-Based Condition Monitoring and Its Application for Degradation Trend Estimation of Low-Speed Slew Bearing, *Machines*:2-28.

Carl, P. and Perterson, B. (2009) Performance Analysis R, Researchgate Publication, 1: 50. https://www.researchgate.net/publication/266407262_Performance_Analysis_in_R Accessed 1st Jun 2018.

Catterson, V. (2013) Understanding data science: feature extraction with R. http://cowlet.org/2013/09/15/understanding-data-science-feature-extraction-with-r.html, Access Date: 1st March 2018.

Cavaye, A. (1996). Case study research: a multi-faceted research approach for IS. *Information Systems Journal*, 6: 227-242.

CPS (2000) Guidelines for Chemical Process Quantitative Risk Analysis, Second Edition. Centre for Chemical Process Safety (CPS) of the American Institute of Chemical engineers, ISBN: 0-8 169-0720-X: 1 - 49.

CCPS (2010) Guidelines for Risk Base Process Safety, *John Wiley & Sons Inc*, ISBN: 978-0-470-16569-0, 1-4.

Chalmers, I. and Glasziou, P (2009) Avoidable waste in the production and reporting of research evidence. *Obst Gynecol*, 114(6):1341–1345.

Chambers, C., Wilday, J. and Turner, S. (2009) A review of Layers of Protection Analysis (LOPA) analyses of overfill of fuel storage tanks. *HSE Research Report* RR716: 1- 59.

Charlwood, M., Turner, S. and Worsell, N. (2004) A methodology for the assignment of safety integrity levels (SILs) to safety-related control functions implemented by safety-related electrical, electronic and programmable electronic control systems of machines. *HSE Research Report* 216, HSE Books ISBN 0 7176 2832 9: 9 – 83.

Chaudhuri, A. (2018) Predictive Maintenance for Industrial IoT of Vehicle Fleets using Hierarchical Modified Fuzzy Support Vector Machine. *A ResearchGate Publication*: 1 -18.

Chen, C.-C., Wang, T.-C., Chen, L.-Y., Dai, J.-H., Shu, C.-M. (2010) Loss prevention in the petrochemical and chemical-process high-tech industries in Taiwan *Journal of Loss Prevention in the Process Industries*, 23(4):531-538.

Chen, R.Q., 2013. Advances in data-driven monitoring methods for complex process. *In Applied Mechanics and Materials*, 423: 2448-2451.

Cheng, C-W, Yao, H-Q & Wu, T-C (2013) Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 22 (6): 1-10.

Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H. and Stoddart, K. (2016) A review of cyber security risk assessment methods for SCADA systems. *Comput. Secur*. 56, 1–27.

Chiang, L., Lu, B. and Castillo, I. (2017) Big Data Analytics in Chemical Engineering. *The Annual Review of Chemical and Biomolecular Engineering* 8:63–85.

Christenfeld, N.J, Sloan, R.P, Carroll, D. and Greenland, S. (2004) Risk factors, confounding, and the illusion of statistical control. *Psychosom Med*, 66:868–75.

Christou, M.D. and Papadakis, G.A., (1998) Risk Assessment & Management in the Context of the Seveso II Directive 6 (1): 27- 46. https://www.elsevier.com/books/risk-assessment-and-management-in-the-context-of-the-seveso-ii-directive/christou/978-0-444-82881-1 , Access Date: 1st January 2017

HSE (2011) Buncefield: Why did it happen? The underlying causes of the explosion and fire at the Buncefield oil storage depot, Hemel Hempstead, Hertfordshire on 11 December 2005. *Control of Major Accident Hazards (COMAH)/HSE Publication*: 1 -36.

Copeland, J.B. (2017) Risk Analysis vs. Risk Assessment: What's the Difference? FAIR Institute Blog publication. Source: https://www.fairinstitute.org/blog/risk-analysis-vs.-risk-assessment-whats-the-difference, Access Date 12th October 2017.

Coze, J.C.L. (2010) Accident in a French dynamite factory: An example of an organisational investigation, *Safety Science*, 48: 80-90.

Cozzani, V. and Salzano, E., (2004) The quantitative assessment of domino effects caused by overpressure: Part I. Probit models. *Journal of Hazardous Materials*, 107(3): 67-80.

CCPS (2000) Guidelines for Chemical Process Quantitative Risk Analysis, Second Edition. *Center for Chemical Process Safety (CPS) of the American Institute of Chemical Engineers* ISBN: 0-8 169-0720-X: 1 - 49.

CSB (2003) Incident data. U.S CSB *Reactive Hazard Investigation Report No. 2003-15-D*: 1 – 37.

CSB (2006) CSB report 2006-H-1, Combustible Dust Hazard Study, US CSB Combustible Dust Hazard Investigation report, http://www.csb.gov/combustible-dust-hazard-investigation/ Access date: 25 January 2017.

CSB (2009) Investigation Report: Sugar dust Explosion and Fire, Imperial Sugar Company, Port Wentworth, Georgia, CSB Investigation Report No. 2008-05I-GA, http://www.csb.gov/assets/1/19/Imperial_Sugar_Report_Final_updated.pdf, Last Access Date: 8th September 2017.

CSB (2017) Factual Investigation Update on Loy Lange Explosion, US CSB Investigation Statement Report, http://www.csb.gov/assets/1/19/Statement__-_Final.pdf , Access date: 26th May 2017.

CSB Safety recommendations to prevent the recurrence, CSB incident investigation database, http://www.csb.gov/recommendations/?F_All=y , Access Date: 20th March 2017.

CSChE (2004) Risk Assessment – Recommended Practices for Municipalities and Industry. Risk Assessment – Recommended Practices, *Canadian Society for Chemical Engineering (CSChE) publication* ISBN No. 0-920804-92-6: 1 – 82.

Cullina, T. L., Dastidar, A. G. and Theis, A. E.(2017) Case Study: A Risk-Based Approach for Combustible Dust Hazard Mitigation, Fauske & Associates, LLC Publication. Source: https://www.fauske.com/blog/case-study-a-risk-based-approach-for-combustible-dust-hazard-mitigation, Access date: 13th October 2017.

de Salis, C. F. (2012) HAZOP and LOPA the Odd Couple. IChemE Symposium Series No. 58. *Hazards* XXIII. 183 – 186.

Delacre, M., Lakens, D. and Leys, C. (2017) Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test . *International Review of Social Psychology* 30 (1): 92–101.

Deloux, E., Castanier, B. and B´erenguer, C. (2009) Predictive maintenance policy for a gradually deteriorating system subject to stress. *Reliability Engineering & System Safety* 94(2): 418-431

Deng, W. Q., Asma, S., and Paré, G. (2014) Meta-analysis of SNPs involved in variance heterogeneity using Levene's test for equal variances. *European Journal of Human Genetics* 22(3): 427–430.

Ding, S.X., 2012. Data-driven design of model-based fault diagnosis systems. *IFAC Proceedings Volumes*, 45(15): 840-847.

DOE (2004) Process Safety Management for Highly Hazardous Chemicals. *US Department of Energy Handbook*-1101-2004: 1-166.

Ebadat, V. (2017) Dust Explosion Hazards Management, Requirements of NFPA 652: Standard on the Fundamentals of Combustible Dust. Conference presentation, *AIHA NE Regional Conference Princeton*: 1 – 62.

Eftekharnejad, B. Carrasco, M.R, Charnley, B., Mba, D. (2011) The application of spectral kurtosis on Acoustic Emission and vibrations from a defective bearing, *Mechanical Systems and Signal Processing 25*: 266–284.

Eruhimov, V., Martyanov, V., Tuv, E. and Runger, G. C. (2007) Change-point Detection with Supervise Learning and Feature Selection. ICINCO 2007 - *International Conference on Informatics in Control, Automation and Robotics*: 359 – 363.

EU (2016) Regulation (EU) 2016/679 of the European Parliament and on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Document 32016R0679: 1 - 88. Source: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN. Last access Date: 7th October 2016.

Everdij, M. H. C. and Blom, H. A. P. (2016) Safety Methods Database. Netherlands Aerospace Centre NLR, 1.1: 1 – 261. Source: http://www.nlr.nl/documents/flyers/SATdb.pdf, Download date: 10th June 2018.

Fearnley, J. and Nair, S. R. (2009) Determining Process Safety Performance Indicators for Major Accident Hazards using Site Process Hazard information. *IChemE Symposium Series* No. 155, Hazards XXI: 221- 225.

Feinerer, I (2008). An introduction to text mining in R. *R News*, 8(2):19–22. *http://CRAN.R-project.org/doc/Rnews/*. Access date: 16th April 2017.

Feinerer, I. (2017) Introduction to the tm Package, Text Mining in R, 1-8, *https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf* Access date: 16th April 2017.

Fleming, M. & Lardner, R (2002) Strategies to promote safe behaviour as part of a health and safety management system*, Health & Safety Executive Contract Research Report 430*, ISBN 0 7176 2352 1.

Ford, C. (2016) Reading PDF files into R for text mining, University of Virginia library News 8 Dec 2016, *http://data.library.virginia.edu/reading-pdf-files-into-r-for-text-mining/* Access date: 14th April 2017.

Freeman, R.A. (1990) CCPS guidelines for chemical process quantitative risk analysis. *Plant/Operations Progress*, 9(4): 231-235.

Gabriel, D. (2011) Methods and methodology. *Black Bloggers UK*, May 2011. http://deborahgabriel.com/2011/05/13/methods-and-methodology/

Gadd, S., Keeley, D. and Balmforth, H. (2003) Good practice and pitfalls in risk assessment, *HSE Research Report 151*, 1-54.

Gadd, S.A., Keeley, D.M. and Balmforth, H.F. (2004) Pitfalls in risk assessment: examples from the UK. *Safety Science*, 42(9): 841-857.

Gasparrini, A. (2011) R software: advantages and opportunities. Centre for Statistical Methodology (CSM) Forum presentation. Source: https://csm.lshtm.ac.uk/wp-content/uploads/sites/6/2016/04/Antonio-14-10-2011.pdf, Access date: 20th January 2018.

Ghahramanzadeh, M. (2013). Managing risk of construction projects: a case study of Iran Doctoral dissertation, *University of East London*.

Giannini, F. M.,Monti, M. S., Ansaldi, S. P., and Bragatto, P. (2006). P.L.M., to Support Hazard Identification in Chemical Plant Design. *Innovation in Life Cycle Engineering and Sustainable Development*: 349 - 362.

Goerlandt, F., Khakzad, N. and Reniers, G. (2017) Validity and validation of safety-related quantitative risk analysis: A review. Safety Science 99 (B): 127-139.

Grace-Martin, K. (2018) What's in a Name? Moderation and Interaction, Independent and Predictor Variables, The Analysis Factor, https://www.theanalysisfactor.com/whats-in-a-name-moderation-and-interaction-independent-and-predictor-variables/, Access Date: 26th June 2018.

Graham, S. (2017) Extracting Text from PDFs; Doing OCR; All within R, Electric Archaeology website, https://electricarchaeology.ca/2014/07/15/doing-ocr-within-r/ Access date: 17th April 2017.

Graney, B. P. and Starry, P. (2011) Rolling Element Bearing Analysis. *Materials Evaluations, 70(1):* 78-85.

Grant, C. & Osanloo, A. (2014) Understanding, Selecting, and Integrating a Theoretical Framework in Dissertation Research: Creating the Blueprint for Your "House". *Administrative Issues Journal: Connecting Education, Practice, and Research,* 4(2): 12- 36.

Grote, G. (2012) Safety management in different high-risk domains–All the same? *Safety Science*, 50(10): 1983-1992.

Guldenmund, F.W. (2000) The Nature of Safety Culture: a Review of Theory and Research. *Safety Science* 34: 215 – 257.

Hancıoğlu, Y., Doğan, Ü. B. and, Yıldırım, Ş. S. (2014) Relationship between Uncertainty Avoidance Culture, Entrepreneurial Activity and Economic Development, *Procedia - Social and Behavioural Sciences* 150: 908 – 916

Harding, S. (1987) Feminism & Science: The Method Question. *Hypatia* 2 (3): 19-35.

Harding, S. (1988) Feminism and Methodology: Social Science Issues. Indiana University Press. ISBN13 9780253204448: 1 -14.

Hart, A. (2002) Case Studies on Quantitative Risk Assessment, *Central Science Laboratory Final Project Report* -PROJECT PN0920: 2-39.

Hashemian, H. M. and Bean, W. C. (2011) State-of-the-Art Predictive Maintenance Techniques, *IEEE Transactions on Instrumentation and Measurement* 60(10): 3480 – 3492

Herbert, J. G. M., Iniyan, S. and Goic, R. (2010) Performance, reliability and failure analysis of wind farm in a developing Country. *Renewable Energy* 35: 2739 – 2751.

HID Inspection Guide Offshore: Inspection of Maintenance Management, HSE UK Operation Guidance e-book: 1-25. Source: http://www.hse.gov.uk/offshore/ed-maintenance-management.pdf, Date download: 11th September 2018.

Hlavac, M. (2014) Ready-Made Regression Tables from the Stargazer Package in Statistics Education. 1-15. Download date: 20th June 2018. Available at SSRN: https://ssrn.com/abstract=2407759 or http://dx.doi.org/10.2139/ssrn.2407759,

Hlavac, Marek (2018) stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Hofstede, G. and Bond, M. H., (1984), Hofstede's Culture Dimensions: An Independent Validation Using Rokeach's Value Survey, *Journal of Cross– Cultural Psychology* 15 (4), pp. 417-433.

Hofstede, G., (2001), Culture's Consequences: Comparing Values, Behaviours, Institutions, and Organizations Across Nations (Second Edition), Thousand Oaks, California: *Sage Publications, Inc.*: 351-373.

Hong, T. and Purucker, S. T. (2018) Spatiotemporal Sensitivity Analysis of Vertical Transport of Pesticides in Soil. *Environmental Modelling & Software* 105: 24 – 38.

Hongping, C., Wenwen, Z. Xinping, Y., Peng, W., McGrath, S. P. and Fang-Jie, Z. (2018) Effective Methods to Reduce Cadmium Accumulation in Rice Grain. *Chemosphere* 207: 699 – 707.

Hou, Z. and Zhao, P., (2016) Based on Fuzzy Bayesian Network of Oil Wharf Handling Risk Assessment. *Mathematical Problems in Engineering*, 2016:1-10.

HSE (2000) Designing and operating safe chemical reaction processes. *HSE Books*, ISBN 978 0 7176 1051 8: 1 -64.

HSE (2001) Reducing risks, protecting people HSE's decision-making process. *UK Health and Safety Executive publication*, ISBN 0 7176 2151 0: 1 – 78.

HSE (2002) Strategies to Promote Safe Behaviour as Part of a Health and Safety Management System. HSE Contract Research report 430: 1- 74. Source: http://www.hse.gov.uk/research/crr_pdf/2002/crr02430.pdf. Access Date 13th May 2016.

HSE (2014) Chemical Reaction Hazards and the Risk of Thermal Runaway. *HSE Report* INDG254(rev1): 1 – 6. Source: http://www.hse.gov.uk/pubns/indg254.pdf, Access Date: 2nd February 2018.

HSE (2014) Risk assessment: A brief guide to controlling risks in the workplace. *HSE publication INDG* 163(rev4): 1- 5. Source: http://www.hse.gov.uk/pubns/indg163.pdf. Access date: 10th September 2017.

HSE (2015) The Control of Major Accident Hazards (COMAH) Regulations 2015, (L111, Third edition), *UK Health and Safety Executive Publication* L111, Third edition ISBN 978 0 7176 6605 8. http://www.hse.gov.uk/pubns/priced/l111.pdf, Access date August 2015.

HSE (2019) Inspection of Electrical, Control and Instrumentation. HSE- COMAH ECI Operational Delivery Guide v2_01: 1 – 30. Source: http://www.hse.gov.uk/eci/eci-delivery-guide.pdf, Access date 17th May 2019.

IChemE (2015) Process Safety Competency – A Model. Retrieved From, IChemE Safety Centre Publication: 1- 32. Source: http://www.ichemesafetycentre.org/~/media/Documents/icheme/Safety%20Centre/process-safety-competency.pdf, Access Date 5th June 2018.

Ingram, D. (2014) The Difference Between Risk and Loss, Willis Towers Watson Wire Bloggers, https://blog.willis.com/2014/12/the-difference-between-risk-and-loss/ Access date: 22nd March 2016.

Kadhem, A. A, Wahab, N. I. A, Aris, I, Jasni, J. and Abdalla, A. N. (2017) Advanced Wind Speed Prediction Model Based on a Combination of Weibull Distribution and an Artificial Neural Network. *Energies* 10: 1 – 17.

Kamaras, K. and Dimitrakopoulos, I. (2016) Vibration Analysis of Rolling Element Bearings using Spectral Kurtosis and Envelope Analysis, *A white paper by FNT Advanced Services Ltd,* 1:1-14.

Kerin, T. (2017) Managing Process Safety. The Core Body of Knowledge for Generalist OHS Professionals. Tullamarine, VIC. *Safety Institute of Australia*. ISBN 978 0 9808743 2 7: 1 – 64.

Khakzad, N., Khan, F. I., Amyotte, P. (2011) Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. *Reliability Engineering & System Safety*, 96(8): 925-932.

Khakzad, N., Khan, F. and Amyotte, P. (2012) Dynamic risk analysis using bow-tie approach. *Reliability Engineering & System Safety*, 104: 36-44.

Khan, F., Rathnayaka, S., Ahmed, S. (2015) Methods and models in process safety and risk management: past, present and future. *Process Safety and Environmental Protection* 98: 116–147.

Khan, K.S., Kunz, R., Kleijnen, J. and Antes, G. (2003) Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3):118-121.

Khoshroo, A., Emrouznejad, A., Ghaffarizadeh, A., Kasraei, M. and Omid, M. (2018) Sensitivity Analysis of Energy Inputs in Crop Production using Artificial Neural Networks. *Journal of Cleaner Production* 197: 992 – 998.

Kidam K, Hurme M. (2012) Origin of equipment design and operation errors, *Journal of Loss Prevention in Process Industries*, 25: 937–949.

Kidam, K., Sahak, H.A., Hassim, M.H., Hashim, H. and Hurme, M. (2015) Method for identifying errors in chemical process development and design base on accidents knowledge, *Process Safety and Environment Protection*, 97: 49–60.

Kidmo, D. K., Danwe, R., Doka, S. Y. and Djongyang, N. (2015) Statistical analysis of wind speed distribution based on six Weibull Methods for wind power evaluation in Garoua, Cameroon. *Renewable Energies* 18(1): 105 – 125.

Killick R, Eckley I. A, Jonathan P, Ewans K (2010). "Detection of Changes in the Characteristics of Oceanographic Time-Series using Statistical Change Point Analysis." Ocean Engineering, 37(13), 1120–1126.

Killick, R. and Eckley, I. A. (2014) Changepoint: An R Package for Changepoint Analysis. Journal of Statistical Software 58(3): 1 – 19.

Kim, S.J., Sohn, J.M. and Paik, J. (2017) An Advanced Procedure for the Quantitative Risk Assessment of Offshore Installations in Explosions. The Royal Institution of Naval Architects Trans RINA, Vol 159, Part A2, *Intl J Maritime Eng*, A-123 – A-138.

Kim, S. J., Lee, J., Paik, J. K., Seo, J. K., Shin, W. H. and Park, J. S. (2016) A Study on Fire Design Accidental Loads for Aluminium Safety Helidecks. *International Journal of Naval Architecture and Ocean Engineering* (8): 519 – 529

Kim, S. J., Sohn, J. M. and Paik, J. K. (2017) An Advanced Procedure for the Quantitative Risk Assessment of Offshore Installations in Explosions. Trans RINA, Trans RINA, Vol 159, Part A2, Intl J Maritime Eng: A-123 -A138.

Kim, Y. J and Cribbie, R. A. (2018) ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *The British Journal of Mathematics and Statistical Psychology*, 71(1):1-12.

Kingsley, B. J. and Kaelin, D. E. S. (2012) Is Your Process Safety Documentation Adequate? Chemical Processing News, https://www.chemicalprocessing.com/articles/2012/is-your-process-safety-documentation-adequate/, last access date: 3rd May 2016.

Kitchel, T. and Ball, A. L. (2014) Quantitative Theoretical and Conceptual Framework Use in Agricultural Education Research. *Journal of Agricultural Education*, 55(1): 186-199.

Kletz, T. A. (1994) What Went Wrong? Case Histories of Process Plant Disasters, Fourth Edition, *Gulf Publishing Company*, ISBN *1098765.*

Kletz, T. (2001) Learning from Accidents, 3rd Edition *Gulf Professional Publishing* ISBN 978-0-7506-4883-7.

Kotek, L. and Tabas, M., 2012. HAZOP study with qualitative risk analysis for prioritization of corrective and preventive actions. *Procedia Engineering*, 42: 808-815.

Koteswara, R. G and Kiran, Y. (2016) Analysis of Accidents in Chemical Process Industries in the period 1998-2015. *International Journal of ChemTech Research* 9(4): 177-191.

Kwok, F. T. (2018) An Automated Energy Detection Algorithm Based on Kurtosis-Histogram Excision, *US Army Research Laboratory Publication*, 1-9.

Lekka, C. & Sugden, C. (2011) The Successes and Challenges of Implementing High Reliability Principles: A Case Study of a UK Oil Refinery, *Process Safety and Environmental Protection*, 89: 443-451.

Leveson, N. G. and Stephanopoulos, G. (2014) A Systematic-theoretic, Control-inspired view and Approach to Process Safety. *AiChE Journal* 60 (1): 2-14.

Li, D. (2018) What's the difference between Error, Risk and Loss? Data Science blog https://datascience.stackexchange.com/questions/35928/whats-the-difference-between-error-risk-and-loss Access Date 18th December 2018.

Li, H., Li, Y. and Yu, H. (2019) A Novel Health Indicator Based on Cointegration for Rolling Bearings' Run-To-Failure Process. *Sensors,* (19). 1-24.

Li, S., Meng, Q., Qu, X., 2012. An overview of maritime waterway quantitative risk assessment models. *Risk Anal*. 32, 496–512.

Li, X., Qiu, W., Morrow, J., DeMeo, D. L., Weiss, S. T., Fu, Y. and Wang, X. (2015) A Comparative Study of Tests for Homogeneity of Variances with Application to DNA Methylation Data, *PL0S One*, 10: 1-12.

Liuzzo, G., Bentley, S., Giacometti, F., Bonfante, E. and Serraino, A. (2014) The Term Risk: Etymology, Legal Definition and Various Traits. *Italian Journal of Food Safety* 3(2269): 36 – 39.

Lunt, J. A., Sheffield, D., Bell, N. Bell, Bennett, V. & Morris, L. A. (2011) Review of preventative behavioural interventions for dermal and respiratory hazards, *Occupational Medicine*, 61: 311- 320.

Marsupial, D (2010) What is the difference between data mining, statistics, machine learning and AI?, Stack Exchange 1st December 201, Access date 4th November 2017 https://stats.stackexchange.com/q/5064.

Martínez-Córcoles, M., Gracia, F., Tomas, I. & Peiro J.M. (2011) Leadership and Employees' Perceived Safety Behaviours in a Nuclear Power Plant: A structural Equation Model, *Safety Science,* 49: 1118-1129.

Marwick, B. (2015) Convert PDFs to text files or CSV files (DfR format) with R, *GitHubGist*, https://gist.github.com/benmarwick/11333467 Access date: 17th April 2017.

Maxel, O.J.M., (2013) Managerial Challenge to Cross Cultural Management of Diversity, *European Journal of Business and Management,* Vol. 5 (20): 177-184.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, Access date: 15th November 2018.

McMurdie, P.J. and Holmes, S. (2013) Phyloseq: an R package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PloS one*, 8(4): 1-12.

McNamee R. (2003) Confounding and confounders. Contrasts competing definitions of a confounder, including those based on data and those based on notions of comparability. *Occup Environ Med*; 60:227-34.

Meek, M. E., Boobis, A. R., Crofton, K. M., Heinemeyer, G., Van Raaij, M. and Vickers, C. (2011) Risk assessment of combined exposure to multiple chemicals: A WHO/IPCS framework. *Regulatory Toxicology and Pharmacology,* 60: 1-14

MIB (2008) The Buncefield Incident 11 December 2005: The final report of the *Major Incident Investigation Board* (MIB) Volume 1, ISBN 978 0 7176 6270 8: 1- 106.

Müller, V. C. and Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion', in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Synthese Library; Berlin: Springer), 553-571.

NASA (2011) System Failure Case Studies, *Dust to Dust* 5(2): 1-4.

Natu, M. (2013) Bearing Fault Analysis Using Frequency Analysis and Wavelet Analysis, International *Journal of Innovation, Management and Technology, 4(1):* 90- 92.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N. and Varnier, C. (2012) PRONOSTIA: An Experimental Platform for Bearings Accelerated Life Test. *IEEE International Conference on Prognostics and Health Management*: 1-11.

Nguyen, T.H., Marmier, F. and Gourc, D. (2013) A decision-making tool to maximize chances of meeting project commitments. *International Journal of Production Economics*, 142(2): 214-224.

Nicol, J. (2001). Have Australia's major hazard facilities learnt from the Longford disaster? An evaluation of the impact of the 1998 ESSO Longford explosion on the petrochemical industry in 2001. *The Institution of Engineers*, Australia. ISBN 085825 738 6.

Niewoehner, G. (2008) How to Ensure Safety in Process Plants. MAVERICK Technologies LLC White Paper: 1- 5.

Nishiguchi, J. and Takai, T. (2010) IPL2 and 3 performance improvement method for process safety using event correlation analysis, *Computers and Chemical Engineering*, 34: 2007-2013.

Niskanen, T., Louhelainen, K. & Hirvonen, M.L. (2014) Results of the Finnish national survey investigating safety management, collaboration and work environment in the chemical industry, *Safety Science*, 70: 233-245

NIST e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/apr/section1/apr162.htm, Access Date 14 November 2017.

Nistane, V.M. and Harsha, S.P. (2016) Failure Evaluation of Ball Bearing for Prognostics, *Procedia Technology* 23: 179-186.

Noda, M., Takai, T. and Higuchi, F. (2012) Operation Analysis of Ethylene Plant by Event Correlation Analysis of Operation Log Data. *In Proc. of FOCAPO*, 8-11.

Nordstokke , D. W and Zumbo, B. D. (2007) A Cautionary Tale About Levene's Tests for Equal Variances*, Journal of Educational Research & Policy Studies*, 7(1): 1- 14.

Nordstokke, David W. and Colp, S. M. (2014) Investigating the robustness of the nonparametric Levene test with more than two groups. *Psychological*, 35: 361- 383.

NTC (2001) Criticality analysis for maintenance purposes. *Norwegian Standard for Oil and Gas Industry (NORSOK standard Z-008)*, *Norwegian Technology Centre*, 2: 1- 30.

Okoh, C., Roy, R., Mehnen, J. and Redding, L. (2014) Overview of Remaining Useful Life Prediction Techniques in Through-Life Engineering Services. *Procedia* CIRP 16: 158 – 163.

OSHA (1994) Process Safety Management - Guidelines for Compliance, US Occupational Safety and Health Administration (OSHA) publication OSHA 3133: 1- 39. https://www.osha.gov/Publications/osha3133.pdf last access date: 2nd May 2016.

OSHA (2000) Process Safety Management, US Occupational Safety and Health Administration (OSHA) publication OSHA 3132: 7-26. https://www.osha.gov/Publications/osha3132.pdf Access date: 2nd May 2016

Page (1963) Controlling Stand Deviation by CUSUMs and Warning Lines. *Institute of Statistics Mimeo Series* No. 356.:1 – 18.

Paltrinieri, N., Tugnoli, A., Buston, J., Wardman, M. and Cozzani, V. (2013), Dynamic Procedure for Atypical Scenarios Identification (DyPASI): A new systematic HAZID tool. *Journal of Loss Prevention in the Process Industries*, 26(4), 683-695.

Paltrinieri, N., Khan, F., Amyotte, P. and Cozzani, V. (2014) Dynamic approach to risk management: Application to the Hoeganaes metal dust accidents. *Process Safety and Environmental Protection*, 92(6): 669-679.

Paradis, E. (2010) pegas: an R package for Population Genetics with an Integrated–modular Approach. *Bioinformatics*, 26(3): 419-420.

Parhami, B. (1997) Defect, Fault, Error, …, or Failure. *IEEE Transaction on Reality* 46 (4): 450 – 451.

Pasman, H. & Rogers, W. (2014) How can we use the Information Provided by Process Safety Performance Indicators? Possibilities and Limitations. *Journal of Loss Prevention in the Process Industries*, 30: 197-206.

Pasman, H., Reniers, G., 2014. Past, present and future of Quantitative Risk Assessment (QRA) and the incentive it obtained from Land-Use Planning (LUP). *J. Loss Prev. Process Ind*. 28, 2–9.

Pasman, H.J., Knegtering, B. and Rogers, W.J. (2013) A holistic approach to control process safety risks: Possible ways forward. *Reliability Engineering & System Safety*, 117: 21-29.

