Durham E-Theses

Exploiting Robust Multivariate Statistics and Data Driven Techniques for Prognosis and Health Management

GODWIN, JAMIE,LEIGH

Exploiting Robust Multivariate Statistics and Data Driven Techniques for Prognosis and Health Management

Jamie L. Godwin

Durham University

School of Engineering and Computing Sciences
Durham University

A thesis submitted for the degree of
Doctor of Philosophy

May 31, 2015

Abstract

This thesis explores state of the art robust multivariate statistical methods and data driven techniques to holistically perform prognostics and health management (PHM). This provides a means to enable the early detection, diagnosis and prognosis of future asset failures. In this thesis, the developed PHM methodology is applied to wind turbine drive train components, specifically focussed on planetary gearbox bearings and gears.

A novel methodology for the identification of relevant time-domain statistical features based upon robust statistical process control charts is presented for high frequency bearing accelerometer data. In total, 28 time-domain statistical features were evaluated for their capabilities as leading indicators of degradation. The results of this analysis describe the extensible multivariate "Moments' model" for the encapsulation of bearing operational behaviour. This is presented, enabling the early degradation of detection, predictive diagnostics and estimation of remaining useful life (RUL).

Following this, an extended physics of failure model based upon low frequency SCADA data for the quantification of wind turbine gearbox condition is described. This extends the state of the art, whilst defining robust performance charts for quantifying component condition. Normalisation against loading of the turbine and transient states based upon empirical data is performed in the bivariate domain, with extensibility into the multivariate domain if necessary. Prognosis of asset condition is found to be possible with the assistance of artificial neural networks in order to provide business intelligence to the planning and scheduling of effective maintenance actions.

These multivariate condition models are explored with multivariate distance and similarity metrics for to exploit traditional data mining techniques for tacit knowledge extraction, ensemble diagnosis and prognosis. Estimation of bearing remaining useful life is found to be possible, with the derived technique correlating strongly to bearing life ($r = .96$).

Declaration

The work in this thesis is based on research carried out by Jamie Leigh Godwin under the supervision of Dr. Peter Matthews and Prof. Peter Tavner within the School of Engineering and Computing Sciences at the University of Durham in the United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification. It all my own work unless referenced to the contrary in the text.

Dedication

This thesis is dedicated to my Grandmother Christine, my late Grandfather Bob, my Mother Val, my Father Mark and my Brother John.

Thank you all for believing in me.

Acknowledgements

List of publications

The work presented in this thesis appears in part in the following internationally recognised, peer-reviewed publications:

- Godwin, J. L. & Matthews, P. C. (2014). *Robust Statistical Methods for Rapid Data Labelling*. In Bhatnatgar, V. Editor, *Data Mining and Analysis in the Engineering Field* (pp. 107–141). Hershey, Pennsylvania (USA): IGI Global.

- Godwin, J. L., & Matthews, P. (2014, January). *Rapid labelling of SCADA data to extract transparent rules using RIPPER*. In Reliability and Maintainability Symposium (RAMS), 2014 Annual (pp. 1-7). IEEE, Colorado Springs, CO, USA.
  T.L. Fagan Award Recipient

- Godwin, J.L.; Matthews, P., "Prognosis of wind turbine gearbox failures by utilising robust multivariate statistical techniques," Prognostics and Health Management (PHM), 2013 IEEE Conference on , vol., no., pp.1,8, 24-27 June 2013 doi: 10.1109/ICPHM.2013.6621428

- Jamie L. Godwin & Peter Matthews. (2014, September). "Accurate empirical encapsulation of asset operational behaviour utilizing elitist memetic algorithms for reliability analysis." In proceedings of the 27th International Congress of Condition Monitoring and Diagnostic Engineering (COMADEM). pp. 1 – 8. 16 - 18 September, Brisbane, Australia.
  Best Paper Recipient

- Godwin, J. L., Matthews, P. C. & Watson, C. (2014) "Robust Multivariate Statistical Ensembles for Bearing Fault Detection and Identification". In proceedings of the IEEE Prognostics and health management (PHM) conference. pp. 1 – 8. 22 - 24 June, Spokane, Washington State, USA

- Godwin, J.L., Matthews, P.C., and Chen, B. (2014, March). *Prediction of Wind turbine Gearbox Condition Based on Hybrid Prognostic Techniques with Robust*

Multivariate Statistics and Artificial Neural Networks. In proceedings of the annual conference of the European Wind Energy Association (EWEA). pp.1 -10. 10-13 March. Barcelona, Spain.

Furthermore, the following additional publications have arisen as a direct result of the research undertaken throughout this thesis:

- Godwin, J. L., & Matthews, P. Classification and Detection of Wind Turbine Pitch Faults Through SCADA Data Analysis. International Journal of prognostics and health management, 4(2), 16.