Patel, P. and Sohani, N. (2013) Review of Available System Safety Assessment Tools and Techniques-Integrated Approaches for Accident Prevention in Process Industry. *International Journal of Latest Trends in Engineering and Technology (IJLTET),* 2(4): 251-258.

Patel, R. K. and Giri, V. K. (2017) Condition monitoring of induction motor bearing based on bearing damage index. *Archives of Electrical Engineering*, 66 (1): 105 – 119.

Payne, S. C., Bergman, M. E., Rodríguez, J. M., Beus, J. M. and Henning, J. B. (2010) Leading and lagging: Process safety climate–incident relationships at one year. *Journal of Loss Prevention in the Process Industries*, 23(6): 806-812.

Perl, I., Mulyukin, A. and Kossovich, T. (2017) Continuous execution of system dynamics models on input data stream, *Proceeding of the 20th Conference of Fruct Association*: 371 – 376.

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., Cornilly, D., Hung, E., Lestel, M., Balkissoon, K. and Wuertz, D. (2018) Econometric Tools for Performance and Risk Analysis. Cran.r-

project. Source: https://cran.r-project.org/web/packages/PerformanceAnalytics/PerformanceAnalytics.pdf. Access Date 11th June 2018

Pezier, J. (2002). A Constructive Review of Basel's Proposals on Operational Risk. *ISMA Discussion Papers in Finance*. 2002-20:1-28

Pitacco, E. (2014) The Insurer's Perspective: Managing Risks, EAA Series (2014): 1-18. http://www.springer.com/978-3-319-12234-2, Access date: 4th March 2017.

Pour, H. A., Heidari, M. R., Norouzzadeh, R., Rahimi, F., Kazemnejad, A. and Fallahi, F. (2016) Psychometric Evaluation of The Sex After Myocardial Infarction Knowledge Test in Iranian Context. *Perspectives In Psychiatric Care*, 54(2): 1- 8.

Pourhoseingholi, M.A., Baghestani, A.R. and Vahedi, M. (2012) How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from bed to bench*, 5(2): 79-84.

Prem, K. P., Ng, D. and Mannan, M. S. (2010) Harnessing database resources for understanding the profile of chemical process industry incidents, *Journal of Loss Prevention in the Process Industries*, 28: 549-560.

Preventive and Predictive Maintenance, 700ZB00102, https://www.lce.com/pdfs/The-PMPdM-Program-124.pdf. Download date: 13th September 2018.

Qiu, H., Lee, J., Lin, J. and Yu, G. (2006) Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics, *Journal of Sound and Vibration 289*, (4–5): 1066-1090.

Quinn, G. D. and Quinn, J. B. (2010) A practical and systematic review of Weibull statistics for reporting strengths of dental materials. *Dent Mater* 26(2): 135–147.

Rae, A. McDermid, J. and Alexander, R. (2012) The Science and Superstition of Quantitative Risk Assessment. *In Proceedings of PSAM 11 & ESREL*. 3: 2292-2301.

Ragab, A., Ouali, M-S., Yacout, S. and Osman, H. (2014) Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation. *Journal of Intelligent Manufacturing*: 1- 16.

Rajkumar, K., Kundu, K., Aravindan S. and Kulkarni, M.S. (2011) Accelerated Wear Testing for Evaluating the Life Characteristics of Copper–graphite Tribological Composite. *Materials and Design* 32: 3029–3035.

Reddy, P. (2018) The Differences Between Research Methods and Research Methodology. DifferenceBetween.net. May 31, 2018. http://www.differencebetween.net/science/the-differences-between-research-methods-and-research-methodology/

Reeves, J. and Chen, J. (2007) A Review and Comparison of Change-point Detection Techniques for Climate Data, *Journal of Applied Metrology and Climatology* 46: 900 – 912.

Reniers, G.L.L, Cremer, K. & Buytaert (2011) Continuously and simultaneously optimizing an organization's safety and security culture and climate: the Improvement Diamond for Excellence Achievement and Leadership in Safety & Security (IDEAL S&S) model, *Journal of Cleaner Production,* 19: 1239-1249

Rezvanizaniani, S. M., Dempsey, J. and Lee, J. (2014) An Effective Predictive Maintenance Approach based on Historical Maintenance Data using a Probabilistic Risk Assessment: PHM14 Data Challenge. *International Journal of Prognostics and Health Management*, 18: 1 – 13.

Riganti, V., Timidei, A., Girolamo, C. D and Mazzei, N. (2007) Case Studies on the explosion of Organic Powder in the Pharmaceutical Industry and in the Foundry Sand. *Journal of Commodity Science Technology and Quality* 46 (I-IV): 1- 14.

Rippel, O., Snoek, J. and Adams, R. P. (2015) Spectral Representations for Convolutional Neural Networks. *In Advances in Neural Information Processing Systems*: 2449-2457.

Rocco, T. S. and Plankhotnik, M. S. (2009) Literature Reviews, Conceptual Frameworks, and Theoretical Frameworks: Terms, Functions, and Distinctions. *Human Resource Development Review*, 8(1): 121- 130.

Rogers, R. L., Radandt, S., and Schwartzbach, C. (2000) Methodology for the Risk Assessment of Unit Operations and Equipment for Use in Potentially Explosive Atmospheres, *The RASE Project Report*, EU Project No: SMT4-CT97-2169: 5-89.

Salh, S. M. (2014) Using weibull distribution in the forecasting by applying on real data of the number of traffic accidents in sulaimani during the period (2010-2013), *International Journal of Advancements in Research & Technology* 3(10): 58 – 68.

Saunders, M., Lewis, P. and Thornhill, A., (2007) Research methods. *Business Students 4th edition Pearson Education Limited*, England. ISBN: 978-1-292-20878-7. Chapter 4: 128- 170.

Seeber, R. and Ulrici, A. (2016) Analog and digital worlds: Part 1. Signal sampling and Fourier Transform, *ChemText,* 2(18): 1-12.

Shah, D. S. and Patel, V. N. (2014) A Review of Dynamic Modelling and Fault Identifications Methods for Rolling Element Bearing. *Procedia Technology*, 14: 447 – 456.

Sharma, S., Swayne, D. A. and Obimbo, C. (2016) Trend analysis and change point techniques: a survey. Energy, Ecology and Environment 1 (3): 123–130.

Skowron-Grabowska, B. and Sobociński, M. (2018) Behaviour Based Safety (BBS) - Advantages and Criticism. *Production Engineering Archives* 20: 12-15.

Sloan, S. (2007) Risk Management vs. Safety Management: Can't we all just get along? *ASSE Professional Development Conference*, Florida:1 – 7.

Smith, T. A. (2010) Continual Renewal and Improvement and System/Safety Performance, a New Theory for Managing Safety in the 21st Century. *A White Paper by Mocal, Inc*. (248): 1- 14.

Snellenburg, J., Laptenok, S., Seger, R., Mullen, K. and Van Stokkum, I. (2012.) Glotaran: a Java-based Graphical user Interface for the R package TIMP. *Journal of Statistical Software*, 49(3): 1 – 22.

Soeder, D.J., Sharma, S., Pekney, N., Hopkinson, L., Dilmore, R., Kutchko, B., Stewart, B., Carter, K., Hakala, A. and Capo, R., (2014) An approach for assessing engineering risk from shale gas wells in the United States. *International Journal of Coal Geology*, 126: 4-19.

Sohaib, M. Kim, C-H and Kim, J-M (2017) A Hybrid Feature Model and Deep-Learning-Based Bearing Fault Diagnosis. *Sensors, 17(12):* 1-16.

Sprint, G., Cook, D., Weeks, D., Dahmen, J. and La Fleur, A. (2017) Analysing Sensor-Based Time Series Data to Track Changes in Physical Activity during Inpatient Rehabilitation. *Sensors 17(10)*: 1 - 20.

Sullivan, M.J. (2016) Controlling for "confounders" in psychosocial pain research. *Pain*, 157(4): 775-776.

Sulzer-Azaroff, B. and Austin, J. (2000). Does BBS Work? Professional Safety. *American Society of Safety Engineers.*

Swuste, P., van Gulijk, C., Zwaard, W. and Oostendorp, Y. (2014) Occupational safety theories, models and metaphors in the three decades since World War II, in the United States, Britain and the Netherlands: A literature review. *Safety Science* 62: 16 – 27.

Tan, Z., Li, J., Wu, Z., Zheng, J. and He, W. (2011) An evaluation of maintenance strategy using risk-based inspection. *Safety Science*, 49(6): 852-860.

Taroun, A., 2014. Towards a better modelling and assessment of construction risk: insights from a literature review. *Int. J. Project Manage* 32, 101–115.

Tiryakioglu, M. and Campbell, J. (2010) Weibull Analysis of Mechanical Data for Castings: A Guide to the Interpretation of Probability Plots. Metallurgical and Materials Transactions A, 41: 3121-3129.

Tixier, J., Dusserre, G., Salvi, O. and Gaston, D. (2002), Review of 62 risk analysis methodologies of industrial plants. *Journal of Loss Prevention in the Process Industries*, 15: 291–303.

Tobon-Mejia, D.A., Medjaher, K., Zerhouni, N. and Tripot, G. (2012) A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models. *IEEE Transactions on reliability*, 61(2): 491-503.

Truong, C., Oudre, L. and Vayatis, N. (2018) A review of change point detection methods, *A ResearchGate Publication*: 1-31.

Upadhyay, R.K., Kumaraswamidhas, L.A. Azam, Md-S (2013) Rolling element bearing failure analysis: A case study, Case Studies in Engineering. *Failure Analysis* 1(1): 15-17

US CSB (2011) Hoeganaes Corporation: Gallatin, TN Metal Dust Flash Fires and Hydrogen Explosion, *CSB Hoeganaes Corporation Case Study Report*, http://www.csb.gov/assets/1/19/CSB_Case_Study_Hoeganaes_Feb3_300-1.pdf (Accessed date: 1st February 2017).

Veritas, D. N. (2001) Marine Risk Assessment, HSE Offshore Technology Report 063, *ISBN 0 7176*, 2231 2: 1- 72.

Villa, V., Paltrinieri N., Cozzani, V. (2015) Overview on Dynamic Approaches to Risk Management in Process Facilities. *Chemical Engineering Transactions*, 43: 1-6.

Villa, V., Paltrinieri, N., Khan, F. and Cozzani, V. (2016) Towards dynamic risk analysis: a review of the risk assessment approach and its limitations in the chemical process industry. *Safety Science*, 89: 77-93.

Vinnem, J. E. (2010) Risk indicators for major hazards on offshore installations, *Safety Science*, 48: 770-787.

Vorderbrueggen, J. (2010) Imperial Sugar Refinery Combustible Dust Explosion Investigation. Conference Presentation AIChE Spring Meeting and Global Congress on Process Safety.

https://www.aiche.org/academy/videos/conference-presentations/imperial-sugar-refinery-combustible-dust-explosion-investigation, Access date: 3rd December 2016.

Wang, Y.F., Li, Y.L., Zhang, B., Yan, P.N. and Zhang, L. (2015) Quantitative Risk Analysis of Offshore Fire and Explosion Based on the Analysis of Human and Organizational Factors. *Mathematical Problems in Engineering*, 2015: 1-10.

Wang, T. (2012) Bearing life prediction based on vibration signals: A case study and lessons learned, IEEE Conference on Prognostics and Health Management (PHM), *ResearchGate Publication*: 1-8.

Wang, Y., Ma, G., Ding, S.X. and Li, C. (2011) Subspace aided data-driven design of robust fault detection and isolation systems. *Automatica*, 47(11):2474-2480.

Wang, Y., Ziedins, I., Holmes, M. and Challands, N. (2012) Tree Models for Difference and Change Detection in a Complex Environment. *The Annals of Applied Statistics* 6(3):1162–1184.

Wasserman, L (2014). "Rise of the Machines." *Past, Present and Future of Statistical Science*. 1 – 12.

Weber, P., Medina-Oliva, G., Simon, C. and Iung, B., 2012. Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, 25(4): 671-682.

Weinberg, C.R. (1993) Toward a clearer definition of confounding. *American Journal of Epidemiology*, 137(1): 1-8.

Whewell, I. (2012) Performance Indicators in major hazard industries– An Offshore Regulator's perspective. *CSB public hearing on process safety performance indicators* Houston Texas 23rd -24th July 2012: 1 -9.

White, G. H. (2008) Basics of Estimating Measurement Uncertainty. *The Clinical Biochemist Reviews*, 29(1): 53– 60.

Wilde, G. J. S. (1982) The Theory of Risk Homeostasis: Implications for Safety and Health. *Risk Analysis*, 2(4): 209-225.

Wilkins, D. J. (2002) The Bathtub Curve and Product Failure Behavior Part One - The Bathtub Curve, Infant Mortality and Burn-in. Reliability HotWire: *eMagazine for the Reliability Professional* (21). Source: https://www.weibull.com/hotwire/issue21/hottopics21.htm, Access date: 7th September 2017.

Witten, I. H., Frank, E., and Hall, M. A. (2011) Data mining: Practical machine learning tools and techniques. Second Edition, Los Altos, CA: Morgan Kaufmann; ISBN: 0-12-088407-0:1-524.

Wu, S. H., Chi, J. H., Huang, C. C., Lin, N. K., Peng, J. J. and Shu, C. M. (2010) Thermal hazard analyses and incompatible reaction evaluation of hydrogen peroxide by DSC. *Journal of thermal analysis and calorimetry*, 102(2): 563-568.

Wu, S., Zhang, L., Zheng, W., Liu, Y. and Lundteigen, M. A. (2016). A DBN-based risk assessment model for prediction and diagnosis of offshore drilling incidents. *Journal of Natural Gas Science and Engineering*, 34: 139-158.

Xu, T, Tang, T., Wang, H. and Yuan, T. (2013) Risk-Based Predictive Maintenance for Safety-Critical Systems by Using Probabilistic Inference. *Mathematical Problems in Engineering* 2013: 1 – 9.

Yildiza, A. E., Dikmen, I. and Birgonul, M. T. (2014) Using Expert Opinion for Risk Assessment: a Case Study of a Construction Project Utilizing a Risk Mapping Tool. 27th IPMA World Congress, *Procedia - Social and Behavioural Sciences* 119: 519 – 528.

Yin, S., Ding, S.X., Abandan Sari, A.H. and Hao, H. (2013) Data-driven monitoring for stochastic systems and its application on batch process. *International Journal of Systems Science*, 44(7): 1366-1376.

Yin, S., Ding, S.X., Xie, X. and Luo, H. (2014) A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, 61(11): 6418-6428.

Yin, S., Wang, G. and Karimi, H.R. (2014) Data-driven design of robust fault detection system for wind turbines. *Mechatronics*, 24(4): 298-306.

Yin, S., Yang, X. and Karimi, H. R. (2012) Data-Driven Adaptive Observer for Fault Diagnosis, *Mathematical Problems in Engineering*: 1- 21.

Zarenistanak, M., Dhorde, A. and Kripalani, R. H. (2014) Trend analysis and change point detection of annual and seasonal precipitation and temperature series over southwest Iran. *J. Earth Syst. Sci.* 123(2): 281–295.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software* 7 (2): 1–38.

Zhai, L. Y., Lu, W.F., Liu, Y., Li, X. and Vachtsevanos, G. (2013) Analysis of Time-to-Failure Data with Weibull Model in Product Life Cycle Management. In: *Re-engineering Manufacturing for Sustainability*: 699-703.

Zhang, B., Zhu, G., Lv, B. and Yan, G. (2018) A Novel and Effective Method for Coal Slime Reduction of Thermal Coal Processing. *Journal of Cleaner Production* 198: 19 – 23.

Zheng, Y. Chen, Y., Xie, X. and Ma, W-Y. (2010) Understanding transportation modes based on GPS data for Web applications - Microsoft Research. *ACM Trans Web* 4(1): 1-36.

Appendix - Tables

Table 3.2d: Behavioural program elements

Behavioural program elements	Abbreviations	Point
Technical aspects like:		
engineering design,	ED	1
equipment's for communication during operation	ECDO	1
maintenance	M	1
Human aspects such as:		
Identification of at-risk /unsafe behaviour	IU	1
observation of behaviours	OB	1
feedback/ nature of reinforcement of behaviours	FNRB	1
encouraging safe and removing unsafe behaviours	ESB/RUB	1
knowledge to respond appropriately to developing incidents	KRADI	1
relation and networks	RN	1
values, attitudes and competence	VAC	1
work practice	WP	1
Organisation aspect like:		
influence of management	MI	1
quality of products the products of systems e.g. quality of safety audit system	QP	1
availability of operator, sufficient/insufficient	AOP	1
number of solutions being enforced	NSBE	1

Table 3.2e: Scoring criteria

Effectiveness of program	Point criteria
Reduction in accidents	+1
No Change	0
Worse situation than before	-1
Not specified (because element article)	-

Table 3.2f: Additional criteria

Criteria	Qualifier	Point criteria
Study duration	Per year of duration	1
Sample size	Bigger sample size affects reliability of study (point awarded per 100 samples)	1
Sampling methodology	Well defined data collection/study design	1

Table 3.2g: Effectiveness ranking

Author	Technical aspect			Human aspect								Organisation aspect				Points
	ED	ECDO	M	IU	OB	FNRB	ESB /RUB	KRAD I	RN	VAC	WP	MI	QP	ASIO	NSBE	
Coze (2010)	-1	-	-1	0	-1	-1	-1	-	-	-1	-1	-1	-1	0	-1	-10
Lekka & Sugden (2011)	-	0	-	0	-	1	-	0	1	-	-	1	1	-	0	4
Martinez-Corcoles et al (2011)	-	1	-	1	1	1	1	-	-	-	-	1	-	-	1	7
Niskanen et al (2014)	-	-	-	0	0	-	0	-	0	-	-	0	-	-	-	0
Reniers, et al. (2011)	0	-	0	-	-	0	0	0	-	0	-	0	0	-	-	0
Vinnem (2010)	0	-	0	-	-	-	-	-	-	-	-	-	-	-	-	0

Point: <0 = worse situation than before; 0 = No change; >0 = Reduction in accident/incidents

Table 3.2h: Process scoring

Author	Technical aspect			Human aspect								Organisation aspect				Points
	ED	ECDO	M	IU	OB	FNRB	ESB /RUB	KRAD I	RN	VAC	WP	MI	QP	ASIO	NSB E	
Coze (2010)	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	12
Lekka & Sugden (2011)	0	1	0	1	0	1	0	1	1	0	0	1	1	0	1	8
Martinez-Corcoles et al (2011)	0	1	0	1	1	1	1	0	0	0	0	1	0	0	1	7
Niskanen et al (2014)	0	0	0	1	1	0	1	0	1	0	0	1	0	0	0	5
Reniers, et al (2011)	1	0	1	0	0	1	1	1	0	1		1	1	0	0	8
Vinnem (2010)	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	3

Point: 0-5 = low; 6-10= medium; >10= High

Table 3.2i: Applying additional criteria

Article	Study Duration	Sample Size	Sampling Methodology	Safety Issues Address	Method Used to Address Safety	Reported Incident	Research Limitations
Coze (2010)	2-3 months	30 people from one company	Qualitative (direct interview of different hierarchy levels)	OHS & Process safety	Occupational health and safety (OHS) & Process safety	Explosion in manufacturing process plant	Method based on probable scenarios other reason/parameters may be missed. Research therefore prone to bias.
Lekka & Sugden (2011)	1 month	21 people from one company	Qualitative case study approach (direct interview of different hierarchy levels)	OHS & Process safety	OHS; Engineering design; Process safety	No reported accident or incident (NRAI)	Impact of reliability-enhanced method not validated; Extent to which data collection methods affect the method was not examined; Emerging issues not scrutinize; study relied on volunteers and focus groups hence risk of bias in the research.
Martinez-Corcoles et al (2011)	1 year	566 from one plant	Qualitative-questionnaire	OHS; Process safety & Behavioural safety	Behavioural safety programs & General health and safety	NRAI	Determination of appropriate behaviour at all hierarchy levels unknown, self-reported instrument data used therefore result may have been exaggerated; safe behaviours not tested hence risk of bias.
Niskanen et al (2014)	1 month	258 form 350 companies	Qualitative -online questionnaire	Process safety	Training and operating procedures; General process safety techniques	NRAI	Method didn't allow conclusion to be drawn with absolute confidence; not possible to validate response; self-report measures hence prone to bias; casual relationships not determined; lark of measurement calibrations.
Reniers, et al (2011)	Not specified	7 companies	Qualitative interview	Operational safety and security	Security & Process safety	NRAI	Optimum safety and security parameters for operation and management not well defined.
Vinnem (2010)	5 years	186 installations	Variable interviews	Process safety & OHS	Process safety	NRAI	High awareness of risk maintained. Data collection scheme is limited and insufficient

Points: <0= Low; 0-5= Medium; >5= High

Table 3.3a: Databases searched and reason

Databases	Reasons
Chemical incidents databases Searched	United States Chemical Safety Board (CSB) Investigations; Occupational Safety and Health Administration (OSHA); National Response Center (NRC's) Database of Spills and Accidents; National Transportation Safety Board (NTSB) Accident Reports; The Right-To-Know Network; Toxics Release Inventory; Major Accident Reporting System (eMars); Failure and Accidents Technical information System (FACTS); Process Safety Incident Database (PSID); International Powered Access Federation (IPAF); United States National Response Centre (NRC); US Environmental Protection (EPA); Agency for Toxic Substances and Disease Registry (ATSDR); ProcessNet; Safety to Safety (S2S); Fire and Blast Information Group (FABIG), were access for any references captured in reports relating to process incidents.
Chemical process safety databases	Because incidents within process industries can also affect the public, Chemical process safety databases including Compendex; ProQuest; SciFinder; Knovel; TOXNET; Canadian Centre for Occupational Health and Safety (CCOHS), were searched for records on toxicology, hazardous chemicals, environmental health, and toxic releases, public health, safety, and industrial hygiene and records covering all areas of engineering, information on substances, reactions, and provides structure and substructure as well as engineering references were searched.
Conference papers Database	Since most development or research information within the areas of process safety are shared via conferences and seminars, databases which provide conference papers relating process safety conference literature such as Mary K. O'Connor Process Safety Centre Conference Proceedings; HAZARDS Proceedings; Proceedings of Loss Prevention Symposia & CCPS International Conferences; Compendex; PapersFirst were searched.
Process safety journals	Other information's are sourced from process safety journals such as Fire Safety Journal; Journal of Hazardous Materials; Journal of Loss Prevention in the Process Industries; Loss Prevention Bulletin; Process Safety and Environmental Protection; Process Safety Progress; Safety Science Chemical Hazards in Industry; Risk Analysis were searched.
Domain Related Databases	Databases of other domain related organisations including UK Health and Safety (UK HSE); British standards online (BSOL); European Process Safety Centre (EPSC); European Federation of Chemical Engineering (EFCE); European Union Labour Force Survey (EU-LFS); Chemical Process Safety and American Institute of Chemical Engineers (CCPS & AIChE); SAI Global Standards; American Petroleum Institute (API); Occupational Safety and Health Administration (OSHA) Regulations/Standards, were all searched for related literature.
EBSCO Host	Of the databases under, the search for published articles was performed from Academic Search Complete because it features multi-disciplinary full-text and comprehensive scholarly peer-reviewed journals. Since IEEE Xplore digital library database has networks of books, journals, conference proceedings, and standards only selected publication titles were searched. These publications together with the reason for their inclusion are detailed in Table 11.

Table 3.3b: Databases Search and number of review articles

Database	Reviewed Articles
Wiley Online library	2
IEEE explore	8993
EBSCO host	7679
Science Direct	211
Chemical Incidents Databases	129
Chemical process safety	3145
Process Safety Conference Literature	78
Engineering standards	0
Organisation	813

Table 3.3c: Class of QRA methodologies adopted from (Tixier et al. 2002; Patel & Sohani, 2013).

Deterministic	Probabilistic	Deterministic & Probabilistic
Accident Hazard Analysis [AHI]	DEFI method	AVRIM2
Annex 6 of SEVESO II Directive	Event Tree Analysis (ETA)	Facility Risk Review
Chemical Runaway Reaction Hazard Index (RRHI)	Fault Tree Analysis (FTA)	Failure Mode Effect Criticality Analysis [FMECA]
Dow's Chemical Exposure Index (CEI)	Maintenance Analysis	IDEF3
Dow's Fire and Explosion Index (FEI)	Short Cut Risk Assessment	International Study Group on Risk Analysis [ISGRA]
Fire and Explosion Damage Index (FEDI)	Work Process Analysis Model WPAM	IPO Risico Berekening Methodiek (IPORBM)
Hazard Identification and Ranking (HIRA)	-	Method Organised Systematic Analysis of Risk (MOSAR)
Instantaneous fractional loss index (IFAL)	-	Optimal Risk Assessment (ORA)
Methodology of domino effects analysis	-	Probabilistic Safety Analysis PSA
Methods of potential risk determination and evaluation	-	Quantitative Risk Assessment (QRA)
Mond Fire Explosion and Toxicity Index (FETI)	-	Rapid Ranking (RR)
SAATY methodology	-	Rapid Risk Analysis Based Design (RRABD)
Toxic Damage Index (TDI)	-	Risk Level Indicators (RLI)

Table 3.3d: Criteria for inclusion and exclusion

Inclusion criteria	Exclusion criteria
Evaluation of risk assessment methodologies around process industries.	Risk analysis methodologies applied to construction industries (e.g. residential, road, offices); health (e.g. hospitals), nuclear/radioactive industry, Waste and recycle industries dealing with biological waste.
Risk assessment methodology must predict future hazards	Risk assessments performed as one-time process without any forecasting element.
Incidents and/or incident investigations with successful outcome.	Incident investigations that did not include the outcome.
Title must meet potential research aim.	Titles that do not meet aim of proposed research.
Risk analysis that involves use of sufficient use of data?	
Risk analysis methodology that does not involve or predict external risk/ hazards (e.g. flood).	Risk analysis or prediction methodology that investigates or predict external influence.
Risk assessment methodologies that evaluate conditions before and after predictions were made.	Risk assessment methods which are not designed towards process safety management.
Must fulfil full risk assessment citation research.	Case study research and literature reviews citations are excluded.
Risk assessments that focus on process safety management.	Risks assessment methodology that focuses on the wider occupational health and safety.
Publications from 2007	Publications before 2007 are excluded

Table 3.3e: Search strings and their corresponding number of citations

Search	Number of Citations
Data Informed Risk Prediction in highly hazardous industrial processes	1765
(Data OR Inform* OR Idea) AND (Danger* OR Hazard* OR Accident* OR Incident* OR Safe*) AND (Predict* OR Forecast* OR Estimat* OR Assess* OR Project* Or Model*) AND (Process* OR Manufactur* OR Product*) AND (Industry* OR Company* OR Reaction)	9854
(Data) AND (Inform*) AND (Risk* OR Safe* OR Hazard* OR Danger* OR impact) AND (Predict* OR identify OR investigate*OR analysis OR assess OR assessment* OR model*) AND (protect OR protection OR prevent) AND (hazardous Industry* OR Manufacturing Process*)	6758
(In-process data) AND (Identify OR Investigate*OR Predict* OR Analysis OR Assess* OR Model*) AND (Safe* OR Hazard* OR Danger OR Impact) AND	4478

Search	Number of Citations
(Protect* OR Prevent*) AND (Chemical* OR Product* OR Manufactur*) AND (Environment* OR Pollut* OR Community*)	
(Process-specific OR In-process OR Batch process* OR Chemical* Process OR Chemical* reaction OR Product*) AND (Data OR Inform*) AND (Identify* OR Investigat* OR Predict* OR Analys* OR Assess* OR Model*) AND (Risk* OR Safe* OR Hazard* OR Danger* OR Impact* OR Protect* OR Prevent*) NOT (Environment* OR Pollut* OR Communit*) NOT (Construction OR General health and safety OR Medical OR Biological OR Natural disaster) NOT (Disease OR Illness OR Sickness) NOT (Transport* OR Supply OR Consumer* chain OR Behaviour* OR Manager*) NOT (Climat* change OR Global warming OR Greenhouse OR Atmospheric carbon dioxide) NOT (Cyber OR Internet OR Terror attack* OR Natural disaster)	5124

Table 3.3f: Journals and number of Publication

Journals/Proceedings	Number of Articles
Mathematical Problems in Engineering	3
Process Safety and Environmental Protection	2
Reliability Engineering and System Safety	1
Computers and Chemical Engineering	1
In Proc. of FOCAPO	1
Loss Prevention in the Process Industries	1
Journal of Natural Gas Science and Engineering	1
IEEE Transactions on reliability	1

Table 3.3h: Article and Applied QRA Methods

Article	Model	Methods
Hou et al. (2016)	Bayesian network (BN) and Analytical Network Process (ANP)	CDP (FTA)
Kalantarnia et al. (2010)	Dynamic Risk assessment (DyRA)	CDP (FTA & ETA)
Khakzad et al. (2012)	Bow-tie (BT)	CDP (ETA & FTA)
Nishiguchi & Takai (2010)	Event coloration analysis (ECA)	CDP
Noda, Takai & Higuchi (2012)	ECA	CDP
Paltrinieri, et al. (2013)	Dynamic Procedure for Atypical Scenarios Identification (DyPASI) & Dynamic Risk Assessment (DyRA)	CDP (ETA & FTA)
Shahriar et al. (2012)	Bow-tie analysis	CDP (FTA & ETA)
Tobon-Mejia et al. (2012)	Failure Prognostics Method	CDP (FMECA)
Wang et al. (2015)	Bayesine network	CDP (FTA & ETA)
Wu et. al. (2016)	Bayesian network	CDP (ETA & FTA)
Yin et al. (2012)	FDI	CDP (FMECA)

Table 3.3i: Publication and Corresponding Study Objectives

Article	Objective of the Studies
Hou & Zhao, (2016)	Identify security weaknesses in oil wharf handling process
Kalantarnia et al. (2010)	Evaluate whether the BP accident could have been predicted early by learning from process history
Khakzad et al. (2012)	Demonstrate importance of dynamic BTs in real-time risk analysis.
Nishiguchi & Takai (2010)	Use data-based evaluation of alarm and operation log data to prevent accident within a process plant.
Noda et al. (2012)	Evaluate robustness of improved method of EC in identifying similarities between two physically related events.
Paltrinieri, et al. (2013)	Produce complete hazard identification and calculate the overall risk
Shahriar et al. (2012)	Characterize severity of risk by introducing utility value for consequences to improve ET and FT risk analysis methodologies.
Tobon-Mejia et al. (2012)	Estimate time to failure
Wang et al. (2015)	Investigate dynamic effect of human operation error (HOE) on fire/explosion risk.
Wu et. al. (2016)	Perform predictive, diagnostic and sensitivity analysis for risk assessment.
Yin et al. (2012)	Apply uncertain or normal variation parameters existence within systems with for fault detection and diagnosis.

Table 3.3j: Publication and Research Method

Article	Research method	Points
Hou et al. (2016)	Case study	-1
Kalantarnia et al. (2010)	Case study	-1
Khakzad et al. (2012)	Case study	-1
Nishiguchi et al. (2010)	Case study, Experimental & Surveys	1
Noda et al. (2012)	Case study & Experimental	0
Paltrinieri, et al. (2013)	Case study	-1
Shahriar et al. (2012)	Surveys	-1
Tobon-Mejia et al. (2012)	Experimental	-1
Wang et al. (2015)	Case study	-1
Wu et. al. (2016)	Case study	-1
Yin et al. (2012)	Case study & Experimental	0

Table 3.3k: Publication and Risk Detection

Article	Risk Detection	Points
Hou & Zhao, (2016)	No	0
Kalantarnia et al. (2010)	Yes	1
Khakzad et al. (2012)	No	0
Nishiguchi & Takai (2010)	No	0
Noda et al. (2012)	No	0
Paltrinieri, et al. (2013)	Yes	1
Shahriar et al. (2012)	Yes	1
Tobon-Mejia et al. (2012)	Yes	1
Wang et al. (2015)	No	1
Wu et. al. (2016)	Yes	1
Yin et al. (2012)	No	0

Table 3.3l: Type of data and their components

Data Type	Components
Operations description data	Failure data; Failure type; Failure frequency; Failure probability & Failure rate
Production/reaction data	Reaction parameters; Reaction kinetics; reactor operation conditions; reaction conditions & type and physicochemical properties of reactants and materials
Toxicological data	Toxicity of reactants; Toxicity data on reaction intermediates; Toxicological data of final products & Exposure to toxicological materials during production
Piping and Instrumentation Identification (P&ID) data	Data on condition of piping and instrumentation during operation.

Table 3.3m: Publications with Type and Amount of Data used

Article	Process/other data	Amount of data	Data type	Points
Hou & Zhao, (2016)	Other	47	Operation, production and P&ID	0
Kalantarnia et al. (2010)	Other	12	Operation, production and P&ID	0
Khakzad et al. (2012)	Process	18	Operation & production	1
Nishiguchi & Takai (2010)	Process	1267	Operation	2
Noda et al. (2012)	Process	4011	Operation	2
Paltrinieri, et al. (2013)	Process	12	Operation	1
Shahriar et al. (2012)	Other	66	Operation and P&ID	0
Tobon-Mejia et al. (2012)	Process	20480	Operation	2
Wang et al. (2015)	Other	70	Operation and P&ID	0
Wu et. al. (2016)	Other	21	Operation and P&ID	0
Yin et al. (2012)	Other	2500	Operation and P&ID	0

Table 3.3n: Article and Statistical Analysis Method

Article	Statistical method	Points
Hou & Zhao, (2016)	MCDA	1
Kalantarnia et al. (2010)	PPP	1
Khakzad et al. (2012)	LR, MLE, & BT	3
Nishiguchi & Takai (2010)	ECA	1
Noda et al. (2012)	CA	1
Paltrinieri, et al. (2013)	LR, BS, PD, PP & BS	5
Shahriar et al. (2012)	MP, PC & SRC	3
Tobon-Mejia et al. (2012)	GP & HMM	2
Wang et al. (2015)	BN & PM	2
Wu et. al. (2016)	BN & PM	2
Yin et al. (2012)	KDE	1

Table 3.3o: Publication and Method validation

Article	Method Validation	Points
Hou & Zhao, (2016)	Case study: Oil wharf handling	1
Kalantarnia et al. (2010)	Case study: Oil refinery	1
Khakzad et al. (2012)	Case study: Offshore reliability	1
Nishiguchi & Takai (2010)	Case study: Chemical plant	1
Noda et al. (2012)	Case Study: Ethylene Plant	1
Paltrinieri, et al. (2013)	Case Study: Dust explosion	1
Shahriar et al. (2012)	Relatively not clear	0
Tobon-Mejia et al. (2012)	Experimental: Bearing data set	1
Wang et al. (2015)	Case Study: Offshore Platform	1
Wu et. al. (2016)	Case Study: Offshore drilling well	1
Yin et al. (2012)	Case Study: Three-Tank System	1

Table 3.3p: Publications and Uncertainty

Article	Handling Uncertainty	Points
Hou & Zhao, (2016)	Ineffective	0
Kalantarnia et al. (2010)	Ineffective	0
Khakzad et al. (2012)	Ineffective	0
Nishiguchi & Takai (2010)	Ineffective	0
Noda et al. (2012)	Ineffective	0
Paltrinieri, et al. (2013)	Ineffective	0
Shahriar et al. (2012)	Effective	1
Tobon-Mejia et al. (2012)	Ineffective	0
Wang et al. (2015)	Ineffective	0
Wu et. al. (2016)	Effective	1
Yin et al. (2012)	Ineffective	0

Table 3.3q: Articles and Nature of Events

Article	Nature of perceive event
Hou et al. (2016)	Gasoline fire accident and explosion
Kalantarnia et al. (2010)	Fluid release
Khakzad et al. (2012)	Faults in design, maintenance of equipment, inadequate housekeeping, deficiency in safety measures
Nishiguchi et al. (2010)	Alarm event within operation system
Noda et al. (2012)	Alarm event within operation system
Paltrinieri, et al. (2013)	Maintenance of equipment, inadequate housekeeping, deficiency in safety measures
Shahriar et al. (2012)	Puncture, fracture and natural gas release
Tobon-Mejia et al. (2012)	Degradation in the form of hidden health conditions
Wang et al. (2015)	Release of gas, failure of equipment, ignition
Wu et. al. (2016)	Tripping activities and high pump pressure
Yin et al. (2012)	Faults in the process

Table 3.3r: Publication and Research Limitations

Article	Research Limitations	Points
Hou et al. (2016)	Method limitation	1
Kalantarnia et al. (2010)	Data limitation	1
Khakzad et al. (2012)	Method limitation	1
Nishiguchi et al. (2010)	Data and method limitations	1
Noda et al. (2012)	Failed to highlight any limitations	0
Paltrinieri, et al. (2013)	Method limitation	1
Shahriar et al. (2012)	Method limitation	1
Tobon-Mejia et al. (2012)	Method limitation	1
Wang et al. (2015)	Limitations not acknowledged.	0
Wu et. al. (2016)	Method limitation	1
Yin et al. (2012)	Limitations not acknowledged	0

Table 3.3s: Citations and their ranking

Article	Data	Research Method	Risk Detection	Statistical Method	Validation	Uncertainty	No. of Events	Limitations	Points (+ve)	Points (-ve)
Hou et al. (2016)	0	-1	0	1	1	0	-1	1	3	-3
Kalantarnia et al. (2010)	0	-1	1	1	1	0	-1	1	4	-3
Khakzad et al. (2012)	1	-1	0	3	1	0	-1	1	6	-3
Nishiguchi et al. (2010)	2	1	0	1	1	0	-1	1	5	-1
Noda et al. (2012)	2	0	0	1	1	0	1	0	4	0
Paltrinieri, et al. (2013)	1	-1	1	5	1	0	-1	1	8	-3
Shahriar et al. (2012)	2	-1	1	3	0	1	-1	1	6	-3
Tobon-Mejia et al. (2012)	0	-1	1	2	1	0	-1	1	7	-2
Wang et al. (2015)	0	-1	1	2	1	0	-1	0	4	-3
Wu et. al. (2016)	0	-1	1	2	1	1	-1	1	7	-3
Yin et al. (2012)	0	0	0	1	1	0	-1	0	2	-1

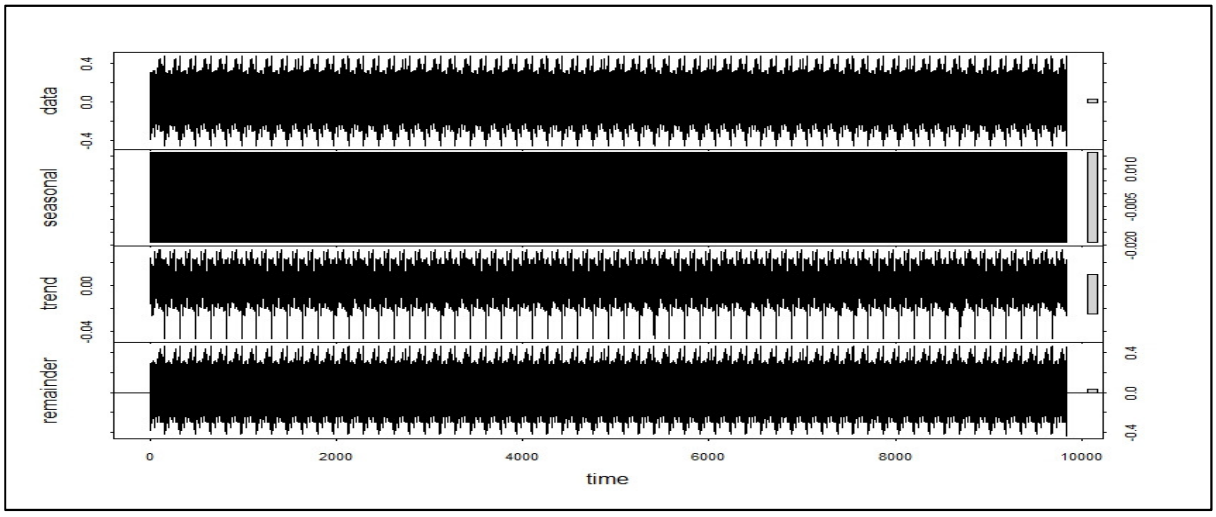


Figure 7.3d: Time series decomposition in vibration of Bearing 2 in Training Dataset

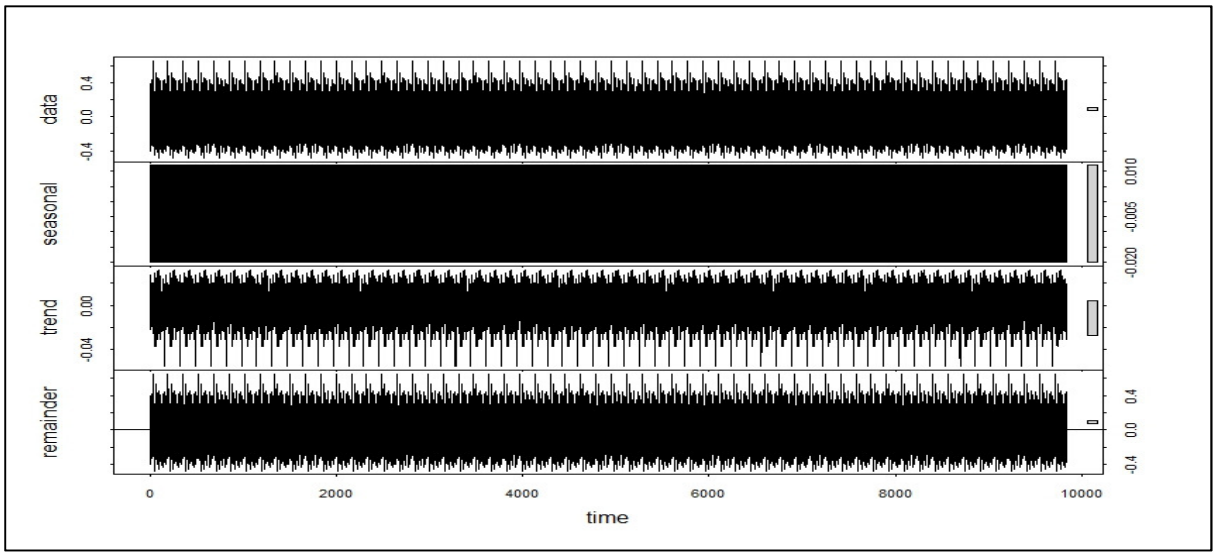


Figure 7.3e: Time series decomposition in vibration of Bearing 3 for Training Dataset

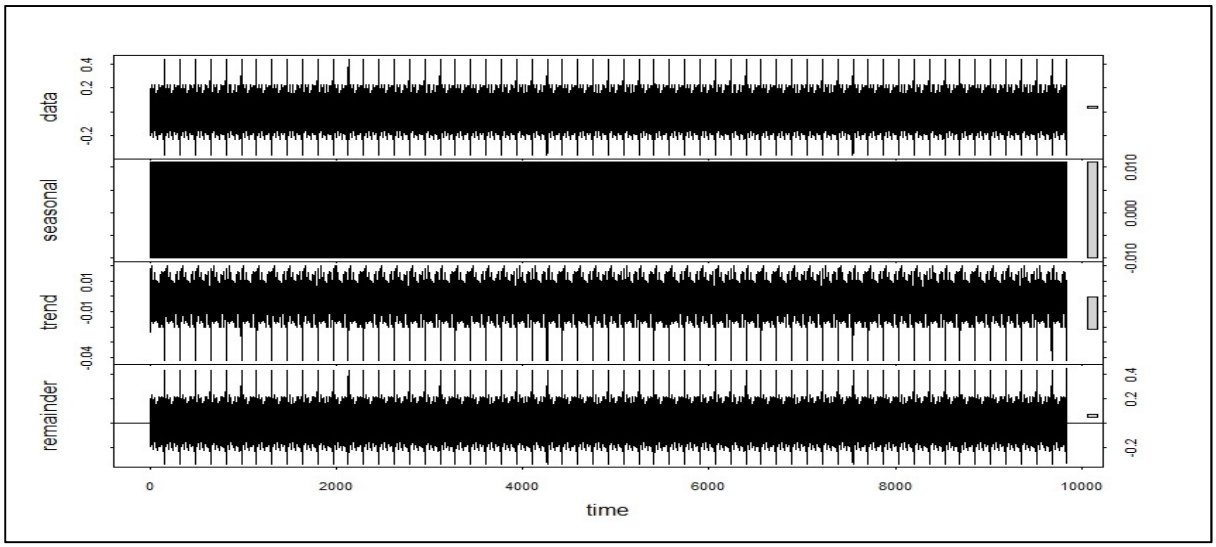


Figure 7.3f: Time series decomposition in vibration of Bering 4 for Training Dataset

```
Top 4 "power" frequencies
      freq      spec
475  0.0475 0.3700145
1383 0.1383 0.3276637
479  0.0479 0.2947734
1001 0.1001 0.2692245
Converting frequency to time periods
 [1] 21.052632  7.230658 20.876827  9.990010
```

Figure 7.3i: Highest frequencies and times in Bearing 1 of Training Dataset

```
Top 4 "power" frequencies
      freq      spec
814  0.0814 0.1955174
3483 0.3483 0.1639471
851  0.0851 0.1557373
502  0.0502 0.1504070

Converting frequency to time periods
 [1] 12.285012  2.871088 11.750881 19.920319
```

Figure 7.3j: Highest frequencies and times in Bearing 2 of Training Dataset

```
Top 4 "power" frequencies
      freq      spec
878  0.0878 0.1852174
2442 0.2442 0.1816113
3632 0.3632 0.1770600
4512 0.4512 0.1594953
Converting frequency to time periods
 [1] 11.389522  4.095004  2.753304  2.216312
```

Figure 7.3k: Highest frequencies and times in Bearing 3 of Training Dataset

```
Top 4 "power" frequencies
      freq       spec
1280 0.1280 0.07178228
3020 0.3020 0.05738450
4336 0.4336 0.05530395
2599 0.2599 0.05276151
Converting frequency to time periods
 [1] 7.812500 3.311258 2.306273 3.847634
```

Figure 7.3l: Highest frequencies and times in Bearing 4 of Training Dataset

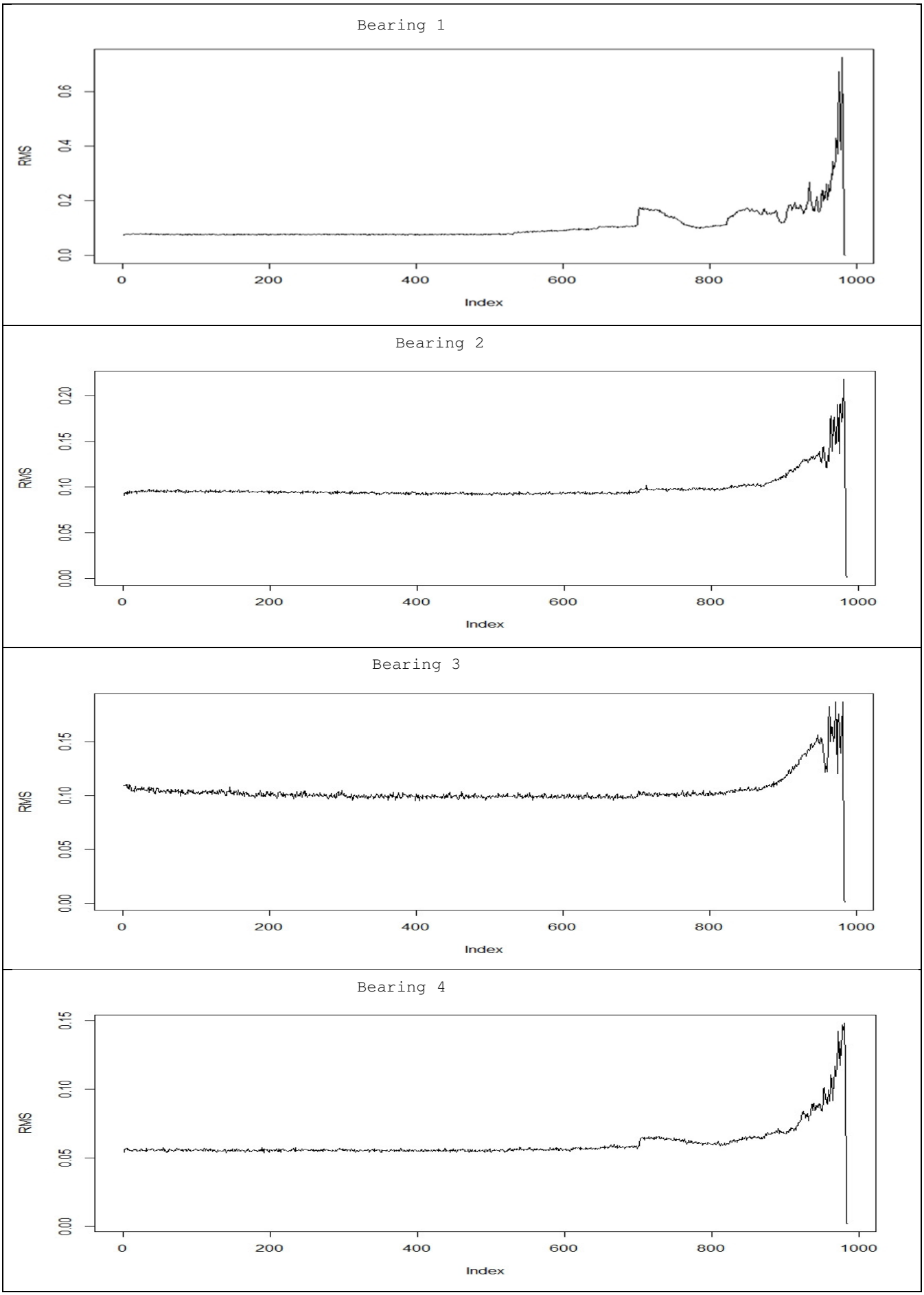
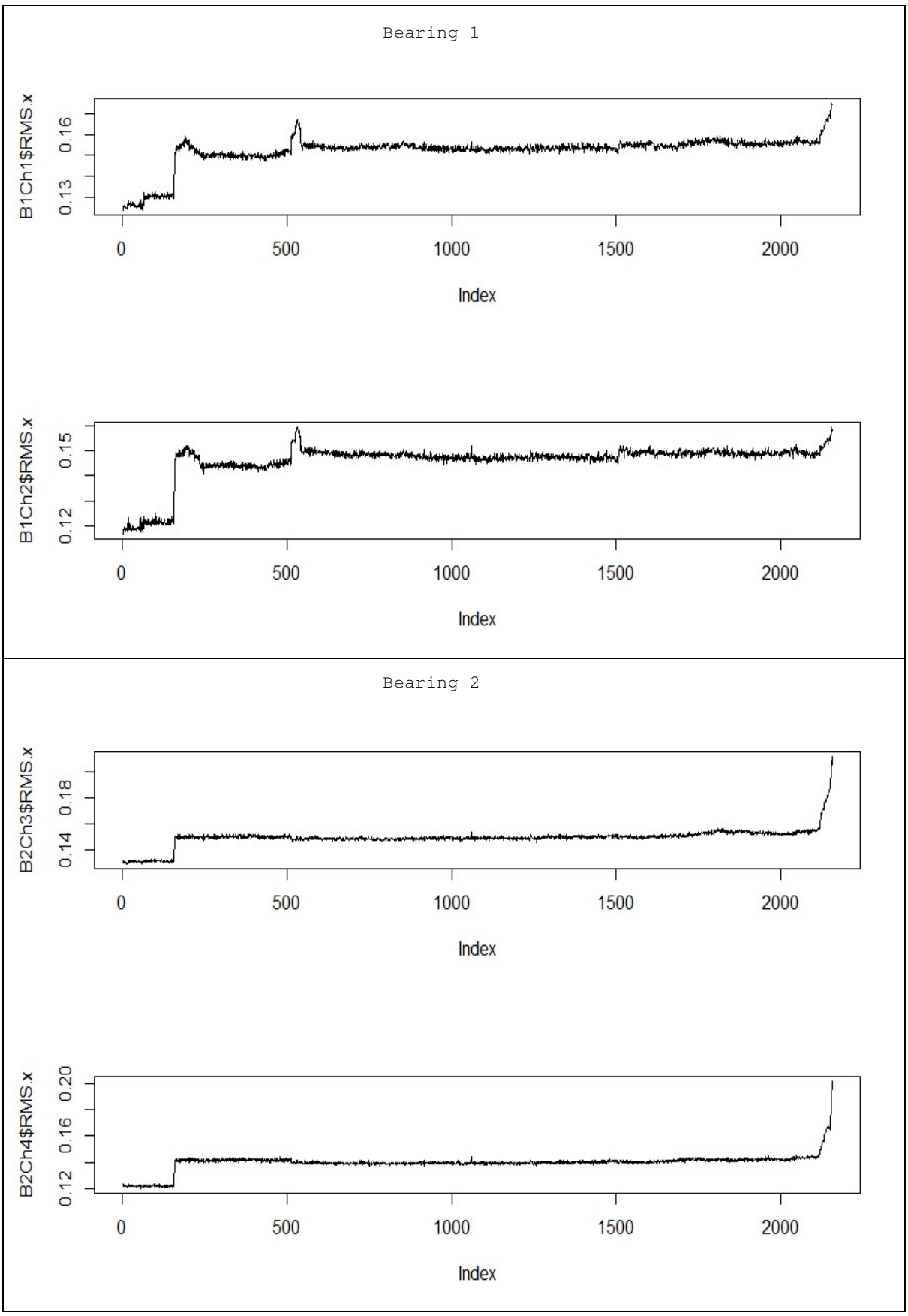
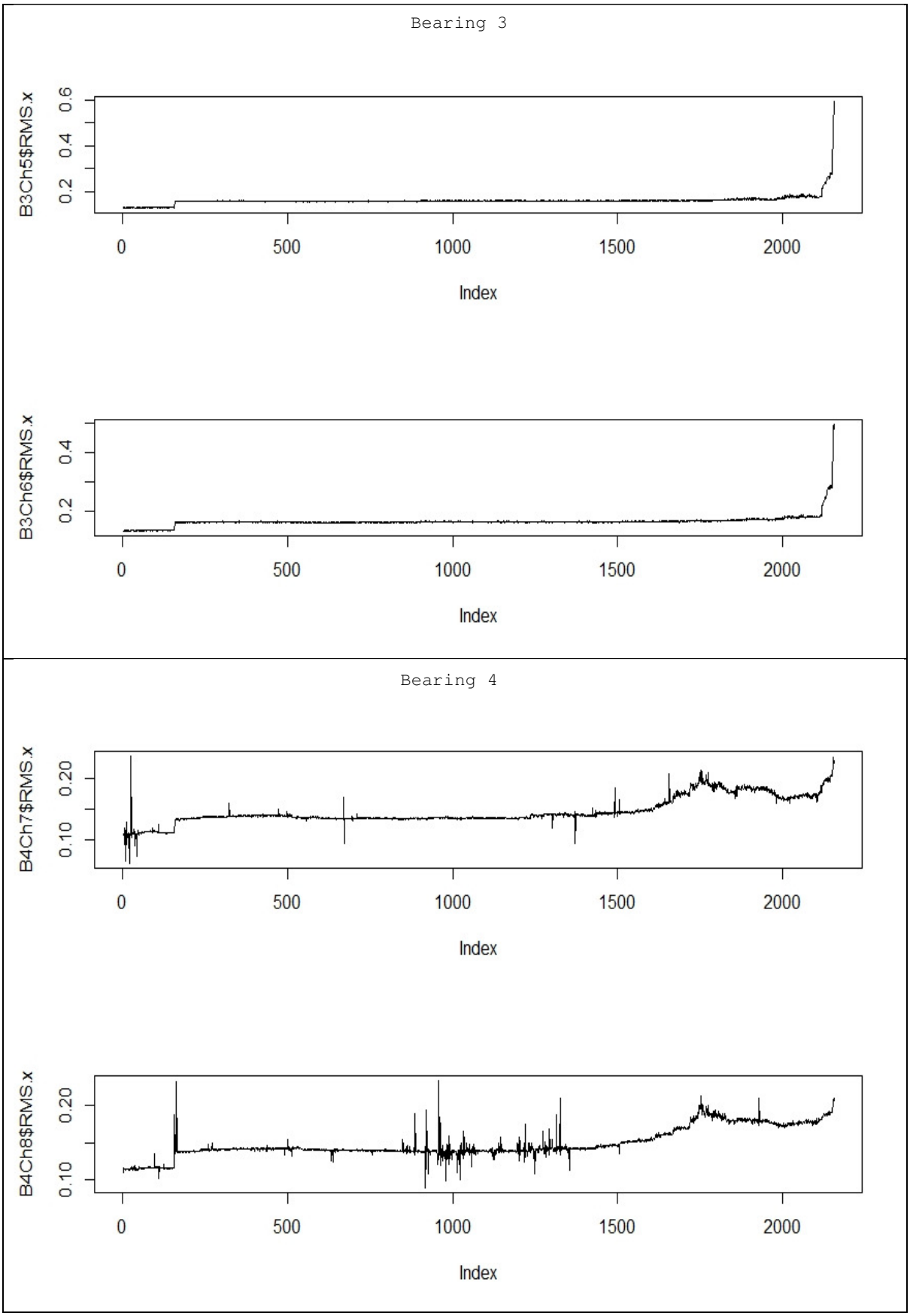


Figure 9.2d: Plot of RMS of all four bearings Training Dataset

Figure RMS Plots for Case Study Dataset 2




Bearing 3
B3Ch5$RMS.x
0.2
0.4
0.6
0
500
1000
1500
2000
Index
B3Ch6$RMS.x
0.2
0.4
0
500
1000
1500
2000
Index
Bearing 4
B4Ch7$RMS.x
0.10
0.20
0
500
1000
1500
2000
Index
B4Ch8$RMS.x
0.10
0.20
0
500
1000
1500
2000
Index