- Jamie L. Godwin, Peter Matthews and Christopher Watson. (2013). Classification and Detection of Electrical Control System Faults Through SCADA Data Analysis. Chemical Engineering Transactions 33(1). pp. 1 – 6.

- Jamie L. Godwin & Peter Matthews. (2013). On The Use of Robust Multivariate Statistical Methods for the Prognosis of Wind Turbine Pitch Faults. In proceedings of the 26th International Congress of Condition Monitoring and Diagnostic Engineering (COMADEM). pp. 1 – 8. 11 - 13 June, Helsinki, Finland.

- Godwin, J. L. & Matthews, P. C. (2014, September) "A Low Frequency Univariate Model for the Effective Diagnosis and Prognosis of Bearing Signals Based Upon High Frequency Data". In proceedings of the annual conference of the prognostics and health management society (PHMSociety). pp. 20 – 29. September 29 - 02 October, Fort Worth, Texas, USA.

- Jamie L. Godwin & Peter Matthews. (2014). "Wind turbine health: metrics for operational analysis". In proceedings of the 2015 European Wind Energy Association Analysis of operating wind farms workshop. 9 - 10 December, Malmo Sweden.

- Marc Hilbert & Jamie L. Godwin. (2014, May). Comparison of state of the art SCADA and High Sampled Data Condition Monitoring for Wind Turbines. In proceedings of the annual conference of the machinery failure and prevention technology (MFPT) society. 20 -22 May, Virginia Beach, Virginia, USA.

Contents

List of Figures

List of Tables

Nomenclature

χ_p^2	χ^2 distribution with p degrees of freedom
L_ϵ	ϵ insensitive loss function
$AR(p)$	AR model of order q
$ARDL(p,r)$	$ARDL$ model of orders p, r
$ARMA(p,q)$	$ARMA$ model of orders p, q
SB_i	i^{th} Sideband
$MA(q)$	MA model of order q
$X \sim F_{p,q}$	X is distributed according to a Fisher distribution with p, q DOF
$|x|$	Absolute value of x
A_x	Ambiguity function
$\overline{TSA_{angular}}(t)$	Angular re-sampling of signal t for TSA
$F'(x)$	Approximation of function $f'(x)$
$\bar{x}$	Arithmetic mean
M	Arithmetic mean
F_a	Assembly phase frequency
$\beta(p,q)$	Beta distribution with p, q degrees of freedom
$b_{i(l)}$	Bias weight from neuron i in layer l
$\mathcal{O}$	Big-Oh
$BP(x)$	Breusch-Pagan test statistic

$D_{Canberra}(X,Y)$	Canberra distance
H_{CS}	Chao-Shen entropy
$D_{Chebychev}(X,Y)$	Chebyshev distance
κ	Cohens κ statistic
$CC(x)$	Complex cepstrum of x
$W(a,b)$	Continuous wavelet transform given shape a and time b
$D_{Cosine}(X,Y)$	Cosine distance
Σ	Covariance matrix
x_{CF}	Crest factor
$\det(\Sigma)$	Determinant of Σ
DF	Dickey Fuller unit root test statistic
$\overline{RUL_t}$	Estimate of component remaining useful life at time t
Σ_h	Estimate of sample covariance matrix given dataset h
$\hat{\mu}$	Estimate of the mean
$D_{Euclidean}(X,Y)$	Euclidean distance
$\dot{\lambda}$	EWMA chart weighting constant
z_i	EWMA value at instance i
λ	Failure rate
FFT_k	Fast Fourier transform
S_f	Feature score
$F_{p,q}$	Fisher distribution with p,q degrees of freedom
$\hat{f}(\xi)$	Fourier transform of ξ
η_{gear}	Gear efficiency
F_{rg}	Gear frequency
F_m	Gear mesh frequency

$\Delta\omega$ Gradient increment in Levenberg-Marquardt training

q Heat generated

T Henze and Zirklers' T

μ_6 Hyperflatness

μ_5 Hyperskewness

$f'(x)$ Imaginary arbitrary function

x_{IQR} Interquartile range

$f(\xi)$ Inverse Fourier transform of ξ

X^{-1} Inverse of $\mathbf{X}$

$F - MCD_{iteration}$ Iterations performed by the Fast-MCD algorithm.