Final R- Code Training Dataset

```
#-----Section 01-----------------------------------------
# get the data
# set working directory

setwd()
getwd()
files <-list.files()
#--------------Section 02----------
# Import first data file

Trainingdir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Training Dataset/test2/"
data <- read.table(paste0(Trainingdir,"2004.02.18.15.22.39"), header=FALSE, sep="\t")

# Re-name column names
colnames(data) <- c("Bearing1","Bearing2","Bearing3","Bearing4")
# explore the data
str(data)    #see the structure of the data
head(data)       # The top of the data
class(data)
sapply(data,class) #print out the class of variables

# Inspect up to twenty randomly selected data files using descriptive statistics
library(pastecs)
stat.desc(data)

# Inspect the spread using inter quitile range
IQR(data$Bearing1)
IQR(data$Bearing2)
IQR(data$Bearing3)
IQR(data$Bearing4)

#skewness  to measure asymmetry
library(moments)
skewness(data)

#kurtosis to measure peakedness compare with a gausian distribution
kurtosis(data)
```

```
# visualise bivariate relationships among transformed: scatterplot matrix
pairs(data)

# more informative scatterplot matrix
library(psych)
pairs.panels(data)

# Correlations/covariances among numeric variables in
# Correlations with significance levels
library(Hmisc)

# type can be pearson or spearman
rcorr(as.matrix(data))

# Visual inspection using histogram for checking normality
library(MVN)
par(mfrow=c(2,2))
hist(data, type = "histogram") # creates univariate histograms
mtext("Histogram-Plot: Training Dataset", line = 0.5, outer = TRUE)

# Visual inspection using boxplot for checking normality
par(mfrow=c(1,1))
boxplot(data, horizontal=TRUE)

#qqnorm() plots for linearity
par(mfrow=c(1,4))
qqnorm(data$Bearing1) # creates univariate qqplot
qqline(data$Bearing1, col = "red", lwd = 2)
qqnorm(data$Bearing2)
qqline(data$Bearing2, col = "red", lwd = 2)
qqnorm(data$Bearing3)
qqline(data$Bearing3, col = "red", lwd = 2)
qqnorm(data$Bearing4)
qqline(data$Bearing4, col = "red", lwd = 2)

## Generate sequence plot of the data.
a = data$Bearing1
```

```
par(mfrow = c(1, 4),
    oma = c(0, 0, 2, 0),
    mar = c(5.1, 4.1, 2.1, 2.1))
plot(a,ylab="A",xlab="Bearing1")
b = data$Bearing2
plot(b,ylab="B",xlab="Bearing2")
c = data$Bearing3
plot(c,ylab="C",xlab="Bearing3")
d = data$Bearing4
plot(d,ylab="D",xlab="Bearing4")

#generate lag plot
plot(a,lag(a),xlab="Bearing 1[i-1]",ylab="Bearing 1[i]")
plot(b,lag(b),xlab="Bearing 2[i-1]",ylab="Bearing 2[i]")
plot(c,lag(c),xlab="Bearing 3[i-1]",ylab="Bearing 3[i]")
plot(d,lag(d),xlab="Bearing 4[i-1]",ylab="Bearing 4[i]")
mtext("Lag-Plot: Training Dataset", line = 0.5, outer = TRUE)

# Slice data into chunks of 2000 and analyse
data2000 <-data[c(1:2000),]
data4000 <-data[c(1:4000),]
data6000 <-data[c(1:6000),]
data8000 <-data[c(1:8000),]

# boxplot of sliced data
boxplot(data2000, horizontal=TRUE)
boxplot(data4000, horizontal=TRUE)
boxplot(data6000, horizontal=TRUE)
boxplot(data8000, horizontal=TRUE)

#--------------Section 04 Calculating and Ploting Key Frequencies using Modified Carsons code ----------

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
```

```
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

B1.fft.amps <- fft.spectrum(data$Bearing1)
features <- c(B1.fft.amps[freq2index(ftf)],
              B1.fft.amps[freq2index(bpfi)],
              B1.fft.amps[freq2index(bpfo)],
              B1.fft.amps[freq2index(bsf)])
features
# calculate Key frequencies

# Strongest frequencies
n <- 5
frequencies <- seq(0, 10000, length.out=length(B1.fft.amps))
sorted <- sort.int(B1.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
features <- append(features, frequencies[top.ind]) # convert indexes to frequencies
```

```
# Power in frequency bands
vhf <- freq2index(6000):length(B1.fft.amps)    # 6kHz plus
hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
lf <- 0:(freq2index(1250)-1)                # forcing frequency band

powers <- c(sum(B1.fft.amps[vhf]), sum(B1.fft.amps[hf]), sum(B1.fft.amps[mf]), sum(B1.fft.amps[lf]))
features <- append(features, powers)
features

# For Bearing 2
B2.fft.amps <- fft.spectrum(data$Bearing2)
B2.features <- c(B2.fft.amps[freq2index(ftf)],
                 B2.fft.amps[freq2index(bpfi)],
                 B2.fft.amps[freq2index(bpfo)],
                 B2.fft.amps[freq2index(bsf)])
B2.features
# calculate Key frequencies

# Strongest frequencies
n <- 5
B2.frequencies <- seq(0, 10000, length.out=length(B2.fft.amps))
sorted <- sort.int(B2.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
B2.features <- append(B2.features, B2.frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
B2.vhf <- freq2index(6000):length(B2.fft.amps)    # 6kHz plus
B2.hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
B2.mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
B2.lf <- 0:(freq2index(1250)-1)                # forcing frequency band

B2.powers <- c(sum(B2.fft.amps[vhf]), sum(B2.fft.amps[hf]), sum(B2.fft.amps[mf]), sum(B2.fft.amps[lf]))
B2.features <- append(B2.features, B2.powers)
B2.features

# For Bearing 3
```

```
B3.fft.amps <- fft.spectrum(data$Bearing3)
B3.features <- c(B3.fft.amps[freq2index(ftf)],
                 B3.fft.amps[freq2index(bpfi)],
                 B3.fft.amps[freq2index(bpfo)],
                 B3.fft.amps[freq2index(bsf)])
B3.features
# calculate Key frequencies

# Strongest frequencies
n <- 5
B3.frequencies <- seq(0, 10000, length.out=length(B3.fft.amps))
sorted <- sort.int(B3.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
B3.features <- append(B3.features, B3.frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
B3.vhf <- freq2index(6000):length(B3.fft.amps)    # 6kHz plus
B3.hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
B3.mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
B3.lf <- 0:(freq2index(1250)-1)                # forcing frequency band

B3.powers <- c(sum(B3.fft.amps[vhf]), sum(B3.fft.amps[hf]), sum(B3.fft.amps[mf]), sum(B3.fft.amps[lf]))
B3.features <- append(B3.features, B3.powers)
B3.features

# For Bearing 4

B4.fft.amps <- fft.spectrum(data$Bearing4)
B4.features <- c(B4.fft.amps[freq2index(ftf)],
                 B4.fft.amps[freq2index(bpfi)],
                 B4.fft.amps[freq2index(bpfo)],
                 B4.fft.amps[freq2index(bsf)])
B4.features
# calculate Key frequencies

# Strongest frequencies
n <- 5
B4.frequencies <- seq(0, 10000, length.out=length(B4.fft.amps))
```

```
sorted <- sort.int(B4.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
B4.features <- append(B4.features, B4.frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
B4.vhf <- freq2index(6000):length(B4.fft.amps)    # 6kHz plus
B4.hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
B4.mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
B4.lf <- 0:(freq2index(1250)-1)                # forcing frequency band

B4.powers <- c(sum(B4.fft.amps[vhf]), sum(B4.fft.amps[hf]), sum(B4.fft.amps[mf]), sum(B4.fft.amps[lf]))
B4.features <- append(B4.features, B4.powers)
B4.features

# Initial analysis using data mining techniques
summary(data$Bearing1)
summary(data$Bearing2)
summary(data$Bearing3)
summary(data$Bearing4)

# Plot variables
par(mfrow=c(4,1))
plot(data$Bearing1, t="l",ylab="frequency",ylim = c(-0.7,0.7),xlab="time index") # t="l" means line plot
plot(data$Bearing2, t="l",ylab="frequency",ylim = c(-0.7,0.7),xlab="time index")
plot(data$Bearing3, t="l",ylab="frequency",ylim = c(-0.7,0.7),xlab="time index")
plot(data$Bearing4, t="l",ylab="frequency",ylim = c(-0.7,0.7),xlab="time index")

# Apply feature extraction to reduce data
# 1. Format the full dataset
B1.fft <- fft(data$Bearing1)
# Ignore the 2nd half, which are complex conjugates of the 1st half,
# and calculate the Mod (magnitude of each complex number)
amplitude <- Mod(B1.fft[1:(length(B1.fft)/2)])

# Calculate the frequencies
B1freq <- seq(0, 10000, length.out=length(B1.fft)/2)

#Repeat for the other Bearings
```

```
B2.fft <- fft(data$Bearing2)
B3.fft <- fft(data$Bearing3)
B4.fft <- fft(data$Bearing4)

#Calculate amplitude
amplitude2 <- Mod(B2.fft[1:(length(B2.fft)/2)])
amplitude3 <- Mod(B3.fft[1:(length(B3.fft)/2)])
amplitude4 <- Mod(B4.fft[1:(length(B4.fft)/2)])

# Calculate frequencies
B2freq <- seq(0, 10000, length.out=length(B2.fft)/2)
B3freq <- seq(0, 10000, length.out=length(B3.fft)/2)
B4freq <- seq(0, 10000, length.out=length(B4.fft)/2)

# Plot
plot(amplitude ~ B1freq, t="l")
plot(amplitude2 ~ B2freq, t="l")
plot(amplitude3 ~ B3freq, t="l")
plot(amplitude4 ~ B4freq, t="l")

# 2.focus on the lower frequencies
plot(amplitude ~ Frequency, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

plot(amplitude2 ~ B2freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)

plot(amplitude3 ~ B3freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)

plot(amplitude4 ~ B4freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(1,1))
# For Bearing 1
B1freq <- seq(0, 500, length.out=length(data$Bearing1)/2)
```

```
plot(B1.fft.amps[1:(length(data$Bearing1)/2)] ~ B1freq, t="l", xlab="Frequency", ylab="Relative power",ylim =
c(0,700))
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Bearing 2
B2freq <- seq(0, 500, length.out=length(data$Bearing2)/2)
plot(B2.fft.amps[1:(length(data$Bearing2)/2)] ~ B2freq, t="l", xlab="Frequency",ylab="Relative power",ylim =
c(0,700))
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Bearing 3
B3freq <- seq(0, 500, length.out=length(data$Bearing3)/2)
plot(B3.fft.amps[1:(length(data$Bearing3)/2)] ~ B3freq, t="l", xlab="Frequency",ylab="Relative power",ylim =
c(0,700))
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Bearing 4
B4freq <- seq(0, 500, length.out=length(data$Bearing4)/2)
plot(B4.fft.amps[1:(length(data$Bearing4)/2)] ~ B4freq, t="l", xlab="Frequency",ylab="Relative power",ylim =
c(0,700))
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# 3.Tabulating the top 15 frequencies
sorted <- sort.int(amplitude, decreasing=TRUE, index.return=TRUE)
top15 <- sorted$ix[1:15] # indexes of the largest 15
B1.top15f <- B1freq[top15] # convert indexes to frequencies
```

```
B1.top15f

B2.sorted <- sort.int(amplitude2, decreasing=TRUE, index.return=TRUE)
B2.top15 <- B2.sorted$ix[1:15] # indexes of the largest 15
B2.top15f <- B2freq[top15] # convert indexes to frequencies
B2.top15f

B3.sorted <- sort.int(amplitude3, decreasing=TRUE, index.return=TRUE)
B3.top15 <- B3.sorted$ix[1:15] # indexes of the largest 15
B3.top15f <- B3freq[top15] # convert indexes to frequencies
B3.top15f

B4.sorted <- sort.int(amplitude4, decreasing=TRUE, index.return=TRUE)
B4.top15 <- B4.sorted$ix[1:15] # indexes of the largest 15
B4.top15f <- B4freq[top15] # convert indexes to frequencies
B4.top15f

fft.profile <- function (dataset, n)
{
  fft.data <- fft(dataset)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  amplitude <- Mod(fft.data[1:(length(fft.data)/2)])
  # Calculate the frequencies
  frequencies <- seq(0, 10000, length.out=length(fft.data)/2)

  sorted <- sort.int(amplitude, decreasing=TRUE, index.return=TRUE)
  top <- sorted$ix[1:n] # indexes of the largest n components
  return (frequencies[top]) # convert indexes to frequencies
}


# How many FFT components should we grab as features?
n <- 5

# Set up storage for bearing-grouped data
b1 <- matrix(nrow=0, ncol=(2*n+1))
b2 <- matrix(nrow=0, ncol=(2*n+1))
```

```
b3 <- matrix(nrow=0, ncol=(2*n+1))
b4 <- matrix(nrow=0, ncol=(2*n+1))

for (filename in list.files(Beari))
{
  cat("Processing file ", filename, "\n")

  timestamp <- as.character(strptime(filename, format="%Y.%m.%d.%H.%M.%S"))

  data <- read.table(paste0(Trainingdir, filename), header=FALSE, sep="\t")
  colnames(data) <- c("b1.x", "b2.x", "b3.x", "b4.x")

  # Bind the new rows to the bearing matrices
  b1 <- rbind(b1, c(timestamp, fft.profile(data$b1.x, n)))
  b2 <- rbind(b2, c(timestamp, fft.profile(data$b2.x, n)))
  b3 <- rbind(b3, c(timestamp, fft.profile(data$b3.x, n)))
  b4 <- rbind(b4, c(timestamp, fft.profile(data$b4.x, n)))

}

write.table(b1, file=paste0(Trainingdir, "../b1.csv"), sep=",", row.names=FALSE, col.names=FALSE)
write.table(b2, file=paste0(Trainingdir, "../b2.csv"), sep=",", row.names=FALSE, col.names=FALSE)
write.table(b3, file=paste0(Trainingdir, "../b3.csv"), sep=",", row.names=FALSE, col.names=FALSE)
write.table(b4, file=paste0(Trainingdir, "../b4.csv"), sep=",", row.names=FALSE, col.names=FALSE)

rm(list=ls())
#--------------Section 05 Final Feature Extraction Approach ----------
# Re-name column names
colnames(data) <- c("Bearing1","Bearing2","Bearing3","Bearing4")

library(e1071)

# Helper functions
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
```

```
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Bearing data
Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz

ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)


all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])
```

```
  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Set up storage for bearing-grouped data
b1m <- matrix(nrow=0, ncol=(1*23))
b2m <- matrix(nrow=0, ncol=(1*23))
b3m <- matrix(nrow=0, ncol=(1*23))
b4m <- matrix(nrow=0, ncol=(1*23))

# and for timestamps
timestamp <- vector()

for (filename in list.files(Trainingdir))
{
  cat("Processing file ", filename, "\n")

  ts <- as.character(strptime(filename, format="%Y.%m.%d.%H.%M.%S"))

  data <- read.table(paste0(Trainingdir, filename), header=FALSE, sep="\t")
  colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")
```

```
  # Bind the new rows to the bearing matrices
  b1m <- rbind(b1m, c(all.features(data$B1Ch1)))
  b2m <- rbind(b2m, c(all.features(data$B2Ch2)))
  b3m <- rbind(b3m, c(all.features(data$B3Ch3)))
  b4m <- rbind(b4m, c(all.features(data$B4Ch4)))

  timestamp <- c(timestamp, ts)
}

cnames <- c("Min.x", "Qu.1.x", "Median.x", "Qu.3.x", "Max.x", "Mean.x", "SD.x", "Skew.x", "Kurt.x", "RMS.x",
"FTF.x", "BPFI.x", "BPFO.x", "BSF.x", "F1.x", "F2.x", "F3.x", "F4.x", "F5.x", "VHF.pow.x", "HF.pow.x",
"MF.pow.x", "LF.pow.x")
colnames(b1m) <- cnames
colnames(b2m) <- cnames
colnames(b3m) <- cnames
colnames(b4m) <- cnames
B1Ch1 <- data.frame(timestamp, b1m)
B2Ch2 <- data.frame(timestamp, b2m)
B3Ch3 <- data.frame(timestamp, b3m)
B4Ch4 <- data.frame(timestamp, b4m)

write.table(B1Ch1, file=paste0(Trainingdir, "../B1Ch1_all.csv"), sep=",", row.names=FALSE)
write.table(B2Ch2, file=paste0(Trainingdir, "../B2Ch2_all.csv"), sep=",", row.names=FALSE)
write.table(B3Ch3, file=paste0(Trainingdir, "../B3Ch3_all.csv"), sep=",", row.names=FALSE)
write.table(B4Ch4, file=paste0(Trainingdir, "../B4Ch4_all.csv"), sep=",", row.names=FALSE)

rm(list=ls())

#--------------Section 06 Inspect Bearing-specific Datsets----------
#Reload New Bearing-Specific Datasets and inspect

B1Ch1 <- read.table(file=paste0(Trainingdir, "../B1Ch1_all.csv"), sep=",", header=FALSE)
B2Ch2 <- read.table(file=paste0(Trainingdir, "../B2Ch2_all.csv"), sep=",", header=FALSE)
B3Ch3 <- read.table(file=paste0(Trainingdir, "../B3Ch3_all.csv"), sep=",", header=FALSE)
B4Ch4 <- read.table(file=paste0(Trainingdir, "../B4Ch4_all.csv"), sep=",", header=FALSE)
head(B1Ch1)
str(B1Ch1)
```

```
# covert Factor to Numeric
B1Ch1 <- read.csv(file = '../B1Ch1_all.csv', stringsAsFactors = TRUE)
str(B1Ch1)
B2Ch2 <- read.csv(file = '../B2Ch2_all.csv', stringsAsFactors = TRUE)
B3Ch3 <- read.csv(file = '../B3Ch3_all.csv', stringsAsFactors = TRUE)
B4Ch4 <- read.csv(file = '../B4Ch4_all.csv', stringsAsFactors = TRUE)

#Descriptive statistics
library(pastecs)
stat.desc(B1Ch1)
stat.desc(B2Ch2)
stat.desc(B3Ch3)
stat.desc(B4Ch4)

#--------------Section 07 Change-point Analysis by Package changpoint----------
library(changepoint)
#Using change-point using statistical Pruned Exact Linear Time (PELT)
#Remember outer race risk (BPFO) in Bearing 1
# so plot BPFO

par(mfrow=c(2,2))

mvalue1 = cpt.mean(B1Ch1$BPFO, method="PELT")
mvalue1 = cpt.mean(B1Ch1[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue1)
vvalue1 = cpt.var(diff(B1Ch1[, 14]), method="PELT")
cpts(vvalue1)

B1.pelt <- cpt.var(diff(diff(B1Ch1[, 14]), method = "PELT"))
plot(B1.pelt, xlab = "Index")
logLik(B1.pelt)

#For Bearing 2
mvalue2 = cpt.mean(B2Ch2[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue2)
vvalue2 = cpt.var(diff(B2Ch2[, 14]), method="PELT")
cpts(vvalue2)
```

```
B2.pelt <- cpt.var(diff(diff(B2Ch2[, 14]), method = "PELT"))
plot(B2.pelt, xlab = "Index")
logLik(B2.pelt)

#For Bearing 3
mvalue3 = cpt.mean(B3Ch3[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue3)
vvalue3 = cpt.var(diff(B3Ch3[, 14]), method="PELT")
cpts(vvalue3)

B3.pelt <- cpt.var(diff(diff(B3Ch3[, 14]), method = "PELT"))
plot(B3.pelt, xlab = "Index")
logLik(B3.pelt)

#For Bearing 4
mvalue4 = cpt.mean(B4Ch4[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue4)
vvalue4 = cpt.var(diff(B4Ch4[, 14]), method="PELT")
cpts(vvalue4)

B4.pelt <- cpt.var(diff(diff(B4Ch4[, 14]), method = "PELT"))
plot(B4.pelt, xlab = "Index")
logLik(B4.pelt)

#--------------Section 08 Change-point Analysis by Package strucchange----------
library(strucchange)
par(mfrow=c(4,2))

#Bearing 1
B1.ts<- ts(B1Ch1[, 14],frequency=1)
B1.bp <-breakpoints((B1.ts~1))
B1.bp
summary(B1.bp)
plot(B1.bp)

# plot data with breakpoint times
plot(B1.ts)
lines(fitted(B1.bp, breaks = 1), col = 4)
```

```
lines(confint(B1.bp, breaks = 1))

#Bearing 2
B2.ts<- ts(B2Ch2[, 14],frequency=1)
B2.bp <-breakpoints((B2.ts~1))
B2.bp
summary(B2.bp)
plot(B2.bp)

# plot data with breakpoint times
plot(B2.ts)
lines(fitted(B2.bp, breaks = 0), col = 4)
lines(confint(B2.bp, breaks = 0))

#Bearing 3
B3.ts<- ts(B3Ch3[, 14],frequency=1)
B3.bp <-breakpoints((B3.ts~1))
B3.bp
summary(B3.bp)
plot(B3.bp)

# plot data with breakpoint times
plot(B3.ts)
lines(fitted(B3.bp, breaks = 0), col = 4)
lines(confint(B3.bp, breaks = 0))

#Bearing 4
B4.ts<- ts(B4Ch4[, 14],frequency=1)
B4.bp <-breakpoints((B4.ts~1))
B4.bp
summary(B4.bp)
plot(B4.bp)

# plot data with breakpoint times
plot(B4.ts)
lines(fitted(B4.bp, breaks = 0), col = 4)
lines(confint(B4.bp, breaks = 0))
```

```
#F-stats
par(mfrow=c(2,2))
B1.Fstats <- Fstats((B1Ch1[, 14]) ~ 1)
plot(B1.Fstats)

B2.Fstats <- Fstats((B2Ch2[, 14]) ~ 1)
plot(B2.Fstats)

B3.Fstats <- Fstats((B3Ch3[, 14]) ~ 1)
plot(B3.Fstats)

B4.Fstats <- Fstats((B4Ch4[, 14]) ~ 1)
plot(B4.Fstats)

# Significant test p-value
sctest(B1.Fstats, type = "supF")
sctest(B2.Fstats, type = "supF")
sctest(B3.Fstats, type = "supF")
sctest(B4.Fstats, type = "supF")

#--------------Section 09 Explore & Compare Test files at the Cpts-----------

#----Import & first data file (healthy file)---------
data <- read.table(paste0(Trainingdir,"2004.02.18.15.22.39"), header=FALSE, sep="\t")
head(data)

# Re-name column names
colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")
# Apply feature extraction to reduce data and plot
# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)

Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz
ratio <- Bd/Pd * cos(a)
```

```
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

B1.fft.amps <- fft.spectrum(data$B1Ch1)
features <- c(B1.fft.amps[freq2index(ftf)],
              B1.fft.amps[freq2index(bpfi)],
              B1.fft.amps[freq2index(bpfo)],
              B1.fft.amps[freq2index(bsf)])
features

# 1. Format the full dataset
B1.fft <- fft(data$B1Ch1)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
amplitude <- Mod(B1.fft[1:(length(B1.fft)/2)])

# Calculate the frequencies
B1freq <- seq(0, 10000, length.out=length(B1.fft)/2)

# Plot
plot(amplitude ~ B1freq, t="l")
```

```
# 2.focus on the lower frequencies
plot(amplitude ~ B1freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(3,1))
# For Bearing 1
B1freq <- seq(0, 310, length.out=length(data$B1Ch1)/2)
plot(B1.fft.amps[1:(length(data$B1Ch1)/2)] ~ B1freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#-----------Import file representing Changepoint (PELT) 968 ------

# Get timestamp of file representing row 968
B1Ch1[968,]
# timestamp = 2004-02-19 03:42:39; read test file 2004.02.19.03.42.39
data1 <- read.table(paste0(Trainingdir,"2004.02.19.03.42.39"), header=FALSE, sep="\t")
head(data1)

# Re-name column names
colnames(data1) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$B1Ch1)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
              d1B1.fft.amps[freq2index(bpfi)],
              d1B1.fft.amps[freq2index(bpfo)],
              d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$B1Ch1)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])
```

```
# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# Plot
plot(d1amplitude ~ d1B1freq, t="l")

# 2.focus on the lower frequencies
plot(d1amplitude ~ d1B1freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(3,1))
# For Bearing 1
d1B1freq <- seq(0, 310, length.out=length(data1$B1Ch1)/2)
plot(d1B1.fft.amps[1:(length(data1$B1Ch1)/2)] ~ d1B1freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#---------Import Test files @ Changepoint (Strucchange) = 837---------
# Get timestamp of file
B1Ch1[837,]
# timestamp = 2004-02-18 05:52:39; read test file
data2 <- read.table(paste0(Trainingdir,"2004.02.18.05.52.39"), header=FALSE, sep="\t")
head(data2)

# Re-name column names
colnames(data2) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")
# Re-name column names
colnames(data2) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

# Apply feature extraction to reduce data and plot
d2B1.fft.amps <- fft.spectrum(data2$B1Ch1)
d2features <- c(d2B1.fft.amps[freq2index(ftf)],
                d2B1.fft.amps[freq2index(bpfi)],
                d2B1.fft.amps[freq2index(bpfo)],
                d2B1.fft.amps[freq2index(bsf)])
```

```
d2features

# 1. Format the full dataset
d2B1.fft <- fft(data2$B1Ch1)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d2amplitude <- Mod(d2B1.fft[1:(length(d2B1.fft)/2)])

# Calculate the frequencies
d2B1freq <- seq(0, 10000, length.out=length(d2B1.fft)/2)

# Plot
plot(d2amplitude ~ d2B1freq, t="l")

# 2.focus on the lower frequencies
plot(d2amplitude ~ d2B1freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(4,1))
# For Bearing 1
d2B1freq <- seq(0, 310, length.out=length(data2$B1Ch1)/2)
plot(d2B1.fft.amps[1:(length(data2$B1Ch1)/2)] ~ d1B1freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For clarity, zoom in to frequencies up to 236.403-236.405Hz & compare
par(mfrow=c(3,1))
B1freq <- seq(0, 310, length.out=length(data$B1Ch1)/2)
plot(B1.fft.amps[1:(length(data$B1Ch1)/2)] ~ B1freq, t="l", ylim= c(0,150),ylab="Relative power")
abline(v=bpfo,col="blue",lty=3)

d1B1freq <- seq(236.40345, 236.4035, length.out=length(data1$B1Ch1)/2)
plot(d1B1.fft.amps[1:(length(data1$B1Ch1)/2)] ~ d1B1freq, t="l", ylim= c(0,150), ylab="Relative power")
abline(v=bpfo,col="blue",lty=3)

d2B1freq <- seq(236.40345, 236.4035, length.out=length(data2$B1Ch1)/2)
```

```
plot(d2B1.fft.amps[1:(length(data2$B1Ch1)/2)] ~ d1B1freq, t="l", ylim= c(0,150), ylab="Relative power")
abline(v=bpfo,col="blue",lty=3)

#---------------Section 10 Investigating Change-points with trend in RMS of Bearing -------------
par(mfrow=c(1,1))
plot(B1Ch1$RMS.x, t="l",ylab="RMS")
plot(B2Ch2$RMS.x, t="l",ylab="RMS")
plot(B3Ch3$RMS.x, t="l",ylab="RMS")
plot(B4Ch4$RMS.x, t="l",ylab="RMS")

#-----Section 11 Re-load and combine Bearing-specific datasets------------------

library(car)

# Create data frame with columns of interest using column indices
# Displays column 12-15
dfnew1 <- B1Ch1[,c(12:15)]
dfnew2 <- B2Ch2[,c(12:15)]
dfnew3 <- B3Ch3[,c(12:15)]
dfnew4 <- B4Ch4[,c(12:15)]

# Re-name column names
colnames(dfnew1) <- c("FTF.B1", "BPFI.B1", "BPFO.B1"," BSF.B1")
colnames(dfnew2) <- c("FTF.B2", "BPFI.B2", "BPFO.B2"," BSF.B2")
colnames(dfnew3) <- c("FTF.B3", "BPFI.B3", "BPFO.B3"," BSF.B3")
colnames(dfnew4) <- c("FTF.B4", "BPFI.B4", "BPFO.B4"," BSF.B4")

# Merge the  data frames
# Use the cbind function to combine data frames side-by-side:
dfnew <-cbind(dfnew1,dfnew2,dfnew3,dfnew4)
dfnew

#-----Section 12 Investigate Correlation for Interaction Effect------------------------------

# Creat a sub-set for interactions up to time index 968
BPFO968 <-dfnew[c(1:968),c(3,7,11,15)]
str(BPFO968)
```

```
# Inspect
boxplot(BPFO968, horizontal=TRUE)
pairs(BPFO968,panel=panel.smooth)

# Investigate correlation
library("PerformanceAnalytics")
chart.Correlation(BPFO968, histogram=TRUE, pch=19)

# Creat a sub-set for time index 837
BPFO837 <-dfnew[c(1:837),c(3,7,11,15)]
str(BPFO837)

boxplot(BPFO837, horizontal=TRUE)
pairs(BPFO837,panel=panel.smooth)

# Investigate correlation
chart.Correlation(BPFO837, histogram=TRUE, pch=19)


# Apply Decision Tree models to determine predictors and moderators for regression models
library(tree)

# For time index 837
Tmodel837<-tree(BPFO.B1~.,data=BPFO837)
plot(Tmodel837)
text(Tmodel837)

# For time index 968
Tmodel968<-tree(BPFO.B1~.,data=BPFO968)
plot(Tmodel968)
text(Tmodel968)

# Creat Regression model to investigate interaction effect

# for time index 837
model.1 <- lm(BPFO.B1 ~ BPFO.B3+BPFO.B2, data=BPFO837)

model.2 <- lm(BPFO.B1 ~ BPFO.B3*BPFO.B2,data=BPFO837)
```

```
# for time index 968
model.3 <- lm(BPFO.B1 ~ BPFO.B3+BPFO.B2, data=BPFO968)

model.4 <- lm(BPFO.B1 ~ BPFO.B3*BPFO.B2,data=BPFO968)

#Show the results
library(stargazer)
stargazer(model.1,model.2,model.3,model.4,type="text",
          column.labels = c("model.1","model.2","model.3","model.4"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar interaction models
anova(model.1, model.2)
anova(model.3, model.4)

# Investigate significant interaction
install.packages("devtools")
devtools::install_github("jacob-long/jtools")

library(jtools)
library(interactions)
library(ggplot2)

# for time index 837
sim_slopes(model.2, pred = BPFO.B3, modx = BPFO.B2, jnplot = TRUE)

# Visualize interaction effect
interact_plot(model.2, pred = BPFO.B3, modx = BPFO.B2, interval = TRUE)
probe_interaction(model.2, pred = BPFO.B3, modx = BPFO.B2, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# for time index 968
sim_slopes(model.4, pred = BPFO.B3, modx = BPFO.B2, jnplot = TRUE)
```

```
# Visualize interaction effect
interact_plot(model.4, pred = BPFO.B3, modx = BPFO.B2, interval = TRUE)
probe_interaction(model.4, pred = BPFO.B3, modx = BPFO.B2, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)


#----------Section 13 Association between Features----------------------
# Creat subset for bearing defect parameters
BOne837 <-dfnew[c(1:837),c(2:4)]
str(BOne837)
colnames(BOne837) <- c("BPFI_B1", "BPFO_B1","BSF_B1")
str(BOne837)
head(BOne837)

BOne968 <-dfnew[c(1:968),c(2:4)]
str(BOne968)
colnames(BOne968) <- c("BPFI_B1", "BPFO_B1","BSF_B1")
str(BOne968)
head(BOne968)

# Creat a tree models to determine predictor and moderator Bearings
# time index 837
TmodelBOne837<-tree(BPFO_B1~.,data=BOne837)
plot(TmodelBOne837)
text(TmodelBOne837)

#time index 968
TmodelBOne968<-tree(BPFO_B1~.,data=BOne968)
plot(TmodelBOne968)
text(TmodelBOne968)

# singlenode trees therefore no further investigation

#-----Section 14-----------------------------------------
# remove all variables from the environment
rm(list=ls())
```

Final R- Code Case Study Dataset 1

```
#-----Section 01-----------------------------------------
# set working directory
setwd("C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 1/test 3/test 3/")
getwd()
files <-list.files()
#--------------Section 02----------
# Import first data file
CSDOnedir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 1/test 3/test 3/"
data <- read.table(paste0(CSDOnedir, "2004.03.04.09.27.46"), header=FALSE, sep="\t")
head(data)
#--------------Section 03----------
# Randomly Select and inspect up to twenty randomly data files using descriptive statistics
library(pastecs)

#Descriptive statistics
stat.desc(data)

# Re-name column names
colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

#--------------Section 04 Apply Feature Extraction --------------------
library(e1071)

# Helper functions
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}
```

```
# Bearing data
Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz

ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)


all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies
```

```
  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Set up storage for bearing-grouped data
b1m <- matrix(nrow=0, ncol=(1*23))
b2m <- matrix(nrow=0, ncol=(1*23))
b3m <- matrix(nrow=0, ncol=(1*23))
b4m <- matrix(nrow=0, ncol=(1*23))

# and for timestamps
timestamp <- vector()

for (filename in list.files(CSDOnedir))
{
  cat("Processing file ", filename, "\n")

  ts <- as.character(strptime(filename, format="%Y.%m.%d.%H.%M.%S"))

  data <- read.table(paste0(CSDOnedir, filename), header=FALSE, sep="\t")
  colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

  # Bind the new rows to the bearing matrices
  b1m <- rbind(b1m, c(all.features(data$B1Ch1)))
  b2m <- rbind(b2m, c(all.features(data$B2Ch2)))
  b3m <- rbind(b3m, c(all.features(data$B3Ch3)))
  b4m <- rbind(b4m, c(all.features(data$B4Ch4)))

  timestamp <- c(timestamp, ts)
}
```

```
cnames <- c("Min.x", "Qu.1.x", "Median.x", "Qu.3.x", "Max.x", "Mean.x", "SD.x", "Skew.x", "Kurt.x", "RMS.x",
"FTF.x", "BPFI.x", "BPFO.x", "BSF.x", "F1.x", "F2.x", "F3.x", "F4.x", "F5.x", "VHF.pow.x", "HF.pow.x",
"MF.pow.x", "LF.pow.x")
colnames(b1m) <- cnames
colnames(b2m) <- cnames
colnames(b3m) <- cnames
colnames(b4m) <- cnames
B1Ch1 <- data.frame(timestamp, b1m)
B2Ch2 <- data.frame(timestamp, b2m)
B3Ch3 <- data.frame(timestamp, b3m)
B4Ch4 <- data.frame(timestamp, b4m)

write.table(B1Ch1, file=paste0(CSDOnedir, "../B1Ch1_all.csv"), sep=",", row.names=FALSE)
write.table(B2Ch2, file=paste0(CSDOnedir, "../B2Ch2_all.csv"), sep=",", row.names=FALSE)
write.table(B3Ch3, file=paste0(CSDOnedir, "../B3Ch3_all.csv"), sep=",", row.names=FALSE)
write.table(B4Ch4, file=paste0(CSDOnedir, "../B4Ch4_all.csv"), sep=",", row.names=FALSE)

# remove all variables from the environment
rm(list=ls())

#--------------Section 05 Re-load and Inspect Bearing-specific Datsets----------

B1Ch1 <- read.table(file=paste0(CSDOnedir, "../B1Ch1_all.csv"), sep=",", header=FALSE)
B2Ch2 <- read.table(file=paste0(CSDOnedir, "../B2Ch2_all.csv"), sep=",", header=FALSE)
B3Ch3 <- read.table(file=paste0(CSDOnedir, "../B3Ch3_all.csv"), sep=",", header=FALSE)
B4Ch4 <- read.table(file=paste0(CSDOnedir, "../B4Ch4_all.csv"), sep=",", header=FALSE)

#Inspect table
head(B1Ch1)
str(B1Ch1)
# covert Factor to Numeric
B1Ch1 <- read.csv(file = '../B1Ch1_all.csv', stringsAsFactors = TRUE)
str(B1Ch1)
B2Ch2 <- read.csv(file = '../B2Ch2_all.csv', stringsAsFactors = TRUE)
B3Ch3 <- read.csv(file = '../B3Ch3_all.csv', stringsAsFactors = TRUE)
B4Ch4 <- read.csv(file = '../B4Ch4_all.csv', stringsAsFactors = TRUE)
```

```
#Descriptive statistics
library(pastecs)
stat.desc(B1Ch1)
stat.desc(B2Ch2)
stat.desc(B3Ch3)
stat.desc(B4Ch4)

#--------------Section 06 Change-point Analysis by Package changpoint----------
library(changepoint)
#Using change-point using statistical Pruned Exact Linear Time (PELT)
#Remember outer race risk (BPFO) in Bearing 3
# so plot BPFO

mvalue1 = cpt.mean(B1Ch1$BPFO, method="PELT")
mvalue1 = cpt.mean(B1Ch1[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue1)
vvalue1 = cpt.var(diff(B1Ch1[, 14]), method="PELT")
cpts(vvalue1)

B1.pelt <- cpt.var(diff(diff(B1Ch1[, 14]), method = "PELT"))
plot(B1.pelt, xlab = "Index")
logLik(B1.pelt)

#For Bearing 2
mvalue2 = cpt.mean(B2Ch2[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue2)
vvalue2 = cpt.var(diff(B2Ch2[, 14]), method="PELT")
cpts(vvalue2)

B2.pelt <- cpt.var(diff(diff(B2Ch2[, 14]), method = "PELT"))
plot(B2.pelt, xlab = "Index")
logLik(B2.pelt)

#For Bearing 3
mvalue3 = cpt.mean(B3Ch3[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue3)
vvalue3 = cpt.var(diff(B3Ch3[, 14]), method="PELT")
cpts(vvalue3)
```

```
B3.pelt <- cpt.var(diff(diff(B3Ch3[, 14]), method = "PELT"))
plot(B3.pelt, xlab = "Index")
logLik(B3.pelt)

#For Bearing 4
mvalue4 = cpt.mean(B4Ch4[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue4)
vvalue4 = cpt.var(diff(B4Ch4[, 14]), method="PELT")
cpts(vvalue4)

B4.pelt <- cpt.var(diff(diff(B4Ch4[, 14]), method = "PELT"))
plot(B4.pelt, xlab = "Index")
logLik(B4.pelt)

par(mfrow=c(2,2))
#--------------Section 07 Change-point using Package strucchange----------
library(strucchange)
par(mfrow=c(4,2))
#Bearing 1
B1.ts<- ts(B1Ch1[, 14],frequency=1)
B1.bp <-breakpoints((B1.ts~1))
B1.bp
summary(B1.bp)
plot(B1.bp)

# plot data with breakpoint times
plot(B1.ts)
lines(fitted(B1.bp, breaks = 1), col = 4)
lines(confint(B1.bp, breaks = 1))

#Bearing 2
B2.ts<- ts(B2Ch2[, 14],frequency=1)
B2.bp <-breakpoints((B2.ts~1))
B2.bp
summary(B2.bp)
plot(B2.bp)
```

```
# plot data with breakpoint times
plot(B2.ts)
lines(fitted(B2.bp, breaks = 2), col = 4)
lines(confint(B2.bp, breaks = 2))

#Bearing 3
B3.ts<- ts(B3Ch3[, 14],frequency=1)
B3.bp <-breakpoints((B3.ts~1))
B3.bp
summary(B3.bp)
plot(B3.bp)

# plot data with breakpoint times
plot(B3.ts)
lines(fitted(B3.bp, breaks = 2), col = 4)
lines(confint(B3.bp, breaks = 2))

#Bearing 4
B4.ts<- ts(B4Ch4[, 14],frequency=1)
B4.bp <-breakpoints((B4.ts~1))
B4.bp
summary(B4.bp)
plot(B4.bp)

# plot data with breakpoint times
plot(B4.ts)
lines(fitted(B4.bp, breaks = 1), col = 4)
lines(confint(B4.bp, breaks = 1))

par(mfrow=c(2,2))
#F-stats and SupF not required

#---------------Section 08 Investigating change-points with plot of RMS ---------------
par(mfrow=c(1,1))
plot(B3Ch3$RMS.x, t="l",ylab="RMS")
plot(B1Ch1$RMS.x, t="l",ylab="RMS")
plot(B2Ch2$RMS.x, t="l",ylab="RMS")
plot(B4Ch4$RMS.x, t="l",ylab="RMS")
```

```
#---------Section 09 Creat New dataframe by Combining Bearing-specific Data Files---

#1: Create data frame with columns of interest
# Displays columns 12-15
dfnew1 <- B1Ch1[,c(12:15)]
dfnew2 <- B2Ch2[,c(12:15)]
dfnew3 <- B3Ch3[,c(12:15)]
dfnew4 <- B4Ch4[,c(12:15)]

# Re-name column names
colnames(dfnew1) <- c("FTF.B1", "BPFI.B1", "BPFO.B1","BSF.B1")
colnames(dfnew2) <- c("FTF.B2", "BPFI.B2", "BPFO.B2","BSF.B2")
colnames(dfnew3) <- c("FTF.B3", "BPFI.B3", "BPFO.B3"," BSF.B3")
colnames(dfnew4) <- c("FTF.B4", "BPFI.B4", "BPFO.B4"," BSF.B4")

# Merge the  data frames
# Use the cbind function to combine data frames side-by-side:
dfnew <-cbind(dfnew1,dfnew2,dfnew3,dfnew4)
dfnew

#----------Section 10 Investigate Correlations-------------

# Investigate features for interaction effect

# time index 6324
BPFO6324 <-dfnew[c(1:6324),c(3,7,11,15)]
str(BPFO6324)

# Visualise correlation matrix
library("PerformanceAnalytics")
chart.Correlation(BPFO6324, histogram=TRUE, pch=19)

# time index 3270
BPFO3270 <-dfnew[c(1:3270),c(3,7,11,15)]
str(BPFO3270)
chart.Correlation(BPFO3270, histogram=TRUE, pch=19)
```

```
# time index 2778
# Any interaction all variables 2778
BPFO2778 <-dfnew[c(1:2778),c(3,7,11,15)]
str(BPFO2778)
chart.Correlation(BPFO2778, histogram=TRUE, pch=19)

# time index 5374
BPFO5374 <-dfnew[c(1:5374),c(3,7,11,15)]
str(BPFO5374)
chart.Correlation(BPFO5374, histogram=TRUE, pch=19)

# time index 6294
BPFO6294 <-dfnew[c(1:6294),c(3,7,11,15)]
str(BPFO6294)
chart.Correlation(BPFO6294, histogram=TRUE, pch=19)

#-----Section 11 Apply Decision Tree models-------------------------
library(tree)

# time index 2778
Tmodel2778<-tree(BPFO.B3~.,data=BPFO2778)
plot(Tmodel2778)
text(Tmodel2778)

# time index 3270
Tmodel3270<-tree(BPFO.B3~.,data=BPFO3270)
plot(Tmodel3270)
text(Tmodel3270)

# time index 5374
Tmodel5374<-tree(BPFO.B3~.,data=BPFO5374)
plot(Tmodel5374)
text(Tmodel5374)

# time index 6294
Tmodel6294<-tree(BPFO.B3~.,data=BPFO6294)
plot(Tmodel6294)
text(Tmodel6294)
```

```
#----------Section 12 Creat Regression model to investigate prediction-------------

# time index 2778
model.1 <- lm(BPFO.B3 ~ BPFO.B2, data=BPFO2778)

# time index 3270
model.2 <- lm(BPFO.B3 ~ BPFO.B2, data=BPFO3270)

# time index 5374
model.3 <- lm(BPFO.B3 ~ BPFO.B2, data=BPFO5374)

# time index 6294
model.4 <- lm(BPFO.B3 ~ BPFO.B2, data=BPFO6294)

#Show the results
library(stargazer)
stargazer(model.1,model.2,model.3,model.4,type="text",
          column.labels = c("model.1","model.2","model.3","model.4"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# No further investigation

#----------Section 13 Investigate Association between Features----
# time index 2778
BThree2778 <-dfnew[c(1:2778),c(9:12)]
colnames(BThree2778) <- c("FTF_B3", "BPFI_B3", "BPFO_B3","BSF_B3")
str(BThree2778)
head(BThree2778)

# Creat a tree model
TmodelBthree2778 <-tree(BPFO_B3~.,data=Bearing3a)
plot(TmodelBthree2778)
text(TmodelBthree2778)
```

```
# time index 3270
BThree3270 <-dfnew[c(1:3270),c(9:12)]
colnames(BThree3270) <- c("FTF_B3", "BPFI_B3", "BPFO_B3","BSF_B3")
str(BThree3270)
head(BThree3270)

# Creat a tree model
TmodelBThree3270<-tree(BPFO_B3~.,data=BThree3270)
plot(TmodelBThree3270)
text(TmodelBThree3270)

# time index 5374
BThree5374 <-dfnew[c(1:5374),c(9:12)]
colnames(BThree5374) <- c("FTF_B3", "BPFI_B3", "BPFO_B3","BSF_B3")
str(BThree5374)
head(BThree5374)

# Creat a tree model
TmodelBThree5374<-tree(BPFO_B3~.,data=BThree5374)
plot(TmodelBThree5374)
text(TmodelBThree5374)

# time index 6294
BThree6294 <-dfnew[c(1:6294),c(9:12)]
colnames(BThree6294) <- c("FTF_B3", "BPFI_B3", "BPFO_B3","BSF_B3")
str(BThree6294)
head(BThree6294)

# Creat a tree model
TmodelBThree6294<-tree(BPFO_B3~.,data=BThree6294)
plot(TmodelBThree6294)
text(TmodelBThree6294)

#---------14 Regression Model for Interactions------------------
# time index 2778
model.1 <- lm(BPFO_B3 ~ BSF_B3+BPFI_B3, data=Bearing3a)

model.2 <- lm(BPFO_B3 ~ BSF_B3*BPFI_B3,data=Bearing3a)
```

```
#Show the results
library(stargazer)
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# performing slopes analysis
# No significant interaction effect therefore no slope analysis


# For change-point 3270
# Model 1: the defect of Bearing 3 using BSF as predictor
model.1 <- lm(BPFO_B3 ~ BSF_B3+BPFI_B3, data=Bearing3b)

# Model 2: Defect of Bearing 1 (BPFO) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFO_B3 ~ BSF_B3*BPFI_B3,data=Bearing3b)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)


# performing slopes analysis
# No significant interaction effect therefore no slope analysis
```

```
# For change-point 5374
# Model 1: the defect of Bearing 3 using BSF as predictor
model.1 <- lm(BPFO_B3 ~ BSF_B3+BPFI_B3, data=Bearing3c)

# Model 2: Defect of Bearing 1 (BPFO) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFO_B3 ~ BSF_B3*BPFI_B3,data=Bearing3c)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# performing slopes analysis
# No significant interaction effect therefore no slope analysis

# For change-point 6294
# Model 1: the defect of Bearing 3 using BSF as predictor
model.1 <- lm(BPFO_B3 ~ BSF_B3+BPFI_B3, data=Bearing3d)

# Model 2: Defect of Bearing 1 (BPFO) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFO_B3 ~ BSF_B3*BPFI_B3,data=Bearing3d)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
```

```
anova(model.1, model.2)

# performing slopes analysis
# No significant interaction effect therefore no slope analysis

#-----Section 05----------------------------------------
# remove all variables from the environment
rm(list=ls())
```