H_{Shrink} James-Stein shrinkage entropy

D Kolmogorov-Smirnov test statistic

H_{KT} Krichevsky Trofimov entropy

$KPSS$ Kwiatkowski-Shillips-Schmidt-Shin test statistic

$D_{LW}(X, Y)$ Lance-Williams distance

L Limit (in deviations)

$\hat{\lambda}$ Lyapunov exponent

$D_{Mahalanobis}(X, Y)$ Mahalanobis distance between multivariate vectors $\mathbf{X}$ and $\mathbf{Y}$

$D_{Mahalanobis}(X)$ Mahalanobis distance of points $\mathbf{X}$ given $\boldsymbol{\mu}'$ and Σ'

$D_{Manhattan}(X, Y)$ Manhattan distance

x_{max} Maximum

$\tilde{x}$ Median

x_{MAD} Median absolute deviation

H_{MM} Miller-Madow entropy

x_{min} Minimum

$D_{Minkowski}(X, Y, p)$	Minkowski distance of order p between points $\mathbf{X}$ and $\mathbf{Y}$
x_{mode}	Mode
$\psi(x)$	Morlet wavelet transform of x
$\hat{x}_i$	Normalised value of x_i
Z_l	Number of neurons in layer l
$I(k)$	Order of integration of k
$v_{i(l)}$	Output of neuron i in layer l
x_{PAPR}	Peak to average power ratio
r	Pearsons' product moment correlation coefficient
$D_{Penrose}(X, Y)$	Penrose distance
F_{rp}	Pinion frequency
$PC(x)$	Power cepstrum of x
P_{out}	Power output
$\overline{x_{rms}}$	Quadratic mean
x_{range}	Range
$R(w)$	Regularised risk function given weight vector w
Δt	Rise in temperature
$D_{RMD}(X)$	RMD from $\mathbf{X}$ given a robustly defined $\boldsymbol{\mu}'$ and Σ'.
RMD_{Ball}	RMD with centres derived on a ball (roller element) fault
RMD_{Inner}	RMD with centres derived on an inner race fault
RMD_{Outer}	RMD with centres derived on an outer race fault
RMD_{Norm}	RMD with centres derived on normal operational behaviour
Σ'	Robust estimate of sample covariance matrix
G_2	Sample kurtosis
n	Sample size

G_1	Sample skewness
H_S	Shannon entropy
W	Shapiro Wilks' W test statistic
$STFT(x)$	Short time Fourier transform of x
r_s	Spearmans' Rho correlation coefficient
σ	Standard deviation
SD	Standard deviation
Q_p	The pth quartile
F_{tr}	Tooth repeat (hunting tooth) frequency
f_l	Transfer function in layer l
X^T	Transpose of $\mathbf{X}$
x_{TM}	Tri-mean
$\overline{TSA}(t)$	TSA of signal t
σ^2	Variance
$\bar{X}$	Vector of means
$\omega_{i,j(l)}$	Weight from neuron j in layer l to neuron i in the previous layer
WVD_x	Wigner-Ville distribution of x
$\bar{x_w}$	Winsorized mean
w	Work done