Final R- Code Case Study Dataset 2

```
#-----Section 01-----------------------------------------
##
# get the data
getwd()
files <-list.files()
#--------------Section 02----------
# Initial analysis using
# data mining process to underst, clean, store,
# change data format.
# Import first data file

Bearing1dir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 2/Data/"
data <- read.table(paste0(Bearing1dir, "2003.11.25.10.47.32"), header=FALSE, sep="\t")
head(data)

# Re-name column names
colnames(data) <- c("B1Ch1","B1Ch2","B2Ch3","B2Ch4","B3Ch5","B3Ch6","B4Ch7","B4Ch8")

#Descriptive statistics
library(pastecs)
stat.desc(data)


# Creat channel specific dataset for the vertical and horizontal channels
library(e1071)

# Helper functions
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
```

```
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Bearing data
Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz

ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
```

```
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Set up storage for bearing-grouped data
# number of columns = 23,
# where 1 = column for time; n = 23 features
b1c1m <- matrix(nrow=0, ncol=(1*23))
b1c2m <- matrix(nrow=0, ncol=(1*23))
b2c3m <- matrix(nrow=0, ncol=(1*23))
b2c4m <- matrix(nrow=0, ncol=(1*23))
b3c5m <- matrix(nrow=0, ncol=(1*23))
b3c6m <- matrix(nrow=0, ncol=(1*23))
b4c7m <- matrix(nrow=0, ncol=(1*23))
b4c8m <- matrix(nrow=0, ncol=(1*23))

# and for timestamps
timestamp <- vector()

for (filename in list.files(Bearing1dir))
{
  cat("Processing file ", filename, "\n")

  ts <- as.character(strptime(filename, format="%Y.%m.%d.%H.%M.%S"))

  data <- read.table(paste0(Bearing1dir, filename), header=FALSE, sep="\t")
  colnames(data) <- c("B1Ch1","B1Ch2","B2Ch3","B2Ch4","B3Ch5","B3Ch6","B4Ch7","B4Ch8")
```

```
  # Bind the new rows to the bearing matrices
  b1c1m <- rbind(b1c1m, c(all.features(data$B1Ch1)))
  b1c2m <- rbind(b1c2m, c(all.features(data$B1Ch2)))
  b2c3m <- rbind(b2c3m, c(all.features(data$B2Ch3)))
  b2c4m <- rbind(b2c4m, c(all.features(data$B2Ch4)))
  b3c5m <- rbind(b3c5m, c(all.features(data$B3Ch5)))
  b3c6m <- rbind(b3c6m, c(all.features(data$B3Ch6)))
  b4c7m <- rbind(b4c7m, c(all.features(data$B4Ch7)))
  b4c8m <- rbind(b4c8m, c(all.features(data$B4Ch8)))
  timestamp <- c(timestamp, ts)
}

cnames <- c("Min.x", "Qu.1.x", "Median.x", "Qu.3.x", "Max.x", "Mean.x", "SD.x", "Skew.x",
            "Kurt.x", "RMS.x", "FTF.x", "BPFI.x", "BPFO.x", "BSF.x", "F1.x", "F2.x", "F3.x",
            "F4.x", "F5.x", "VHF.pow.x", "HF.pow.x", "MF.pow.x", "LF.pow.x")
colnames(b1c1m) <- cnames
colnames(b1c2m) <- cnames
colnames(b2c3m) <- cnames
colnames(b2c4m) <- cnames
colnames(b3c5m) <- cnames
colnames(b3c6m) <- cnames
colnames(b4c7m) <- cnames
colnames(b4c8m) <- cnames
B1Ch1 <- data.frame(timestamp, b1c1m)
B1Ch2 <- data.frame(timestamp, b1c2m)
B2Ch3 <- data.frame(timestamp, b2c3m)
B2Ch4 <- data.frame(timestamp, b2c4m)
B3Ch5 <- data.frame(timestamp, b3c5m)
B3Ch6 <- data.frame(timestamp, b3c6m)
B4Ch7 <- data.frame(timestamp, b4c7m)
B4Ch8 <- data.frame(timestamp, b4c8m)

write.table(B1Ch1, file=paste0(Bearing1dir, "../B1Ch1_all.csv"), sep=",", row.names=FALSE)
write.table(B1Ch2, file=paste0(Bearing1dir, "../B1Ch2_all.csv"), sep=",", row.names=FALSE)
write.table(B2Ch3, file=paste0(Bearing1dir, "../B2Ch3_all.csv"), sep=",", row.names=FALSE)
write.table(B2Ch4, file=paste0(Bearing1dir, "../B2Ch4_all.csv"), sep=",", row.names=FALSE)
write.table(B3Ch5, file=paste0(Bearing1dir, "../B3Ch5_all.csv"), sep=",", row.names=FALSE)
write.table(B3Ch6, file=paste0(Bearing1dir, "../B3Ch6_all.csv"), sep=",", row.names=FALSE)
```

```
write.table(B4Ch7, file=paste0(Bearing1dir, "../B4Ch7_all.csv"), sep=",", row.names=FALSE)
write.table(B4Ch8, file=paste0(Bearing1dir, "../B4Ch8_all.csv"), sep=",", row.names=FALSE)

# remove all variables from the environment
rm(list=ls())

# -----Section 03 Changepoint Analysis-----------------------------------------
#
#Reload Bearing Speific data files
Bearingdir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 2/Data/"
files <-list.files()

B1Ch1 <- read.table(file=paste0(Bearingdir, "../B1Ch1_all.csv"), sep=",", header=FALSE)
B1Ch2 <- read.table(file=paste0(Bearingdir, "../B1Ch2_all.csv"), sep=",", header=FALSE)
B2Ch3 <- read.table(file=paste0(Bearingdir, "../B2Ch3_all.csv"), sep=",", header=FALSE)
B2Ch4 <- read.table(file=paste0(Bearingdir, "../B2Ch4_all.csv"), sep=",", header=FALSE)
B3Ch5 <- read.table(file=paste0(Bearingdir, "../B3Ch5_all.csv"), sep=",", header=FALSE)
B3Ch6 <- read.table(file=paste0(Bearingdir, "../B3Ch6_all.csv"), sep=",", header=FALSE)
B4Ch7 <- read.table(file=paste0(Bearingdir, "../B4Ch7_all.csv"), sep=",", header=FALSE)
B4Ch8 <- read.table(file=paste0(Bearingdir, "../B4Ch8_all.csv"), sep=",", header=FALSE)

#Inspect table
head(B1Ch1)
str(B1Ch1)

# covert Factor to Numeric & inspect
B1Ch1 <- read.csv(file = '../B1Ch1_all.csv', stringsAsFactors = TRUE)
str(B1Ch1)
B1Ch2 <- read.csv(file = '../B1Ch2_all.csv', stringsAsFactors = TRUE)
B2Ch3 <- read.csv(file = '../B2Ch3_all.csv', stringsAsFactors = TRUE)
B2Ch4 <- read.csv(file = '../B2Ch4_all.csv', stringsAsFactors = TRUE)
B3Ch5 <- read.csv(file = '../B3Ch5_all.csv', stringsAsFactors = TRUE)
B3Ch6 <- read.csv(file = '../B3Ch6_all.csv', stringsAsFactors = TRUE)
B4Ch7 <- read.csv(file = '../B4Ch7_all.csv', stringsAsFactors = TRUE)
B4Ch8 <- read.csv(file = '../B4Ch8_all.csv', stringsAsFactors = TRUE)

#Descriptive statistics
library(pastecs)
```

```
stat.desc(B1Ch1)
stat.desc(B1Ch2)
stat.desc(B2Ch3)
stat.desc(B2Ch4)
stat.desc(B3Ch5)
stat.desc(B3Ch6)
stat.desc(B4Ch7)
stat.desc(B4Ch8)

# changepoint by PELT

# Ignore Bearing 1 and 2
library(changepoint)
par(mfrow=c(2,1))

# Inner race failure in Besring 3== investigate BPFI (V13)
# Bearing 3- Channel 5
vvalueb3c5 = cpt.var(diff(B3Ch5[, 13]), method="PELT")
cpts(vvalueb3c5)

B3c5.pelt <- cpt.var(diff(diff(B3Ch5[, 13]), method = "PELT"))
plot(B3c5.pelt, xlab = "Index")
logLik(B3c5.pelt)

#For Bearing 3- Channel 6
vvalueb3c6 = cpt.var(diff(B3Ch6[, 13]), method="PELT")
cpts(vvalueb3c6)

B3c6.pelt <- cpt.var(diff(diff(B3Ch6[, 13]), method = "PELT"))
plot(B3c6.pelt, xlab = "Index")
logLik(B3c6.pelt)

#Outer race failure observed in Bearing 4== investigate BPFO (V14)
#For Bearing 4 -Channel 7
vvalueb4c7 = cpt.var(diff(B4Ch7[, 14]), method="PELT")
cpts(vvalueb4c7)

B4c7.pelt <- cpt.var(diff(diff(B4Ch7[, 14]), method = "PELT"))
```

```
plot(B4c7.pelt, xlab = "Index")
logLik(B4c7.pelt)

#For Bearing 4- Channel 8
vvalueb4c8 = cpt.var(diff(B4Ch8[, 14]), method="PELT")
cpts(vvalueb4c8)

B4c8.pelt <- cpt.var(diff(diff(B4Ch8[, 14]), method = "PELT"))
plot(B4c8.pelt, xlab = "Index")
logLik(B4c8.pelt)

# Rollar element failure in Bearing 4 == investigate BSF (V15)
#For Bearing 4 -Channel 7
vvalueb4c7 = cpt.var(diff(B4Ch7[, 15]), method="PELT")
cpts(vvalueb4c7)

B4c7.pelt <- cpt.var(diff(diff(B4Ch7[, 15]), method = "PELT"))
plot(B4c7.pelt, xlab = "Index")
logLik(B4c7.pelt)

#For Bearing 4- Channel 8
vvalueb4c8 = cpt.var(diff(B4Ch8[, 15]), method="PELT")
cpts(vvalueb4c8)

B4c8.pelt <- cpt.var(diff(diff(B4Ch8[, 15]), method = "PELT"))
plot(B4c8.pelt, xlab = "Index")
logLik(B4c8.pelt)

#Change-point by Structure Change
library(strucchange)
par(mfrow=c(2,1))

# Inner race failure in Besring 3== investigate BPFI (V13)
#Bearing 3- Channel 5
B3c5.ts<- ts(B3Ch5[, 13],frequency=1)
B3c5.bp <-breakpoints((B3c5.ts~1))
B3c5.bp
```

```
# plot data with breakpoint times
plot(B3c5.ts)
lines(fitted(B3c5.bp, breaks = 2), col = 4)
lines(confint(B3c5.bp, breaks = 2))

#Bearing 3- channel 6
B3c6.ts<- ts(B3Ch6[, 13],frequency=1)
B3c6.bp <-breakpoints((B3c6.ts~1))
B3c6.bp


# plot data with breakpoint times
plot(B3c6.ts)
lines(fitted(B3c6.bp, breaks = 1), col = 4)
lines(confint(B3c6.bp, breaks = 1))

#Outer race failure observed in Bearing 4== investigate BPFO (V14)
#Bearing 4- Channel 7
B4c7.ts<- ts(B4Ch7[, 14],frequency=1)
B4c7.bp <-breakpoints((B4c7.ts~1))
B4c7.bp

# plot data with breakpoint times
plot(B4c7.ts)
lines(fitted(B4c7.bp, breaks = 1), col = 4)
lines(confint(B4c7.bp, breaks = 1))

#Bearing 4- channel 8
B4c8.ts<- ts(B4Ch8[, 14],frequency=1)
B4c8.bp <-breakpoints((B4c8.ts~1))
B4c8.bp

# plot data with breakpoint times
plot(B4c8.ts)
lines(fitted(B4c8.bp, breaks = 2), col = 4)
lines(confint(B4c8.bp, breaks = 2))

# Rollar element failure in Bearing 4 == investigate BSF (V15)
```

```
#Bearing 4- Channel 7
B4c7.ts<- ts(B4Ch7[, 15],frequency=1)
B4c7.bp <-breakpoints((B4c7.ts~1))
B4c7.bp

# plot data with breakpoint times
plot(B4c7.ts)
lines(fitted(B4c7.bp, breaks = 1), col = 4)
lines(confint(B4c7.bp, breaks = 1))

#Bearing 4- channel 8
B4c8.ts<- ts(B4Ch8[, 15],frequency=1)
B4c8.bp <-breakpoints((B4c8.ts~1))
B4c8.bp

# plot data with breakpoint times
plot(B4c8.ts)
lines(fitted(B4c8.bp, breaks = 2), col = 4)
lines(confint(B4c8.bp, breaks = 2))

# F-stats
par(mfrow=c(3,2))

# BPFI- Bearing 3
B3c5.Fstats <- Fstats((B3Ch5[, 13]) ~ 1)
plot(B3c5.Fstats)

B3c6.Fstats <- Fstats((B3Ch6[, 13]) ~ 1)
plot(B3c6.Fstats)

# BPFO- Bearing 4
B4c7.Fstats1 <- Fstats((B4Ch7[, 14]) ~ 1)
plot(B4c7.Fstats1)

B4c8.Fstats1 <- Fstats((B4Ch8[, 14]) ~ 1)
plot(B4c8.Fstats1)

#BSF - Bearing 4
```

```
B4c7.Fstats2 <- Fstats((B4Ch7[, 15]) ~ 1)
plot(B4c7.Fstats2)

B4c8.Fstats2 <- Fstats((B4Ch8[, 15]) ~ 1)
plot(B4c8.Fstats2)

# Significant test p-value

sctest(B3c5.Fstats, type = "supF")
sctest(B3c6.Fstats, type = "supF")
sctest(B4c7.Fstats1, type = "supF")
sctest(B4c8.Fstats1, type = "supF")
sctest(B4c7.Fstats2, type = "supF")
sctest(B4c8.Fstats2, type = "supF")

#---------------Section 04 Investigating trend in RMS & Kurtosis---------------
# Change-points are almost similar for both channe on each of the bearings
# so plot only channel 6 for Bearing 3 and channel 8 for Bearing 4
par(mfrow=c(2,1))

plot(B3Ch5$RMS.x, t="l")
plot(B3Ch6$RMS.x, t="l")

plot(B4Ch7$RMS.x, t="l")
plot(B4Ch8$RMS.x, t="l")

plot(B1Ch1$RMS.x, t="l")
plot(B1Ch2$RMS.x, t="l")

plot(B2Ch3$RMS.x, t="l")
plot(B2Ch4$RMS.x, t="l")

# remove all variables from the environment
rm(list=ls())

#-----Section 05 Explore & Compare Test files at the Cpts--------------
#Reload Bearing data
Bearing1dir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 2/Data/"
```

```
files <-list.files()

B1Ch1 <- read.table(file=paste0(Bearing1dir, "../B1Ch1_all.csv"), sep=",", header=FALSE)
B1Ch2 <- read.table(file=paste0(Bearing1dir, "../B1Ch2_all.csv"), sep=",", header=FALSE)
B2Ch3 <- read.table(file=paste0(Bearing1dir, "../B2Ch3_all.csv"), sep=",", header=FALSE)
B2Ch4 <- read.table(file=paste0(Bearing1dir, "../B2Ch4_all.csv"), sep=",", header=FALSE)
B3Ch5 <- read.table(file=paste0(Bearing1dir, "../B3Ch5_all.csv"), sep=",", header=FALSE)
B3Ch6 <- read.table(file=paste0(Bearing1dir, "../B3Ch6_all.csv"), sep=",", header=FALSE)
B4Ch7 <- read.table(file=paste0(Bearing1dir, "../B4Ch7_all.csv"), sep=",", header=FALSE)
B4Ch8 <- read.table(file=paste0(Bearing1dir, "../B4Ch8_all.csv"), sep=",", header=FALSE)
#Inspect table
head(B1Ch1)
str(B1Ch1)

# covert Factor to Numeric & inspect
B1Ch1 <- read.csv(file = '../B1Ch1_all.csv', stringsAsFactors = TRUE)
str(B1Ch1)
B1Ch2 <- read.csv(file = '../B1Ch2_all.csv', stringsAsFactors = TRUE)
B2Ch3 <- read.csv(file = '../B2Ch3_all.csv', stringsAsFactors = TRUE)
B2Ch4 <- read.csv(file = '../B2Ch4_all.csv', stringsAsFactors = TRUE)
B3Ch5 <- read.csv(file = '../B3Ch5_all.csv', stringsAsFactors = TRUE)
B3Ch6 <- read.csv(file = '../B3Ch6_all.csv', stringsAsFactors = TRUE)
B4Ch7 <- read.csv(file = '../B4Ch7_all.csv', stringsAsFactors = TRUE)
B4Ch8 <- read.csv(file = '../B4Ch8_all.csv', stringsAsFactors = TRUE)

#----------Section 06-----------------------
# Combine the data sets

#1: Create data frame with columns of interest
# Displays column 1, 9-15 & 22
dfnew1 <- B1Ch1[,c(12:15)]
str(dfnew1)
dfnew1
dfnew2 <- B1Ch2[,c(12:15)]
dfnew3 <- B2Ch3[,c(12:15)]
dfnew4 <- B2Ch4[,c(12:15)]
dfnew5 <- B3Ch5[,c(12:15)]
dfnew6 <- B3Ch6[,c(12:15)]
```

```
dfnew7 <- B4Ch7[,c(12:15)]
dfnew8 <- B4Ch8[,c(12:15)]

# Re-name column names
colnames(dfnew1) <- c("FTF.B1C1", "BPFI.B1C1", "BPFO.B1C1"," BSF.B1C1")
colnames(dfnew2) <- c("FTF.B1C2", "BPFI.B1C2", "BPFO.B1C2"," BSF.B1C2")
colnames(dfnew3) <- c("FTF.B2C3", "BPFI.B2C3", "BPFO.B2C3"," BSF.B2C3")
colnames(dfnew4) <- c("FTF.B2C4", "BPFI.B2C4", "BPFO.B2C4"," BSF.B2C4")
colnames(dfnew5) <- c("FTF.B3C5", "BPFI.B3C5", "BPFO.B3C5"," BSF.B3C5")
colnames(dfnew6) <- c("FTF.B3C6", "BPFI.B3C6", "BPFO.B3C6"," BSF.B3C6")
colnames(dfnew7) <- c("FTF.B4C7", "BPFI.B4C7", "BPFO.B4C7"," BSF.B4C7")
colnames(dfnew8) <- c("FTF.B4C8", "BPFI.B4C8", "BPFO.B4C8"," BSF.B4C8")

# Merge the  data frames
# Use the cbind function to combine data frames side-by-side:
dfnew <-cbind(dfnew1,dfnew2,dfnew3,dfnew4,dfnew5,dfnew6,dfnew7,dfnew8)
dfnew

#Write csv table for new data frame
write.table(dfnew, file=paste0(Bearing1dir, "../dfnew.csv"), sep=",", row.names=FALSE, col.names=FALSE)

# remove all variables from the environment
rm(list=ls())

#----------Section 07-----------------------
#Read & inspect new data
Bearing1dir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 2/Data/"

dfnew <- read.table(file=paste0(Bearing1dir, "../dfnew.csv"), sep=",", header=FALSE)
str(dfnew)

# Re-name column names & inspect
colnames(dfnew) <- c("FTF.B1C1", "BPFI.B1C1", "BPFO.B1C1"," BSF.B1C1",
                     "FTF.B1C2", "BPFI.B1C2", "BPFO.B1C2"," BSF.B1C2",
                     "FTF.B2C3", "BPFI.B2C3", "BPFO.B2C3"," BSF.B2C3",
                     "FTF.B2C4", "BPFI.B2C4", "BPFO.B2C4"," BSF.B2C4",
                     "FTF.B3C5", "BPFI.B3C5", "BPFO.B3C5"," BSF.B3C5",
                     "FTF.B3C6", "BPFI.B3C6", "BPFO.B3C6"," BSF.B3C6",
```

```
                      "FTF.B4C7", "BPFI.B4C7", "BPFO.B4C7"," BSF.B4C7",
                      "FTF.B4C8", "BPFI.B4C8", "BPFO.B4C8"," BSF.B4C8")
head(dfnew)
str(dfnew)

#----------Section 08----------------------
# Investigate features for interaction effect
library(car)
library("PerformanceAnalytics")
par(mfrow=c(1,1))


# Correlations at lower end of change-points (i.e. 1710)
Changepoint1710 <-dfnew[c(1:1710),]

# Correlations at lower end of change-points (i.e. 1710)
# For BPFO
Changepoint1710BPFO <-dfnew[c(1:1710),c(3,7,11,15,19,23,27,31)]
str(Changepoint1710BPFO)
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint1710BPFO, histogram=TRUE, pch=19)

# For BSF
Changepoint1710BSF <-dfnew[c(1:1710),c(4,8,12,16,20,24,28,32)]
str(Changepoint1710BSF)
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint1710BSF, histogram=TRUE, pch=19)

# For BPFI
Changepoint1710BPFI <-dfnew[c(1:1710),c(2,6,10,14,18,22,26,30)]
str(Changepoint1710BPFI)
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint1710BPFI, histogram=TRUE, pch=19)

#----------Section 08b----------------------
# Correlations at upper end of change-points (i.e. 2119)
# For BPFO
Changepoint2119BPFO <-dfnew[c(1:2119),c(3,7,11,15,19,23,27,31)]
```

```
str(Changepoint2119BPFO)
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint2119BPFO, histogram=TRUE, pch=19)

# For BSF
Changepoint2119BSF <-dfnew[c(1:2119),c(4,8,12,16,20, 24,28,32)]
str(Changepoint2119BSF)
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint2119BSF, histogram=TRUE, pch=19)

# For BPFI
Changepoint2119BPFI <-dfnew[c(1:2119),c(2,6,10,14,18,22,26,30)]
str(Changepoint2119BPFI)
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint2119BPFI, histogram=TRUE, pch=19)

#-------09---------
# Creat a tree model to determine predictors
library(tree)
# Investigate interaction effect @ change-point 1710 (all three defects)
# Creat a subset of all BPFO's and inspect
# For B4C8
Changepoint1710BPFOEve <-dfnew[c(1:1710),c(7,15,23,31)]
str(Changepoint1710BPFOEve)
colnames(Changepoint1710BPFOEve) <- c("BPFO.B1", "BPFO.B2", "BPFO.B3","BPFO.B4")

Tree1710BPFO<-tree(BPFO.B1~.,data=Changepoint1710BPFOEve)
plot(Tree1710BPFO)
text(Tree1710BPFO)

# Creat a subset of all BSF's and inspect
# For B4C8
Changepoint1710BSFEve <-dfnew[c(1:1710),c(8,16,24,32)]
str(Changepoint1710BSFEve)
colnames(Changepoint1710BSFEve) <- c("BSF.B1", "BSF.B2", "BSF.B3","BSF.B4")

# Creat a tree model to determine independent and moderator Bearings
par(mfrow=c(1,2))
```

```
Tmodel1710BSF<-tree(BSF.B4~.,data=Changepoint1710BSFEve)
plot(Tmodel1710BSF)
text(Tmodel1710BSF)

# Creat a subset of all BPFI's and inspect
# For B4C8
Changepoint1710BPFIEve <-dfnew[c(1:1710),c(6,14,22,30)]
str(Changepoint1710BPFIEve)
colnames(Changepoint1710BPFIEve) <- c("BPFI.B1", "BPFI.B2", "BPFI.B3","BPFI.B4")

# Creat a tree model to determine independent and moderator Bearings
Tmodel1710BPFI<-tree(BPFI.B4~.,data=Changepoint1710BPFIEve)
plot(Tmodel1710BPFI)
text(Tmodel1710BPFI)

# Investigate interaction effect on even channels @ change-point 2119 (only BPFI)
# For B4C8
Changepoint2119BPFIEve <-dfnew[c(1:2119),c(6,14,22,30)]
str(Changepoint2119BPFIEve)
colnames(Changepoint2119BPFIEve) <- c("BPFI.B1", "BPFI.B2", "BPFI.B3","BPFI.B4")

# Creat a tree model to determine independent and moderator Bearings
Tmodel2119BPFI<-tree(BPFI.B4~.,data=Changepoint2119BPFIEve)
plot(Tmodel2119BPFI)
text(Tmodel2119BPFI)

#-------------10-------------------------------
# creat Regression model

#-------------10a ------------------------------

# No tree model fit for BPFO therefore fit model with Bearing 3 as predictor base on location
# Model 1: the defect of Bearing 4 (BPFO) predicted by Bearing 3
model.1 <- lm(BPFO.B4 ~ BPFO.B3+BPFO.B2+BPFO.B1, data=Changepoint1710BPFOEve)

# Model 2: Interaction model for Bearing 4 (BPFO)
model.2 <- lm(BPFO.B4 ~ BPFO.B3*BPFO.B2*BPFO.B1, data=Changepoint1710BPFOEve)
```

```
#Show the results
library(stargazer)
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar with ANOVA
sstable1 <- Anova(model.2, type = 2)
sstable1

# Visualize interaction effect and slop aanalysis
# Not performed because no significant interaction effect observed

# Remove Bearing 1 and re-run
model.3 <- lm(BPFO.B4 ~ BPFO.B3+BPFO.B2, data=Changepoint1710BPFOEve)

model.4 <- lm(BPFO.B4 ~ BPFO.B3*BPFO.B2, data=Changepoint1710BPFOEve)

#Show the results
stargazer(model.3,model.4,type="text",
          column.labels = c("model.3","model.4"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.2, model.4)

# performing slopes analysis
#Anova test shows the interaction is insignificant
# therefore no further probe

#-------------10b------------------------------
# For BSF at 1710
```

```
# Model 1: the defect of Bearing 4 (BSF) depend on main effect of B2; B3 & B4
model.1 <- lm(BSF.B4 ~ BSF.B1+BSF.B2+BSF.B3, data=Changepoint1710BSFEve)

# Model 2: Defect of Bearing 4 (BPSF) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BSF.B4 ~ BSF.B1*BSF.B2*BSF.B3,data=Changepoint1710BSFEve)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# Visualize interaction effect
probe_interaction(model.2, pred = BSF.B1, modx = BSF.B2, mod2 = BSF.B3, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BSF.B1, modx = BSF.B2, mod2 = BSF.B3,  jnplot = TRUE)

# Visualize the coefficients
ss2 <- sim_slopes(model.2, pred = BSF.B1, modx = BSF.B2, mod2 = BSF.B3)
plot(ss2)

# Tabular output
as_huxtable(ss2)

#------------10c-----------------------------
# For BPFI at change-point 1710
# Model 1: the defect of Bearing 3 (BPFI) depend on main effect of B2; B3 & B4
model.1 <- lm(BPFI.B3 ~ BPFI.B2+BPFI.B1+BPFI.B4, data=Changepoint1710BPFIEve)

# Model 2: Defect of Bearing 3 (BPFI) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFI.B3 ~ BPFI.B2*BPFI.B1*BPFI.B4,data=Changepoint1710BPFIEve)
```

```
#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# Visualize interaction effect
probe_interaction(model.2, pred = BPFI.B2, modx = BPFI.B1, mod2 = BPFI.B4, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BPFI.B2, modx = BPFI.B1, mod2 = BPFI.B4,  jnplot = TRUE)

# Visualize the coefficients
ss3 <- sim_slopes(model.2, pred = BPFI.B2, modx = BPFI.B1, mod2 = BPFI.B4)
plot(ss3)

# Tabular output
as_huxtable(ss3)

#-------------10d-------------------------------
# For BPFI at change-point 2119
# Model 1: the defect of Bearing 3 (BPFI) depend on main effect of B2; B3 & B4
model.1 <- lm(BPFI.B3 ~ BPFI.B2+BPFI.B1+BPFI.B4, data=Changepoint2119BPFIEve)

# Model 2: Defect of Bearing 3 (BPFI) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFI.B3 ~ BPFI.B2*BPFI.B1*BPFI.B4,data=Changepoint2119BPFIEve)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
```

```
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# Visualize interaction effect
probe_interaction(model.2, pred = BPFI.B2, modx = BPFI.B1, mod2 = BPFI.B4, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BPFI.B2, modx = BPFI.B1, mod2 = BPFI.B4,  jnplot = TRUE)

# Visualize the coefficients
ss4 <- sim_slopes(model.2, pred = BPFI.B2, modx = BPFI.B1, mod2 = BPFI.B4)
plot(ss4)

# Tabular output
as_huxtable(ss4)


#---------11-------------------------
# Investigate feature associations
# For all features of Bearing 4 at 1710
# Correlations at lower end of change-points (i.e. 1710)
Changepoint1710Bearing4 <-dfnew[c(1:1710),c(29:32)]
str(Changepoint1710Bearing4)
colnames(Changepoint1710Bearing4) <- c("FTF.B4", "BPFI.B4", "BPFO.B4","BSF.B4")

# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint1710Bearing4, histogram=TRUE, pch=19)

# For all features of Bearing 3 at 1710
Changepoint1710Bearing3 <-dfnew[c(1:1710),c(21:24)]
str(Changepoint1710Bearing3)
colnames(Changepoint1710Bearing3) <- c("FTF.B3", "BPFI.B3", "BPFO.B3","BSF.B3")
```

```
# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint1710Bearing3, histogram=TRUE, pch=19)

# For all features of Bearing 3 at 2119
Changepoint2119Bearing3 <-dfnew[c(1:2119),c(21:24)]
str(Changepoint2119Bearing3)
colnames(Changepoint2119Bearing3) <- c("FTF.B3", "BPFI.B3", "BPFO.B3","BSF.B3")

# Visualise correlation matrix with performanceanalytics
chart.Correlation(Changepoint2119Bearing3, histogram=TRUE, pch=19)

#----------12---------
# Creat a tree model to determine predictors
library(tree)

# Investigate interaction effect BPFO @ change-point 1710
Tree1710Bearing4<-tree(BPFO.B4~.,data=Changepoint1710Bearing4)
plot(Tree1710Bearing4)
text(Tree1710Bearing4)

# Investigate interaction effect BSF @ change-point 1710 for Bearung 4
Tree1710Bearing4a<-tree(BSF.B4~.,data=Changepoint1710Bearing4)
plot(Tree1710Bearing4a)
text(Tree1710Bearing4a)

# Investigate interaction effect BPFI @ change-point 1710 for Bearung 3
Tree1710Bearing3a<-tree(BPFI.B3~.,data=Changepoint1710Bearing3)
plot(Tree1710Bearing3a)
text(Tree1710Bearing3a)

# Investigate interaction effect BPFI @ change-point 2119 for Bearung 3
Tree1710Bearing3b<-tree(BPFI.B3~.,data=Changepoint2119Bearing3)
plot(Tree1710Bearing3b)
text(Tree1710Bearing3b)
#-------------13-----------------------------
# creat Regression model

#-------------13a ----------------------------
```

```
# Model 1: the defect of Bearing 4 (BPFO)
model.1 <- lm(BPFO.B4 ~ BSF.B4+BPFI.B4+FTF.B4, data=Changepoint1710Bearing4)

# Model 2: Defect of Bearing 4 (BPSF) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFO.B4 ~ BSF.B4*BPFI.B4*FTF.B4, data=Changepoint1710Bearing4)

#Show the results
library(stargazer)
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# Visualize interaction effect
library(jtools)
probe_interaction(model.2, pred = BSF.B4, modx = BPFI.B4, mod2 = FTF.B4, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BSF.B4, modx = BPFI.B4, mod2 = FTF.B4,  jnplot = TRUE)

# Visualize the coefficients
ss5 <- sim_slopes(model.2, pred = BSF.B4, modx = BPFI.B4, mod2 = FTF.B4)
plot(ss5)

# Tabular output
as_huxtable(ss5)

#-------------13b-------------------------------
# Model 1: the defect of Bearing 4 (BSF)
model.1 <- lm(BSF.B4 ~ BPFO.B4+BPFI.B4+FTF.B4, data=Changepoint1710Bearing4)

# Model 2: Defect of Bearing 4 (BPSF) depend on interaction effect of Bearings 2, 3 & 4
```

```
model.2 <- lm(BSF.B4 ~ BPFO.B4*BPFI.B4*FTF.B4, data=Changepoint1710Bearing4)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# Visualize interaction effect
probe_interaction(model.2, pred = BPFO.B4, modx = BPFI.B4, mod2 = FTF.B4, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BPFO.B4, modx = BPFI.B4, mod2 = FTF.B4,  jnplot = TRUE)

# Visualize the coefficients
ss6 <- sim_slopes(model.2, pred = BPFO.B4, modx = BPFI.B4, mod2 = FTF.B4)
plot(ss6)

# Tabular output
as_huxtable(ss6)

#-------------13c-----------------------------
# For BPFI at change-point 1710
# Model 1: the defect of Bearing 3 (BPFI)
model.1 <- lm(BPFI.B3 ~ BPFO.B3+BSF.B3+FTF.B3, data=Changepoint1710Bearing3)

# Model 2: Defect of Bearing 3 (BPSF) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFI.B3 ~ BPFO.B3*BSF.B3*FTF.B3, data=Changepoint1710Bearing3)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
```

```
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# compar the models
anova(model.1, model.2)

# Visualize interaction effect
probe_interaction(model.2, pred = BPFO.B3, modx = BSF.B3, mod2 = FTF.B3, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BPFO.B3, modx = BSF.B3, mod2 = FTF.B3,  jnplot = TRUE)

# Visualize the coefficients
ss7 <- sim_slopes(model.2, pred = BPFO.B3, modx = BSF.B3, mod2 = FTF.B3)
plot(ss7)

# Tabular output
as_huxtable(ss7)

#-------------13d-------------------------------
# For BPFI at change-point 2119

# Model 1: the defect of Bearing 3 (BPFI)
model.1 <- lm(BPFI.B3 ~ BPFO.B3+BSF.B3+FTF.B3, data=Changepoint2119Bearing3)

# Model 2: Defect of Bearing 3 (BPSF) depend on interaction effect of Bearings 2, 3 & 4
model.2 <- lm(BPFI.B3 ~ BPFO.B3*BSF.B3*FTF.B3, data=Changepoint2119Bearing3)

#Show the results
stargazer(model.1,model.2,type="text",
          column.labels = c("model.1","model.2"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)
```

```
# compar the models
anova(model.1, model.2)

# Visualize interaction effect
probe_interaction(model.2, pred = BPFO.B3, modx = BSF.B3, mod2 = FTF.B3, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BPFO.B3, modx = BSF.B3, mod2 = FTF.B3,  jnplot = TRUE)

# Visualize the coefficients
ss8 <- sim_slopes(model.2, pred = BPFO.B3, modx = BSF.B3, mod2 = FTF.B3)
plot(ss8)

# Tabular output
as_huxtable(ss8)

#-----Section 14-----------------------------------------
# remove all variables from the environment
rm(list=ls())
```

Final R- Code Case Study Dataset 3

```
#-----Section 01-----------------------------------------
# get the data
# set working directory

setwd("C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 3/")
getwd()
files <-list.files()
#-------------Section 02----------
# Import first data file

BearingNdir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 3/"
data <- read.csv(paste0(BearingNdir, "acc_00001.csv"), header=FALSE)

head(data)

#Descriptive statistics
library(pastecs)
stat.desc(data)

# Re-name column names
colnames(data) <- c("Hour", "Minute", "Second", "u_second","Horiz","Vert")
str(data)
head(data)

library(e1071)

# Helper functions
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
```

```
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.137795 # ball diameter (3.5mm), in inches
Pd <- 1.008 # bearing mean diameter = 25.6mm (pitch diameter), in inches
Nb <- 13 # number of rolling elements
a <- 0 # contact angle, in radians Ref: Wang 2015
s <- 2000/60 # rotational frequency, in Hz (challenge info)
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
```

```
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Set up storage for bearing-grouped data
b1m <- matrix(nrow=0, ncol=(1*23))
b2m <- matrix(nrow=0, ncol=(1*23))
b3m <- matrix(nrow=0, ncol=(1*23))
b4m <- matrix(nrow=0, ncol=(1*23))
b5m <- matrix(nrow=0, ncol=(1*23))
b6m <- matrix(nrow=0, ncol=(1*23))
library(lubridate)
strftime(as.POSIXlt(data_tmp2$my_date, format = "%Y-%d-%m"), format="%W")
# and for timestamps
timestamp <- vector()

for (filename in list.files(BearingNdir))
{
  cat("Processing file ", filename, "\n")

  data <- read.csv(paste0(BearingNdir, filename), header=FALSE)
  colnames(data) <- c("Hour", "Minute", "Second", "usecond","Horiz","Vert")

  # Bind the new rows to the bearing matrices
  b1m <- rbind(b1m, c(all.features(data$Hour)))
  b2m <- rbind(b2m, c(all.features(data$Minute)))
  b3m <- rbind(b3m, c(all.features(data$Second)))
```

```
  b4m <- rbind(b4m, c(all.features(data$usecond)))
  b5m <- rbind(b5m, c(all.features(data$Horiz)))
  b6m <- rbind(b6m, c(all.features(data$Vert)))

}

cnames <- c("Min.x", "Qu.1.x", "Median.x", "Qu.3.x", "Max.x", "Mean.x", "SD.x", "Skew.x", "Kurt.x", "RMS.x",
"FTF.x", "BPFI.x", "BPFO.x", "BSF.x", "F1.x", "F2.x", "F3.x", "F4.x", "F5.x", "VHF.pow.x", "HF.pow.x",
"MF.pow.x", "LF.pow.x")
colnames(b1m) <- cnames
colnames(b2m) <- cnames
colnames(b3m) <- cnames
colnames(b4m) <- cnames
colnames(b5m) <- cnames
colnames(b6m) <- cnames

Hour <- data.frame(b1m)
Minute <- data.frame(b2m)
Second <- data.frame(b3m)
usecond <- data.frame(b4m)
Horiz <- data.frame(b5m)
Vert <- data.frame(b6m)

write.table(Hour, file=paste0(BearingNdir, "../Hour.csv"), sep=",", row.names=FALSE)
write.table(Minute, file=paste0(BearingNdir, "../Minute.csv"), sep=",", row.names=FALSE)
write.table(Second, file=paste0(BearingNdir, "../Second.csv"), sep=",", row.names=FALSE)
write.table(usecond, file=paste0(BearingNdir, "../usecond.csv"), sep=",", row.names=FALSE)
write.table(Horiz, file=paste0(BearingNdir, "../Horiz.csv"), sep=",", row.names=FALSE)
write.table(Vert, file=paste0(BearingNdir, "../Vert.csv"), sep=",", row.names=FALSE)

# remove all variables from the environment
rm(list=ls())

#-----Section 03 Changepoint Analysis--------------------
#Reload Bearing data

BearingNdir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 3/"
files <-list.files()
```

```
Horiz <- read.table(file=paste0(BearingNdir, "../Horiz.csv"), sep=",", header=FALSE)
Vert <- read.table(file=paste0(BearingNdir, "../Vert.csv"), sep=",", header=FALSE)

#Inspect table
head(Horiz)
cnames <- c("Min", "Qu.1", "Median", "Qu.3", "Max", "Mean", "SD", "Skew", "Kurt",
            "RMS", "FTF", "BPFI", "BPFO", "BSF", "F1", "F2", "F3", "F4", "F5", "VHF.pow",
            "HF.pow", "MF.pow", "LF.pow")
str(Horiz)
# covert Factor to Numeric
Horiz <- read.csv(file = '../Horiz.csv', stringsAsFactors = TRUE)
str(Horiz)
Vert <- read.csv(file = '../Vert.csv', stringsAsFactors = TRUE)
cnames <- c("Min", "Qu.1", "Median", "Qu.3", "Max", "Mean", "SD", "Skew", "Kurt",
            "RMS", "FTF", "BPFI", "BPFO", "BSF", "F1", "F2", "F3", "F4", "F5", "VHF.pow",
            "HF.pow", "MF.pow", "LF.pow")

#Using changepoint using statistical Pruned Exact Linear Time (PELT)
# Failure type unknown so we investigate all three failures
# Horizontal channel BPFO
# so plot BPFO
Horiz <-Horiz[,c(11:14)]
head(Horiz)
str(Horiz)
is.na(Horiz)
library(changepoint)
par(mfrow=c(2,1))

mvalue1 = cpt.mean(Horiz$BPFO.x, method="PELT") #mean changepoints using PELT
cpts(mvalue1)
vvalue1 = cpt.var(diff(Horiz$BPFO.x), method="PELT")
cpts(vvalue1)

Horiz.pelt <- cpt.var(diff(diff(Horiz$BPFO.x), method = "PELT"))
plot(Horiz.pelt, xlab = "Index")
logLik(Horiz.pelt)
```

```
# Vertical channel BPFO
vvalue2 = cpt.var(diff(Vert$BPFO.x), method="PELT")
cpts(vvalue2)

Vert.pelt <- cpt.var(diff(diff(Vert$BPFO.x), method = "PELT"))
plot(Vert.pelt, xlab = "Index")
logLik(Vert.pelt)

#Structure Change
library(strucchange)
par(mfrow=c(2,2))
#Horiz channel
Horiz.ts<- ts(Horiz$BPFO.x,frequency=1)
Horiz.bp <-breakpoints((Horiz.ts~1))
Horiz.bp
summary(Horiz.bp)
plot(Horiz.bp)
memory.size()
memory.limit(5000)
# plot data with breakpoint times
plot(Horiz.ts)
lines(fitted(Horiz.bp, breaks = 0), col = 4)
lines(confint(Horiz.bp, breaks = 0))

# Vert channel
Vert.ts<- ts(Vert$BPFO.x,frequency=1)
Vert.bp <-breakpoints((Vert.ts~1))
Vert.bp
summary(Vert.bp)
plot(Vert.bp)

# plot data with breakpoint times
plot(Vert.ts)
lines(fitted(Vert.bp, breaks = 2), col = 4)
lines(confint(Vert.bp, breaks = 2))
```

```
#F-stats
par(mfrow=c(1,2))
Horiz.Fstats <- Fstats((Horiz$BPFO.x) ~ 1)
plot(Horiz.Fstats)

Vert.Fstats <- Fstats((Vert$BPFO.x) ~ 1)
plot(Vert.Fstats)

# Significant test p-value
sctest(Horiz.Fstats, type = "supF")
sctest(Vert.Fstats, type = "supF")

#-----Section 03b --------------------
#  for  BSF
library(changepoint)
par(mfrow=c(2,1))

vvalue1 = cpt.var(diff(Horiz$BSF.x), method="PELT")
cpts(vvalue1)

Horiz.pelt <- cpt.var(diff(diff(Horiz$BSF.x), method = "PELT"))
plot(Horiz.pelt, xlab = "Index")
logLik(Horiz.pelt)

# Vertical channel
vvalue2 = cpt.var(diff(Vert$BSF.x), method="PELT")
cpts(vvalue2)

Vert.pelt <- cpt.var(diff(diff(Vert$BSF.x), method = "PELT"))
plot(Vert.pelt, xlab = "Index")
logLik(Vert.pelt)

#Structure Change

par(mfrow=c(2,2))
#Horiz channel
Horiz.ts<- ts(Horiz$BSF.x,frequency=1)
Horiz.bp <-breakpoints((Horiz.ts~1))
```

```
Horiz.bp
summary(Horiz.bp)
plot(Horiz.bp)

# plot data with breakpoint times
plot(Horiz.ts)
lines(fitted(Horiz.bp, breaks = 2), col = 4)
lines(confint(Horiz.bp, breaks = 2))

# Vert channel
Vert.ts<- ts(Vert$BSF.x,frequency=1)
Vert.bp <-breakpoints((Vert.ts~1))
Vert.bp
summary(Vert.bp)
plot(Vert.bp)

# plot data with breakpoint times
plot(Vert.ts)
lines(fitted(Vert.bp, breaks = 2), col = 4)
lines(confint(Vert.bp, breaks = 2))


#F-stats
par(mfrow=c(1,2))
Horiz.Fstats <- Fstats((Horiz$BSF.x) ~ 1)
plot(Horiz.Fstats)

Vert.Fstats <- Fstats((Vert$BSF.x) ~ 1)
plot(Vert.Fstats)

# Significant test p-value
sctest(Horiz.Fstats, type = "supF")
sctest(Vert.Fstats, type = "supF")

#-----Section 03c --------------------
#  for  BPFI
par(mfrow=c(2,1))
```

```
vvalue1 = cpt.var(diff(Horiz$BPFI.x), method="PELT")
cpts(vvalue1)

Horiz.pelt <- cpt.var(diff(diff(Horiz$BPFI.x), method = "PELT"))
plot(Horiz.pelt, xlab = "Index")
logLik(Horiz.pelt)

# Vertical channel
vvalue2 = cpt.var(diff(Vert$BPFI.x), method="PELT")
cpts(vvalue2)

Vert.pelt <- cpt.var(diff(diff(Vert$BPFI.x), method = "PELT"))
plot(Vert.pelt, xlab = "Index")
logLik(Vert.pelt)

#Structure Change

par(mfrow=c(2,2))
#Horiz channel
Horiz.ts<- ts(Horiz$BPFI.x,frequency=1)
Horiz.bp <-breakpoints((Horiz.ts~1))
Horiz.bp
summary(Horiz.bp)
plot(Horiz.bp)

# plot data with breakpoint times
plot(Horiz.ts)
lines(fitted(Horiz.bp, breaks = 2), col = 4)
lines(confint(Horiz.bp, breaks = 2))

# Vert channel
Vert.ts<- ts(Vert$BPFI.x,frequency=1)
Vert.bp <-breakpoints((Vert.ts~1))
Vert.bp
summary(Vert.bp)
plot(Vert.bp)

# plot data with breakpoint times
```

```
plot(Vert.ts)
lines(fitted(Vert.bp, breaks = 2), col = 4)
lines(confint(Vert.bp, breaks = 2))


#F-stats
par(mfrow=c(1,2))
Horiz.Fstats <- Fstats((Horiz$BPFI.x) ~ 1)
plot(Horiz.Fstats)

Vert.Fstats <- Fstats((Vert$BPFI.x) ~ 1)
plot(Vert.Fstats)

# Significant test p-value
sctest(Horiz.Fstats, type = "supF")
sctest(Vert.Fstats, type = "supF")

#-----Section 03d --------------------
#  for  FTF
par(mfrow=c(2,1))

vvalue1 = cpt.var(diff(Horiz$FTF.x), method="PELT")
cpts(vvalue1)

Horiz.pelt <- cpt.var(diff(diff(Horiz$FTF.x), method = "PELT"))
plot(Horiz.pelt, xlab = "Index")
logLik(Horiz.pelt)

# Vertical channel
vvalue2 = cpt.var(diff(Vert$FTF.x), method="PELT")
cpts(vvalue2)

Vert.pelt <- cpt.var(diff(diff(Vert$FTF.x), method = "PELT"))
plot(Vert.pelt, xlab = "Index")
logLik(Vert.pelt)

#Structure Change
```

```
par(mfrow=c(2,2))
#Horiz channel
Horiz.ts<- ts(Horiz$FTF.x,frequency=1)
Horiz.bp <-breakpoints((Horiz.ts~1))
Horiz.bp
summary(Horiz.bp)
plot(Horiz.bp)

# plot data with breakpoint times
plot(Horiz.ts)
lines(fitted(Horiz.bp, breaks = 2), col = 4)
lines(confint(Horiz.bp, breaks = 2))

# Vert channel
Vert.ts<- ts(Vert$FTF.x,frequency=1)
Vert.bp <-breakpoints((Vert.ts~1))
Vert.bp
summary(Vert.bp)
plot(Vert.bp)

# plot data with breakpoint times
plot(Vert.ts)
lines(fitted(Vert.bp, breaks = 2), col = 4)
lines(confint(Vert.bp, breaks = 2))


#F-stats
par(mfrow=c(1,2))
Horiz.Fstats <- Fstats((Horiz$FTF.x) ~ 1)
plot(Horiz.Fstats)

Vert.Fstats <- Fstats((Vert$FTF.x) ~ 1)
plot(Vert.Fstats)

# Significant test p-value
sctest(Horiz.Fstats, type = "supF")
sctest(Vert.Fstats, type = "supF")
```

```
#----------Section 04----------------------
# Combine the data sets
#1: Create data frame with columns of interest
# Displays column 8-15 & 23
dfnew1 <- Horiz[,c(8:15,23)]
dfnew1
dfnew2 <- Vert[,c(8:15,23)]

# Re-name column names
colnames(dfnew1) <- c("Skew.H", "Kurt.H", "RMS.H", "FTF.H", "BPFI.H", "BPFO.H"," BSF.H","HF.H","LF.H")
colnames(dfnew2) <- c("Skew.V", "Kurt.V", "RMS.V", "FTF.V", "BPFI.V", "BPFO.V"," BSF.V","HF.V","LF.V")

# Merge the  data frames
# Use the cbind function to combine data frames side-by-side:
dfnew <-cbind(dfnew1,dfnew2)
dfnew

#Write csv table for new data frame
write.table(dfnew, file=paste0(BearingNdir, "../dfnew.csv"), sep=",", row.names=FALSE, col.names=FALSE)

# remove all variables from the environment
rm(list=ls())


#----------Section 05b----------------------
# Explore & Compare Test files at the Cpts

# Import File for Horizontal channels BPFO cpt (PELT)= 1495
# Get timestamp of file representing row 1495
data1 <- read.table(paste0(BearingNdir,"acc_01495.csv"), sep=",", header=FALSE)
head(data1)

# Re-name column names
colnames(data1) <- c("Hour", "Minute", "Second", "u_second","Horiz","Vert")
str(data1)
head(data1)

# Calculate the four frequencies for the
```

```
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.137795 # ball diameter (3.5mm), in inches
Pd <- 1.008 # bearing mean diameter = 25.6mm (pitch diameter), in inches
Nb <- 13 # number of rolling elements
a <- 0 # contact angle, in radians Ref: Wang 2015
s <- 2000/60 # rotational frequency, in Hz (challenge info)
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
```

```
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Horiz)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Horiz)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])
```

```
# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(2,1))
# For Horiz
Horizfreq <- seq(0, 310, length.out=length(data1$Horiz)/2)
plot(d1B1.fft.amps[1:(length(data1$Horiz)/2)] ~ Horizfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Vertical
# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Vert)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Vert)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 310Hz

# For Vert
Vertfreq <- seq(0, 310, length.out=length(data1$Vert)/2)
plot(d1B1.fft.amps[1:(length(data1$Vert)/2)] ~ Vertfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
```

```
abline(v=ftf,col="violet",lty=3)

#----------Section 05c-----------------------
# Explore & Compare Test files at the Cpts

# Import File for Horizontal channels BPFO cpt (PELT)= 1655
# Get timestamp of file representing row 1655
data1 <- read.table(paste0(BearingNdir,"acc_01655.csv"), sep=",", header=FALSE)
head(data1)

# Re-name column names
colnames(data1) <- c("Hour", "Minute", "Second", "u_second","Horiz","Vert")
str(data1)
head(data1)

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.137795 # ball diameter (3.5mm), in inches
Pd <- 1.008 # bearing mean diameter = 25.6mm (pitch diameter), in inches
Nb <- 13 # number of rolling elements
a <- 0 # contact angle, in radians Ref: Wang 2015
s <- 2000/60 # rotational frequency, in Hz (challenge info)
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
```

```
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
```

```
  return (floor(freq/step))
}

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Horiz)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Horiz)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(2,1))
# For Horiz
Horizfreq <- seq(0, 310, length.out=length(data1$Horiz)/2)
plot(d1B1.fft.amps[1:(length(data1$Horiz)/2)] ~ Horizfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Vertical
# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Vert)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features
```

```
# 1. Format the full dataset
d1B1.fft <- fft(data1$Vert)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 310Hz

# For Vert
Vertfreq <- seq(0, 310, length.out=length(data1$Vert)/2)
plot(d1B1.fft.amps[1:(length(data1$Vert)/2)] ~ Vertfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#----------Section 05d-----------------------
# Explore & Compare Test files at the Cpts

# Import File for Horizontal channels BPFO cpt (PELT)= 1656
# Get timestamp of file representing row 1656
data1 <- read.table(paste0(BearingNdir,"acc_01656.csv"), sep=",", header=FALSE)
head(data1)

# Re-name column names
colnames(data1) <- c("Hour", "Minute", "Second", "u_second","Horiz","Vert")
str(data1)
head(data1)

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.137795 # ball diameter (3.5mm), in inches
Pd <- 1.008 # bearing mean diameter = 25.6mm (pitch diameter), in inches
Nb <- 13 # number of rolling elements
a <- 0 # contact angle, in radians Ref: Wang 2015
```

```
s <- 2000/60 # rotational frequency, in Hz (challenge info)
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)
```

```
  return(features)
}

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Horiz)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Horiz)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(2,1))
# For Horiz
```

```
Horizfreq <- seq(0, 310, length.out=length(data1$Horiz)/2)
plot(d1B1.fft.amps[1:(length(data1$Horiz)/2)] ~ Horizfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Vertical
# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Vert)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Vert)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 310Hz

# For Vert
Vertfreq <- seq(0, 310, length.out=length(data1$Vert)/2)
plot(d1B1.fft.amps[1:(length(data1$Vert)/2)] ~ Vertfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#----------Section 05e-----------------------
# Explore & Compare Test files at the Cpts

# Import File for Horizontal channels BPFO cpt (PELT)= 1662
```

```
# Get timestamp of file representing row 1662
data1 <- read.table(paste0(BearingNdir,"acc_01662.csv"), sep=",", header=FALSE)
head(data1)

# Re-name column names
colnames(data1) <- c("Hour", "Minute", "Second", "u_second","Horiz","Vert")
str(data1)
head(data1)

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.137795 # ball diameter (3.5mm), in inches
Pd <- 1.008 # bearing mean diameter = 25.6mm (pitch diameter), in inches
Nb <- 13 # number of rolling elements
a <- 0 # contact angle, in radians Ref: Wang 2015
s <- 2000/60 # rotational frequency, in Hz (challenge info)
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])
```

```
  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Horiz)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
```

```
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Horiz)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(2,1))
# For Horiz
Horizfreq <- seq(0, 310, length.out=length(data1$Horiz)/2)
plot(d1B1.fft.amps[1:(length(data1$Horiz)/2)] ~ Horizfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Vertical
# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Vert)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Vert)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
```

```
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 310Hz

# For Vert
Vertfreq <- seq(0, 310, length.out=length(data1$Vert)/2)
plot(d1B1.fft.amps[1:(length(data1$Vert)/2)] ~ Vertfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#----------Section 05f----------------------
# Explore & Compare Test files at the Cpts

# Import File for Horizontal channels BPFO cpt (PELT)= 1945
# Get timestamp of file representing row 1945
data1 <- read.table(paste0(BearingNdir,"acc_01945.csv"), sep=",", header=FALSE)
head(data1)

# Re-name column names
colnames(data1) <- c("Hour", "Minute", "Second", "u_second","Horiz","Vert")
str(data1)
head(data1)

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.137795 # ball diameter (3.5mm), in inches
Pd <- 1.008 # bearing mean diameter = 25.6mm (pitch diameter), in inches
Nb <- 13 # number of rolling elements
a <- 0 # contact angle, in radians Ref: Wang 2015
s <- 2000/60 # rotational frequency, in Hz (challenge info)
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)
```

```
all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
  features <- append(features, powers)

  return(features)
}

# Generate the first four features of the bearings
fft.spectrum <- function (d)
```

```
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Horiz)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Horiz)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(2,1))
# For Horiz
Horizfreq <- seq(0, 310, length.out=length(data1$Horiz)/2)
plot(d1B1.fft.amps[1:(length(data1$Horiz)/2)] ~ Horizfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)
```

```
# For Vertical
# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$Vert)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
                d1B1.fft.amps[freq2index(bpfi)],
                d1B1.fft.amps[freq2index(bpfo)],
                d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$Vert)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# For clarity, zoom in to frequencies up to 310Hz

# For Vert
Vertfreq <- seq(0, 310, length.out=length(data1$Vert)/2)
plot(d1B1.fft.amps[1:(length(data1$Vert)/2)] ~ Vertfreq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#---------------Section 06 Investigating trend in RMS Kurtosis---------------
dfnew
head(dfnew)
par(mfrow=c(1,2))
# Horizontal plots
plot(dfnew$RMS.H, t="l")
plot(dfnew$Kurt.H, t="l")

# Vertical plots
plot(dfnew$RMS.V, t="l")
```

```
plot(dfnew$Kurt.V, t="l")
getwd()

#-----Section 07-----------------------------------------
setwd("C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 3/")
BearingNdir <- "C:/Users/gjordan/Desktop/SysRevWD/Projects/Case Study Dataset 3/Data/"
dfnew <- read.table(file=paste0(BearingNdir, "../dfnew.csv"), sep=",", header=FALSE)
head(dfnew)

colnames(dfnew) <- c("Skew.H", "Kurt.H", "RMS.H", "FTF.H", "BPFI.H", "BPFO.H"," BSF.H","HF.H","LF.H",
                     "Skew.V", "Kurt.V", "RMS.V", "FTF.V", "BPFI.V", "BPFO.V"," BSF.V","HF.V","LF.V")

# Investigate relationship at change-points
library(car)

# Any interaction b/n all variables Vertical Channel
# @ change-point 1945
Bearing1945 <-dfnew[c(1:1945),c(14:16)]
str(Bearing1945)

# Visualise correlation matrix with performanceanalytics
library("PerformanceAnalytics")
chart.Correlation(Bearing1945, histogram=TRUE, pch=19)

# @ change-point 1656

Bearing1656 <-dfnew[c(1:1656),c(14:16)]
str(Bearing1656)

# Visualise correlation matrix with performanceanalytics
chart.Correlation(Bearing1656, histogram=TRUE, pch=19)

#----------Section 08-----------------------
# Investigate interaction effect on even channels @ change-point 1662
library(car)
# subset for vertical channelfor BPFO
Vert1945 <-dfnew[c(1:1945),c(14:16)]
str(Vert1945)
```

```
colnames(Vert1945) <- c("BPFI", "BPFO","BSF")

# Fit a tree model for BPFO, BSF and BPFI at the change-point
# Creat a tree model to determine independent and moderator Bearings for BPFO
library(tree)
Tmodel1<-tree(BPFO~.,data=Vert1945)
plot(Tmodel1)
text(Tmodel1)

# Creat a tree model to determine independent and moderator Bearings for BSF
Tmodel2<-tree(BSF~.,data=Vert1945)
plot(Tmodel2)
text(Tmodel2)

# Creat a tree model to determine independent and moderator Bearings for BPFI
Tmodel3<-tree(BPFI~.,data=Vert1945)
plot(Tmodel3)
text(Tmodel3)

# Creat a subset for change-point 1656
Vert1656 <-dfnew[c(1:1656),c(14:16)]
colnames(Vert1656) <- c("BPFI", "BPFO","BSF")
str(Vert1656)
head(Vert1656)

# Creat a tree model to determine independent and moderator Bearings for BPFO
Tmodel4<-tree(BPFO~.,data=Vert1656)
plot(Tmodel4)
text(Tmodel4)

# Creat a tree model to determine independent and moderator Bearings for BSF
Tmodel5<-tree(BSF~.,data=Vert1656)
plot(Tmodel5)
text(Tmodel5)

# Creat a tree model to determine independent and moderator Bearings for BPFI
Tmodel6<-tree(BPFI~.,data=Vert1656)
plot(Tmodel6)
```

```
text(Tmodel6)

#-----------09----------------
# Association between features

#-------------09a Change-point 1656-----------------------------
# For BPFO of Bearing 4

model.1 <- lm(BPFO ~ BSF + BPFI,data=Vert1656)

model.2 <- lm(BPFO ~ BSF*BPFI,data=Vert1656)

# For BSF of Bearing 4
model.3 <- lm(BSF ~ BPFO + BPFI,data=Vert1656)

model.4 <- lm(BSF ~ BPFO*BPFI,data=Vert1656)

# For BPFI of Bearing 3
model.5 <- lm(BPFI ~ BPFO + BSF,data=Vert1656)

# Model 4: Defect of BPFO association BSF as predictor
model.6 <- lm(BPFI ~ BPFO*BSF,data=Vert1656)

#Show the results
library(stargazer)
stargazer(model.1,model.2,model.3,model.4,model.5,model.6,type="text",
          column.labels = c("model.1","model.2","model.3","model.4","model.5","model.6"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# Significant effect but dependant variables differs so no further investigation with ANOVA
anova(model.1, model.2)
anova(model.3, model.4)
anova(model.5, model.6)

#-------------09b Visualise Effect and Slope Analysis-----------------------------
```

```
library(jtools)
# For BPFO
probe_interaction(model.2, pred = BSF, modx = BPFI, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BSF, modx = BPFI, jnplot = TRUE)

# For BSF
probe_interaction(model.4, pred = BPFO, modx = BPFI, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.4, pred = BPFO, modx = BPFI, jnplot = TRUE)

# For BPFI
probe_interaction(model.6, pred = BPFO, modx = BSF, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.6, pred = BPFO, modx = BSF, jnplot = TRUE)

#-------------09c Change-point 1945-------------------------------
# For BPFO of Bearing 4

model.1 <- lm(BPFO ~ BSF + BPFI,data=Vert1945)

model.2 <- lm(BPFO ~ BSF*BPFI,data=Vert1945)

# For BSF of Bearing 4
model.3 <- lm(BSF ~ BPFO + BPFI,data=Vert1945)

model.4 <- lm(BSF ~ BPFO*BPFI,data=Vert1945)

# For BPFI of Bearing 3
model.5 <- lm(BPFI ~ BPFO + BSF,data=Vert1945)

model.6 <- lm(BPFI ~ BPFO*BSF,data=Vert1945)
```

```
#Show the results
library(stargazer)
stargazer(model.1,model.2,model.3,model.4,model.5,model.6,type="text",
          column.labels = c("model.1","model.2","model.3","model.4","model.5","model.6"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)

# Significant effect so further investigation with ANOVA
anova(model.1, model.2)
anova(model.3, model.4)
anova(model.5, model.6)

#-------------09d Visualise Effect and Slope Analysis-----------------------------
library(jtools)
# For BPFO
probe_interaction(model.2, pred = BSF, modx = BPFI, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.2, pred = BSF, modx = BPFI, jnplot = TRUE)

# For BSF
probe_interaction(model.4, pred = BPFO, modx = BPFI, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.4, pred = BPFO, modx = BPFI, jnplot = TRUE)

# For BPFI
probe_interaction(model.6, pred = BPFO, modx = BSF, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# performing slopes analysis
sim_slopes(model.6, pred = BPFO, modx = BSF, jnplot = TRUE)
```

```
#-----Section 10-----------------------------------------
# remove all variables from the environment
rm(list=ls())
```

Final R- Code Case Study Dataset 4

```
#-----Section 01-----------------------------------------
# get the data
# set working directory

setwd("C:/Users/George/Desktop/Main research/WDResearch 2/Projects/Training Dataset/test2/")
getwd()
files <-list.files()
#-------------Section 02----------
# Import first data file

Trainingdir <- "C:/Users/George/Desktop/Main research/WDResearch 2/Projects/Training Dataset/test2/"
data <- read.table(paste0(Trainingdir,"2004.02.18.15.22.39"), header=FALSE, sep="\t")

# Re-name column names
colnames(data) <- c("Bearing1","Bearing2","Bearing3","Bearing4")
# explore the data
str(data)    #see the structure of the data
head(data)       # The top of the data
class(data)
sapply(data,class) #print out the class of variables

# Inspect up to twenty randomly selected data files using descriptive statistics
library(pastecs)
stat.desc(data)

# Inspect the spread using inter quitile range
IQR(data$Bearing1)
IQR(data$Bearing2)
IQR(data$Bearing3)
IQR(data$Bearing4)

#skewness  to measure asymmetry
library(moments)
skewness(data)

#kurtosis to measure peakedness compare with a gausian distribution
kurtosis(data)
```

```
# visualise bivariate relationships among transformed: scatterplot matrix
pairs(data)

# more informative scatterplot matrix
library(psych)
pairs.panels(data)

# Correlations/covariances among numeric variables in
# Correlations with significance levels
library(Hmisc)

# type can be pearson or spearman
rcorr(as.matrix(data))

# Visual inspection using histogram for checking normality
library(MVN)
par(mfrow=c(2,2))
hist(data, type = "histogram") # creates univariate histograms
mtext("Histogram-Plot: Training Dataset", line = 0.5, outer = TRUE)

# Visual inspection using boxplot for checking normality
par(mfrow=c(1,1))
boxplot(data, horizontal=TRUE)

#qqnorm() plots for linearity
par(mfrow=c(1,4))
qqnorm(data$Bearing1) # creates univariate qqplot
qqline(data$Bearing1, col = "red", lwd = 2)
qqnorm(data$Bearing2)
qqline(data$Bearing2, col = "red", lwd = 2)
qqnorm(data$Bearing3)
qqline(data$Bearing3, col = "red", lwd = 2)
qqnorm(data$Bearing4)
qqline(data$Bearing4, col = "red", lwd = 2)

## Generate sequence plot of the data.
a = data$Bearing1
```

```
par(mfrow = c(1, 4),
    oma = c(0, 0, 2, 0),
    mar = c(5.1, 4.1, 2.1, 2.1))
plot(a,ylab="A",xlab="Bearing1")
b = data$Bearing2
plot(b,ylab="B",xlab="Bearing2")
c = data$Bearing3
plot(c,ylab="C",xlab="Bearing3")
d = data$Bearing4
plot(d,ylab="D",xlab="Bearing4")

#generate lag plot
plot(a,lag(a),xlab="Bearing 1[i-1]",ylab="Bearing 1[i]")
plot(b,lag(b),xlab="Bearing 2[i-1]",ylab="Bearing 2[i]")
plot(c,lag(c),xlab="Bearing 3[i-1]",ylab="Bearing 3[i]")
plot(d,lag(d),xlab="Bearing 4[i-1]",ylab="Bearing 4[i]")
mtext("Lag-Plot: Training Dataset", line = 0.5, outer = TRUE)

#--------------Section 04 Calculating and Ploting Key Frequencies using Modified Carsons code ----------
# Re-name column names
colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)
# Reffor info on structure and size: Qiu, H., Lee, J., Lin, J. & Yu, G. (2006)
Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

# Generate the first four features of the bearings
fft.spectrum <- function (d)
```

```
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

B1.fft.amps <- fft.spectrum(data$B1Ch1)
features <- c(B1.fft.amps[freq2index(ftf)],
              B1.fft.amps[freq2index(bpfi)],
              B1.fft.amps[freq2index(bpfo)],
              B1.fft.amps[freq2index(bsf)])
features
# calculate Key frequencies

# Strongest frequencies
n <- 5
frequencies <- seq(0, 10000, length.out=length(B1.fft.amps))
sorted <- sort.int(B1.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
vhf <- freq2index(6000):length(B1.fft.amps)    # 6kHz plus
hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
lf <- 0:(freq2index(1250)-1)                # forcing frequency band

powers <- c(sum(B1.fft.amps[vhf]), sum(B1.fft.amps[hf]), sum(B1.fft.amps[mf]), sum(B1.fft.amps[lf]))
features <- append(features, powers)
features
```

```
# For Bearing 2
B2.fft.amps <- fft.spectrum(data$B2Ch2)
B2.features <- c(B2.fft.amps[freq2index(ftf)],
                 B2.fft.amps[freq2index(bpfi)],
                 B2.fft.amps[freq2index(bpfo)],
                 B2.fft.amps[freq2index(bsf)])
B2.features
# calculate Key frequencies

# Strongest frequencies
n <- 5
B2.frequencies <- seq(0, 10000, length.out=length(B2.fft.amps))
sorted <- sort.int(B2.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
B2.features <- append(B2.features, B2.frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
B2.vhf <- freq2index(6000):length(B2.fft.amps)    # 6kHz plus
B2.hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
B2.mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
B2.lf <- 0:(freq2index(1250)-1)                # forcing frequency band

B2.powers <- c(sum(B2.fft.amps[vhf]), sum(B2.fft.amps[hf]), sum(B2.fft.amps[mf]), sum(B2.fft.amps[lf]))
B2.features <- append(B2.features, B2.powers)
B2.features

# For Bearing 3

B3.fft.amps <- fft.spectrum(data$B3Ch3)
B3.features <- c(B3.fft.amps[freq2index(ftf)],
                 B3.fft.amps[freq2index(bpfi)],
                 B3.fft.amps[freq2index(bpfo)],
                 B3.fft.amps[freq2index(bsf)])
B3.features
# calculate Key frequencies

# Strongest frequencies
n <- 5
```

```
B3.frequencies <- seq(0, 10000, length.out=length(B3.fft.amps))
sorted <- sort.int(B3.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
B3.features <- append(B3.features, B3.frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
B3.vhf <- freq2index(6000):length(B3.fft.amps)    # 6kHz plus
B3.hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
B3.mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
B3.lf <- 0:(freq2index(1250)-1)                # forcing frequency band

B3.powers <- c(sum(B3.fft.amps[vhf]), sum(B3.fft.amps[hf]), sum(B3.fft.amps[mf]), sum(B3.fft.amps[lf]))
B3.features <- append(B3.features, B3.powers)
B3.features

# For Bearing 4

B4.fft.amps <- fft.spectrum(data$B4Ch4)
B4.features <- c(B4.fft.amps[freq2index(ftf)],
                 B4.fft.amps[freq2index(bpfi)],
                 B4.fft.amps[freq2index(bpfo)],
                 B4.fft.amps[freq2index(bsf)])
B4.features
# calculate Key frequencies

# Strongest frequencies
n <- 5
B4.frequencies <- seq(0, 10000, length.out=length(B4.fft.amps))
sorted <- sort.int(B4.fft.amps, decreasing=TRUE, index.return=TRUE)
top.ind <- sorted$ix[1:n] # indexes of the largest n components
B4.features <- append(B4.features, B4.frequencies[top.ind]) # convert indexes to frequencies

# Power in frequency bands
B4.vhf <- freq2index(6000):length(B4.fft.amps)    # 6kHz plus
B4.hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
B4.mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
B4.lf <- 0:(freq2index(1250)-1)                # forcing frequency band
```

```
B4.powers <- c(sum(B4.fft.amps[vhf]), sum(B4.fft.amps[hf]), sum(B4.fft.amps[mf]), sum(B4.fft.amps[lf]))
B4.features <- append(B4.features, B4.powers)
B4.features

# Initial analysis using data mining technique on Bearing 1
summary(data$B1Ch1)
summary(data$B2Ch2)
summary(data$B3Ch3)
summary(data$B4Ch4)

# Plot variables
par(mfrow=c(4,1))
plot(data$B1Ch1, t="l") # t="l" means line plot
plot(data$B1Ch1, t="l")
plot(data$B1Ch1, t="l")
plot(data$B1Ch1, t="l")

# Apply feature extraction to reduce data
# 1. Format the full dataset
B1.fft <- fft(data$B1Ch1)
# Ignore the 2nd half, which are complex conjugates of the 1st half,
# and calculate the Mod (magnitude of each complex number)
amplitude <- Mod(B1.fft[1:(length(B1.fft)/2)])

# Calculate the frequencies
Frequency <- seq(0, 10000, length.out=length(B1.fft)/2)

#Repeat for the other Bearings
B2.fft <- fft(data$B2Ch2)
B3.fft <- fft(data$B3Ch3)
B4.fft <- fft(data$B4Ch4)

#Calculate amplitude
amplitude2 <- Mod(B2.fft[1:(length(B2.fft)/2)])
amplitude3 <- Mod(B3.fft[1:(length(B3.fft)/2)])
amplitude4 <- Mod(B4.fft[1:(length(B4.fft)/2)])

# Calculate frequencies
```

```
B2freq <- seq(0, 10000, length.out=length(B2.fft)/2)
B3freq <- seq(0, 10000, length.out=length(B3.fft)/2)
B4freq <- seq(0, 10000, length.out=length(B4.fft)/2)

# Plot
plot(amplitude ~ B1freq, t="l")
plot(amplitude2 ~ B2freq, t="l")
plot(amplitude3 ~ B3freq, t="l")
plot(amplitude4 ~ B4freq, t="l")

# 2.focus on the lower frequencies
plot(amplitude ~ Frequency, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

plot(amplitude2 ~ B2freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)

plot(amplitude3 ~ B3freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)

plot(amplitude4 ~ B4freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(1,1))
# For Bearing 1
B1freq <- seq(0, 500, length.out=length(data$B1Ch1)/2)
plot(B1.fft.amps[1:(length(data$B1Ch1)/2)] ~ Frequency, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Bearing 2
B2freq <- seq(0, 500, length.out=length(data$B2Ch2)/2)
plot(B2.fft.amps[1:(length(data$B2Ch2)/2)] ~ B2freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
```

```
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Bearing 3
B3freq <- seq(0, 500, length.out=length(data$B3Ch3)/2)
plot(B3.fft.amps[1:(length(data$B3Ch3)/2)] ~ B3freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For Bearing 4
B4freq <- seq(0, 500, length.out=length(data$B4Ch4)/2)
plot(B4.fft.amps[1:(length(data$B4Ch4)/2)] ~ Frequency, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# 3.Tabulating the top 15 frequencies
sorted <- sort.int(amplitude, decreasing=TRUE, index.return=TRUE)
top15 <- sorted$ix[1:15] # indexes of the largest 15
B1.top15f <- B1freq[top15] # convert indexes to frequencies
B1.top15f

B2.sorted <- sort.int(amplitude2, decreasing=TRUE, index.return=TRUE)
B2.top15 <- B2.sorted$ix[1:15] # indexes of the largest 15
B2.top15f <- B2freq[top15] # convert indexes to frequencies
B2.top15f

B3.sorted <- sort.int(amplitude3, decreasing=TRUE, index.return=TRUE)
B3.top15 <- B3.sorted$ix[1:15] # indexes of the largest 15
B3.top15f <- B3freq[top15] # convert indexes to frequencies
B3.top15f

B4.sorted <- sort.int(amplitude4, decreasing=TRUE, index.return=TRUE)
B4.top15 <- B4.sorted$ix[1:15] # indexes of the largest 15
B4.top15f <- B4freq[top15] # convert indexes to frequencies
```

```
B4.top15f

fft.profile <- function (dataset, n)
{
  fft.data <- fft(dataset)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  amplitude <- Mod(fft.data[1:(length(fft.data)/2)])
  # Calculate the frequencies
  frequencies <- seq(0, 10000, length.out=length(fft.data)/2)

  sorted <- sort.int(amplitude, decreasing=TRUE, index.return=TRUE)
  top <- sorted$ix[1:n] # indexes of the largest n components
  return (frequencies[top]) # convert indexes to frequencies
}


# How many FFT components should we grab as features?
n <- 5

# Set up storage for bearing-grouped data
b1 <- matrix(nrow=0, ncol=(2*n+1))
b2 <- matrix(nrow=0, ncol=(2*n+1))
b3 <- matrix(nrow=0, ncol=(2*n+1))
b4 <- matrix(nrow=0, ncol=(2*n+1))

for (filename in list.files(Beari))
{
  cat("Processing file ", filename, "\n")

  timestamp <- as.character(strptime(filename, format="%Y.%m.%d.%H.%M.%S"))

  data <- read.table(paste0(Trainingdir, filename), header=FALSE, sep="\t")
  colnames(data) <- c("b1.x", "b2.x", "b3.x", "b4.x")

  # Bind the new rows to the bearing matrices
  b1 <- rbind(b1, c(timestamp, fft.profile(data$b1.x, n)))
  b2 <- rbind(b2, c(timestamp, fft.profile(data$b2.x, n)))
```

```
  b3 <- rbind(b3, c(timestamp, fft.profile(data$b3.x, n)))
  b4 <- rbind(b4, c(timestamp, fft.profile(data$b4.x, n)))

}

write.table(b1, file=paste0(Trainingdir, "../b1.csv"), sep=",", row.names=FALSE, col.names=FALSE)
write.table(b2, file=paste0(Trainingdir, "../b2.csv"), sep=",", row.names=FALSE, col.names=FALSE)
write.table(b3, file=paste0(Trainingdir, "../b3.csv"), sep=",", row.names=FALSE, col.names=FALSE)
write.table(b4, file=paste0(Trainingdir, "../b4.csv"), sep=",", row.names=FALSE, col.names=FALSE)

rm(list=ls())
#--------------Section 05 Final Feature Extraction Approach ----------
# Re-name column names
colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

library(e1071)

# Helper functions
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}

freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

# Bearing data
Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz
```

```
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)


all.features <- function(d)
{
  # Statistical features
  features <- c(quantile(d, names=FALSE), mean(d), sd(d), skewness(d), kurtosis(d))

  # RMS
  features <- append(features, sqrt(mean(d**2)))

  # Key frequencies
  fft.amps <- fft.spectrum(d)

  features <- append(features, fft.amps[freq2index(ftf)])
  features <- append(features, fft.amps[freq2index(bpfi)])
  features <- append(features, fft.amps[freq2index(bpfo)])
  features <- append(features, fft.amps[freq2index(bsf)])

  # Strongest frequencies
  n <- 5
  frequencies <- seq(0, 10000, length.out=length(fft.amps))
  sorted <- sort.int(fft.amps, decreasing=TRUE, index.return=TRUE)
  top.ind <- sorted$ix[1:n] # indexes of the largest n components
  features <- append(features, frequencies[top.ind]) # convert indexes to frequencies

  # Power in frequency bands
  vhf <- freq2index(6000):length(fft.amps)    # 6kHz plus
  hf <- freq2index(2600):(freq2index(6000)-1) # 2.6kHz to 6kHz
  mf <- freq2index(1250):(freq2index(2600)-1) # 1.25kHz to 2.6kHz
  lf <- 0:(freq2index(1250)-1)                # forcing frequency band

  powers <- c(sum(fft.amps[vhf]), sum(fft.amps[hf]), sum(fft.amps[mf]), sum(fft.amps[lf]))
```

```
  features <- append(features, powers)

  return(features)
}

# Set up storage for bearing-grouped data
b1m <- matrix(nrow=0, ncol=(1*23))
b2m <- matrix(nrow=0, ncol=(1*23))
b3m <- matrix(nrow=0, ncol=(1*23))
b4m <- matrix(nrow=0, ncol=(1*23))

# and for timestamps
timestamp <- vector()

for (filename in list.files(Trainingdir))
{
  cat("Processing file ", filename, "\n")

  ts <- as.character(strptime(filename, format="%Y.%m.%d.%H.%M.%S"))

  data <- read.table(paste0(Trainingdir, filename), header=FALSE, sep="\t")
  colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

  # Bind the new rows to the bearing matrices
  b1m <- rbind(b1m, c(all.features(data$B1Ch1)))
  b2m <- rbind(b2m, c(all.features(data$B2Ch2)))
  b3m <- rbind(b3m, c(all.features(data$B3Ch3)))
  b4m <- rbind(b4m, c(all.features(data$B4Ch4)))

  timestamp <- c(timestamp, ts)
}

cnames <- c("Min.x", "Qu.1.x", "Median.x", "Qu.3.x", "Max.x", "Mean.x", "SD.x", "Skew.x", "Kurt.x", "RMS.x",
"FTF.x", "BPFI.x", "BPFO.x", "BSF.x", "F1.x", "F2.x", "F3.x", "F4.x", "F5.x", "VHF.pow.x", "HF.pow.x",
"MF.pow.x", "LF.pow.x")
colnames(b1m) <- cnames
colnames(b2m) <- cnames
colnames(b3m) <- cnames
```

```
colnames(b4m) <- cnames
B1Ch1 <- data.frame(timestamp, b1m)
B2Ch2 <- data.frame(timestamp, b2m)
B3Ch3 <- data.frame(timestamp, b3m)
B4Ch4 <- data.frame(timestamp, b4m)

write.table(B1Ch1, file=paste0(Trainingdir, "../B1Ch1_all.csv"), sep=",", row.names=FALSE)
write.table(B2Ch2, file=paste0(Trainingdir, "../B2Ch2_all.csv"), sep=",", row.names=FALSE)
write.table(B3Ch3, file=paste0(Trainingdir, "../B3Ch3_all.csv"), sep=",", row.names=FALSE)
write.table(B4Ch4, file=paste0(Trainingdir, "../B4Ch4_all.csv"), sep=",", row.names=FALSE)


rm(list=ls())

#--------------Section 06 Inspect Bearing-specific Datsets----------
#Reload New Bearing-Specific Datasets and inspect

B1Ch1 <- read.table(file=paste0(Trainingdir, "../B1Ch1_all.csv"), sep=",", header=FALSE)
B2Ch2 <- read.table(file=paste0(Trainingdir, "../B2Ch2_all.csv"), sep=",", header=FALSE)
B3Ch3 <- read.table(file=paste0(Trainingdir, "../B3Ch3_all.csv"), sep=",", header=FALSE)
B4Ch4 <- read.table(file=paste0(Trainingdir, "../B4Ch4_all.csv"), sep=",", header=FALSE)
head(B1Ch1)
str(B1Ch1)

# covert Factor to Numeric
B1Ch1 <- read.csv(file = '../B1Ch1_all.csv', stringsAsFactors = TRUE)
str(B1Ch1)
B2Ch2 <- read.csv(file = '../B2Ch2_all.csv', stringsAsFactors = TRUE)
B3Ch3 <- read.csv(file = '../B3Ch3_all.csv', stringsAsFactors = TRUE)
B4Ch4 <- read.csv(file = '../B4Ch4_all.csv', stringsAsFactors = TRUE)

#Descriptive statistics
library(pastecs)
stat.desc(B1Ch1)
stat.desc(B2Ch2)
stat.desc(B3Ch3)
stat.desc(B4Ch4)
```

```
#--------------Section 07 Change-point Analysis by Package changpoint----------
library(changepoint)
#Using change-point using statistical Pruned Exact Linear Time (PELT)
#Remember outer race risk (BPFO) in Bearing 1
# so plot BPFO

par(mfrow=c(2,2))

mvalue1 = cpt.mean(B1Ch1$BPFO, method="PELT")
mvalue1 = cpt.mean(B1Ch1[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue1)
vvalue1 = cpt.var(diff(B1Ch1[, 14]), method="PELT")
cpts(vvalue1)

B1.pelt <- cpt.var(diff(diff(B1Ch1[, 14]), method = "PELT"))
plot(B1.pelt, xlab = "Index")
logLik(B1.pelt)

#For Bearing 2
mvalue2 = cpt.mean(B2Ch2[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue2)
vvalue2 = cpt.var(diff(B2Ch2[, 14]), method="PELT")
cpts(vvalue2)

B2.pelt <- cpt.var(diff(diff(B2Ch2[, 14]), method = "PELT"))
plot(B2.pelt, xlab = "Index")
logLik(B2.pelt)

#For Bearing 3
mvalue3 = cpt.mean(B3Ch3[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue3)
vvalue3 = cpt.var(diff(B3Ch3[, 14]), method="PELT")
cpts(vvalue3)

B3.pelt <- cpt.var(diff(diff(B3Ch3[, 14]), method = "PELT"))
plot(B3.pelt, xlab = "Index")
logLik(B3.pelt)
```

```
#For Bearing 4
mvalue4 = cpt.mean(B4Ch4[, 14], method="PELT") #mean changepoints using PELT
cpts(mvalue4)
vvalue4 = cpt.var(diff(B4Ch4[, 14]), method="PELT")
cpts(vvalue4)

B4.pelt <- cpt.var(diff(diff(B4Ch4[, 14]), method = "PELT"))
plot(B4.pelt, xlab = "Index")
logLik(B4.pelt)

#--------------Section 08 Change-point Analysis by Package strucchange----------
library(strucchange)
par(mfrow=c(4,2))

#Bearing 1
B1.ts<- ts(B1Ch1[, 14],frequency=1)
B1.bp <-breakpoints((B1.ts~1))
B1.bp
summary(B1.bp)
plot(B1.bp)

# plot data with breakpoint times
plot(B1.ts)
lines(fitted(B1.bp, breaks = 1), col = 4)
lines(confint(B1.bp, breaks = 1))

#Bearing 2
B2.ts<- ts(B2Ch2[, 14],frequency=1)
B2.bp <-breakpoints((B2.ts~1))
B2.bp
summary(B2.bp)
plot(B2.bp)

# plot data with breakpoint times
plot(B2.ts)
lines(fitted(B2.bp, breaks = 0), col = 4)
lines(confint(B2.bp, breaks = 0))
```

```
#Bearing 3
B3.ts<- ts(B3Ch3[, 14],frequency=1)
B3.bp <-breakpoints((B3.ts~1))
B3.bp
summary(B3.bp)
plot(B3.bp)

# plot data with breakpoint times
plot(B3.ts)
lines(fitted(B3.bp, breaks = 0), col = 4)
lines(confint(B3.bp, breaks = 0))

#Bearing 4
B4.ts<- ts(B4Ch4[, 14],frequency=1)
B4.bp <-breakpoints((B4.ts~1))
B4.bp
summary(B4.bp)
plot(B4.bp)

# plot data with breakpoint times
plot(B4.ts)
lines(fitted(B4.bp, breaks = 0), col = 4)
lines(confint(B4.bp, breaks = 0))

#F-stats
par(mfrow=c(2,2))
B1.Fstats <- Fstats((B1Ch1[, 14]) ~ 1)
plot(B1.Fstats)

B2.Fstats <- Fstats((B2Ch2[, 14]) ~ 1)
plot(B2.Fstats)

B3.Fstats <- Fstats((B3Ch3[, 14]) ~ 1)
plot(B3.Fstats)

B4.Fstats <- Fstats((B4Ch4[, 14]) ~ 1)
plot(B4.Fstats)
```

```
# Significant test p-value
sctest(B1.Fstats, type = "supF")
sctest(B2.Fstats, type = "supF")
sctest(B3.Fstats, type = "supF")
sctest(B4.Fstats, type = "supF")

#--------------Section 09 Explore & Compare Test files at the Cpts-----------

#----Import & first data file (healthy file)---------
data <- read.table(paste0(Trainingdir,"2004.02.18.15.22.39"), header=FALSE, sep="\t")
head(data)

# Re-name column names
colnames(data) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")
# Apply feature extraction to reduce data and plot
# Calculate the four frequencies for the
# dataset bearings (i.e.BPFO, BPFI, BSF, & FTF)

Bd <- 0.331 # ball diameter, in inches
Pd <- 2.815 # pitch diameter, in inches
Nb <- 16 # number of rolling elements
a <- 15.17*pi/180 # contact angle, in radians
s <- 2000/60 # rotational frequency, in Hz
ratio <- Bd/Pd * cos(a)
ftf <- s/2 * (1 - ratio)
bpfi <- Nb/2 * s * (1 + ratio)
bpfo <- Nb/2 * s * (1 - ratio)
bsf <- Pd/Bd * s/2 * (1 - ratio**2)

# Generate the first four features of the bearings
fft.spectrum <- function (d)
{
  fft.data <- fft(d)
  # Ignore the 2nd half, which are complex conjugates of the 1st half,
  # and calculate the Mod (magnitude of each complex number)
  return (Mod(fft.data[1:(length(fft.data)/2)]))
}
```

```
freq2index <- function(freq)
{
  step <- 10000/10240 # 10kHz over 10240 bins
  return (floor(freq/step))
}

B1.fft.amps <- fft.spectrum(data$B1Ch1)
features <- c(B1.fft.amps[freq2index(ftf)],
              B1.fft.amps[freq2index(bpfi)],
              B1.fft.amps[freq2index(bpfo)],
              B1.fft.amps[freq2index(bsf)])
features

# 1. Format the full dataset
B1.fft <- fft(data$B1Ch1)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
amplitude <- Mod(B1.fft[1:(length(B1.fft)/2)])

# Calculate the frequencies
B1freq <- seq(0, 10000, length.out=length(B1.fft)/2)

# Plot
plot(amplitude ~ B1freq, t="l")

# 2.focus on the lower frequencies
plot(amplitude ~ B1freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(3,1))
# For Bearing 1
B1freq <- seq(0, 310, length.out=length(data$B1Ch1)/2)
plot(B1.fft.amps[1:(length(data$B1Ch1)/2)] ~ B1freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)
```

```
#-----------Import file representing Changepoint (PELT) 968 ------

# Get timestamp of file representing row 968
B1Ch1[968,]
# timestamp = 2004-02-19 03:42:39; read test file 2004.02.19.03.42.39
data1 <- read.table(paste0(Trainingdir,"2004.02.19.03.42.39"), header=FALSE, sep="\t")
head(data1)

# Re-name column names
colnames(data1) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

# Apply feature extraction to reduce data and plot
d1B1.fft.amps <- fft.spectrum(data1$B1Ch1)
d1features <- c(d1B1.fft.amps[freq2index(ftf)],
              d1B1.fft.amps[freq2index(bpfi)],
              d1B1.fft.amps[freq2index(bpfo)],
              d1B1.fft.amps[freq2index(bsf)])
d1features

# 1. Format the full dataset
d1B1.fft <- fft(data1$B1Ch1)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d1amplitude <- Mod(d1B1.fft[1:(length(d1B1.fft)/2)])

# Calculate the frequencies
d1B1freq <- seq(0, 10000, length.out=length(d1B1.fft)/2)

# Plot
plot(d1amplitude ~ d1B1freq, t="l")

# 2.focus on the lower frequencies
plot(d1amplitude ~ d1B1freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(3,1))
# For Bearing 1
d1B1freq <- seq(0, 310, length.out=length(data1$B1Ch1)/2)
```

```
plot(d1B1.fft.amps[1:(length(data1$B1Ch1)/2)] ~ d1B1freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

#---------Import Test files @ Changepoint (Strucchange) = 837---------
# Get timestamp of file
B1Ch1[837,]
# timestamp = 2004-02-18 05:52:39; read test file
data2 <- read.table(paste0(Trainingdir,"2004.02.18.05.52.39"), header=FALSE, sep="\t")
head(data2)

# Re-name column names
colnames(data2) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")
# Re-name column names
colnames(data2) <- c("B1Ch1", "B2Ch2", "B3Ch3", "B4Ch4")

# Apply feature extraction to reduce data and plot
d2B1.fft.amps <- fft.spectrum(data2$B1Ch1)
d2features <- c(d2B1.fft.amps[freq2index(ftf)],
                d2B1.fft.amps[freq2index(bpfi)],
                d2B1.fft.amps[freq2index(bpfo)],
                d2B1.fft.amps[freq2index(bsf)])
d2features

# 1. Format the full dataset
d2B1.fft <- fft(data2$B1Ch1)
# Ignore the 2nd half and calculate the Mod (magnitude of each complex number)
d2amplitude <- Mod(d2B1.fft[1:(length(d2B1.fft)/2)])

# Calculate the frequencies
d2B1freq <- seq(0, 10000, length.out=length(d2B1.fft)/2)

# Plot
plot(d2amplitude ~ d2B1freq, t="l")

# 2.focus on the lower frequencies
```

```
plot(d2amplitude ~ d2B1freq, t="l", xlim=c(0,1000), ylim=c(0,500))
axis(1, at=seq(0,1000,100), labels=FALSE)  # add more ticks

# For clarity, zoom in to frequencies up to 500Hz
par(mfrow=c(4,1))
# For Bearing 1
d2B1freq <- seq(0, 310, length.out=length(data2$B1Ch1)/2)
plot(d2B1.fft.amps[1:(length(data2$B1Ch1)/2)] ~ d1B1freq, t="l", ylab="Relative power")
abline(v=bsf,col="red",lty=3)
abline(v=bpfo,col="blue",lty=3)
abline(v=bpfi,col="green",lty=3)
abline(v=ftf,col="violet",lty=3)

# For clarity, zoom in to frequencies up to 236.403-236.405Hz & compare
par(mfrow=c(3,1))
B1freq <- seq(0, 310, length.out=length(data$B1Ch1)/2)
plot(B1.fft.amps[1:(length(data$B1Ch1)/2)] ~ B1freq, t="l", ylim= c(0,150),ylab="Relative power")
abline(v=bpfo,col="blue",lty=3)

d1B1freq <- seq(236.40345, 236.4035, length.out=length(data1$B1Ch1)/2)
plot(d1B1.fft.amps[1:(length(data1$B1Ch1)/2)] ~ d1B1freq, t="l", ylim= c(0,150), ylab="Relative power")
abline(v=bpfo,col="blue",lty=3)

d2B1freq <- seq(236.40345, 236.4035, length.out=length(data2$B1Ch1)/2)
plot(d2B1.fft.amps[1:(length(data2$B1Ch1)/2)] ~ d1B1freq, t="l", ylim= c(0,150), ylab="Relative power")
abline(v=bpfo,col="blue",lty=3)

#---------------Section 10 Investigating Change-points with trend in RMS of Bearing -------------
par(mfrow=c(1,1))
plot(B1Ch1$RMS.x, t="l",ylab="RMS")
plot(B2Ch2$RMS.x, t="l",ylab="RMS")
plot(B3Ch3$RMS.x, t="l",ylab="RMS")
plot(B4Ch4$RMS.x, t="l",ylab="RMS")

#-----Section 11 Re-load and combine Bearing-specific datasets------------------

library(car)
```

```
# Create data frame with columns of interest using column indices
# Displays column 12-15
dfnew1 <- B1Ch1[,c(12:15)]
dfnew2 <- B2Ch2[,c(12:15)]
dfnew3 <- B3Ch3[,c(12:15)]
dfnew4 <- B4Ch4[,c(12:15)]

# Re-name column names
colnames(dfnew1) <- c("FTF.B1", "BPFI.B1", "BPFO.B1"," BSF.B1")
colnames(dfnew2) <- c("FTF.B2", "BPFI.B2", "BPFO.B2"," BSF.B2")
colnames(dfnew3) <- c("FTF.B3", "BPFI.B3", "BPFO.B3"," BSF.B3")
colnames(dfnew4) <- c("FTF.B4", "BPFI.B4", "BPFO.B4"," BSF.B4")

# Merge the  data frames
# Use the cbind function to combine data frames side-by-side:
dfnew <-cbind(dfnew1,dfnew2,dfnew3,dfnew4)
dfnew

#-----Section 12 Investigate Correlation for Interaction Effect------------------------------

# Creat a sub-set for interactions up to time index 968
BPFO968 <-dfnew[c(1:968),c(3,7,11,15)]
str(BPFO968)

# Inspect
boxplot(BPFO968, horizontal=TRUE)
pairs(BPFO968,panel=panel.smooth)

# Investigate correlation
library("PerformanceAnalytics")
chart.Correlation(BPFO968, histogram=TRUE, pch=19)

# Creat a sub-set for time index 837
BPFO837 <-dfnew[c(1:837),c(3,7,11,15)]
str(BPFO837)

boxplot(BPFO837, horizontal=TRUE)
pairs(BPFO837,panel=panel.smooth)
```

```
# Investigate correlation
chart.Correlation(BPFO837, histogram=TRUE, pch=19)


# Apply Decision Tree models to determine predictors and moderators for regression models
library(tree)

# For time index 837
Tmodel837<-tree(BPFO.B1~.,data=BPFO837)
plot(Tmodel837)
text(Tmodel837)

# For time index 968
Tmodel968<-tree(BPFO.B1~.,data=BPFO968)
plot(Tmodel968)
text(Tmodel968)

# Creat Regression model to investigate interaction effect

# for time index 837
model.1 <- lm(BPFO.B1 ~ BPFO.B3+BPFO.B2, data=BPFO837)

model.2 <- lm(BPFO.B1 ~ BPFO.B3*BPFO.B2,data=BPFO837)

# for time index 968
model.3 <- lm(BPFO.B1 ~ BPFO.B3+BPFO.B2, data=BPFO968)

model.4 <- lm(BPFO.B1 ~ BPFO.B3*BPFO.B2,data=BPFO968)

#Show the results
library(stargazer)
stargazer(model.1,model.2,model.3,model.4,type="text",
          column.labels = c("model.1","model.2","model.3","model.4"),
          intercept.bottom = FALSE,
          single.row=FALSE,
          notes.append = FALSE,
          header=FALSE)
```

```
# compar interaction models
anova(model.1, model.2)
anova(model.3, model.4)

# Investigate significant interaction
install.packages("devtools")
devtools::install_github("jacob-long/jtools")

library(jtools)
library(interactions)
library(ggplot2)

# for time index 837
sim_slopes(model.2, pred = BPFO.B3, modx = BPFO.B2, jnplot = TRUE)

# Visualize interaction effect
interact_plot(model.2, pred = BPFO.B3, modx = BPFO.B2, interval = TRUE)
probe_interaction(model.2, pred = BPFO.B3, modx = BPFO.B2, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)

# for time index 968
sim_slopes(model.4, pred = BPFO.B3, modx = BPFO.B2, jnplot = TRUE)

# Visualize interaction effect
interact_plot(model.4, pred = BPFO.B3, modx = BPFO.B2, interval = TRUE)
probe_interaction(model.4, pred = BPFO.B3, modx = BPFO.B2, cond.int = TRUE,
                  interval = TRUE,  jnplot = TRUE)


#----------Section 13 Association between Features-----------------------
# Creat subset for bearing defect parameters
BOne837 <-dfnew[c(1:837),c(2:4)]
str(BOne837)
colnames(BOne837) <- c("BPFI_B1", "BPFO_B1","BSF_B1")
str(BOne837)
head(BOne837)
```

```
BOne968 <-dfnew[c(1:968),c(2:4)]
str(BOne968)
colnames(BOne968) <- c("BPFI_B1", "BPFO_B1","BSF_B1")
str(BOne968)
head(BOne968)

# Creat a tree models to determine predictor and moderator Bearings
# time index 837
TmodelBOne837<-tree(BPFO_B1~.,data=BOne837)
plot(TmodelBOne837)
text(TmodelBOne837)

#time index 968
TmodelBOne968<-tree(BPFO_B1~.,data=BOne968)
plot(TmodelBOne968)
text(TmodelBOne968)

# singlenode trees therefore no further investigation

#-----Section 14-----------------------------------------
# remove all variables from the environment
rm(list=ls())
```