List of abbreviations

AANN . . . . Auto-associative neural network

AIC . . . . Akaike information criterion

ANN . . . . Artificial neural network

ANOVA . . . . Analysis of variance

AR . . . . Auto-regressive

ARIMA . . . . Auto-regressive integrated moving average

BPFI . . . . Ball pass frequency inner

BPFO . . . . Ball pass frequency outer

BSF . . . . Ball spin frequency

CBM . . . . Condition based maintenance

CBR . . . . Case based reasoning

CF . . . . Crest factor

CMS . . . . Condition monitoring system

CNN . . . . Condensed nearest neighbour rule

CNN . . . . Cascade neural network

CO$_2$ . . . . Carbon dioxide

CPM . . . . Change process monitoring

CWT . . . . Continuous wavelet transform

DECC . . . . Department for energy and climate change

DL .... Distributed lag

DOF .... Degrees of freedom

DTM .... Decision table majority

EMA .... Elitist memetic algorithm

EWMA .... Exponentially weighted moving average

FFT .... Fast Fourier transform

FMEA .... Failure methods and effects analysis

FMECA .... Failure methods and effects criticality analysis

FN .... False negative

FP .... False positive

FTF .... Fundamental train frequency

GA .... Genetic algorithm

GPM .... General path model

HFCMS .... High frequency condition monitoring system

HMM .... Hidden Markov model

HSS .... High speed shaft

IMS .... Intermediate shaft

IQR .... Interquartile range

kNN .... k-Nearest neighbour

KPSS .... Kwiatkowski-Phillips-Schmidt-Shin

KT .... Krichevsky-Trofimov

LCoE .... Levelised cost of energy

LSS .... Low speed shaft

MA .... Moving average

MAD .... Median absolute deviation

MAS .... Multi-agent system

MCD .... Minimum covariance determinant

MLE .... Maximal Lyapunov exponent

MSE .... Mean squared error

MTTF .... Mean time to failure

MTTR .... Mean time to repair

MVE .... Minimum volume ellipsoid

NCL .... Neighbourhood cleaning rule

NFF .... No fault found

NREL .... National renewable energy laboratory

NSET .... Non-linear state estimation

O&M .... Operations and maintenance

ODD .... On demand diagnosis

OEM .... Original equipment manufacturer

OSS .... One sided selection

PAPR .... Peak to average power ratio

PCA .... Principal component analysis

PDF .... Probability density function

PHM .... Prognostics and health management

PPM .... Parts per million

PSA .... Power spectrum amplitude

PV .... Photovoltaic

RCM .... Reliability centred maintenance

RMD .... Robust Mahalanobis distance

RMS .... Root mean square

RO .... Research objective

ROI .... Return on investment

RQ .... Research question

RUL .... Remaining useful life

SARIMA .... Seasonal ARIMA

SBPF .... Side band power factor

SCADA .... Supervisory control and data acquisition

SHM .... Structural health monitoring

SMOTE .... Synthetic minority oversampling technique

SOM .... Self organising map

SOP .... Standard operating procedure

SPC .... Statistical process control

SPR .... Statistical pattern recognition

SSA .... Singular spectrum analysis

STFT .... Short time Fourier transform

SVD .... Singular value decomposition

SVM .... Support vector machine

SVMR .... Support vector machine regression

TFR .... Time frequency representation

TN .... True negative

TP .... True positive

TSA .... Time synchronous averaging

WSD .... Wind speed difference

WT .... Wind turbine

WVD .... Wigner-Ville distribution

Chapter 1

The Need for Viable Renewable Energy

This chapter briefly motivates the research included in this thesis. An overview of renewable energy technologies, the cost of both traditional and renewable energy is discussed and the impact of climate change is detailed. Following this, a short discussion of wind turbine reliability (with a specific focus on offshore wind) and the challenges for operations and maintenance (O&M) is given. Finally, the contributions of this thesis in particular and the structure of the thesis are detailed.

1.1 Introduction

Renewable sources of energy have been exploited for thousands of years. However, in recent years, interest in renewable energy has been increasing. Historically, biomass
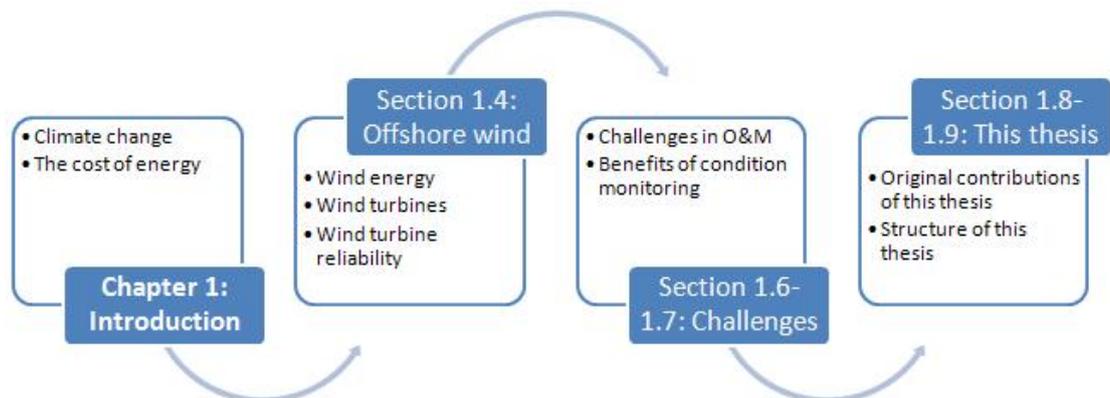
Figure 1.1: Overview of the structure of chapter 1.

has been used to fuel fires, wind has been used to power sail boats across both rivers and oceans, and also to mill grain for flour.

Up until the late 1800's, renewable energies dominated the energy mix within the US (U.S. Energy Information Administration, 2011). It wasn't until the advent of the industrial revolution and the necessity to utilise materials with a higher energy density than wooden biomass (for instance, coal) that fossil fuels became the dominant source of energy. At the turn of the 20th century, oil and natural gas contributed approximately the same energy as wind and solar energy contributed at the turn of the 21st century. The energy mix of the US over the last 236 years is presented in figure 1.2.

Ever since the domination of fossil fuels in the energy mix, there has been pressure from many groups to curb their use and return to replenishable sources of energy. The concern of exhausting fossil fuels has been an issue throughout their use, with Professor Augustine Mouchot stating in April of 1870 that:

> *"The time will arrive when the industry of Europe will cease to find those natural resources, so necessary for it. Petroleum springs and coal mines are not inexhaustible but are rapidly diminishing in many places. Will man, then, return to the power of water and wind? Or will he emigrate where the most powerful source of heat sends its rays to all? History will show what will come." (Ploutz, 2011).*
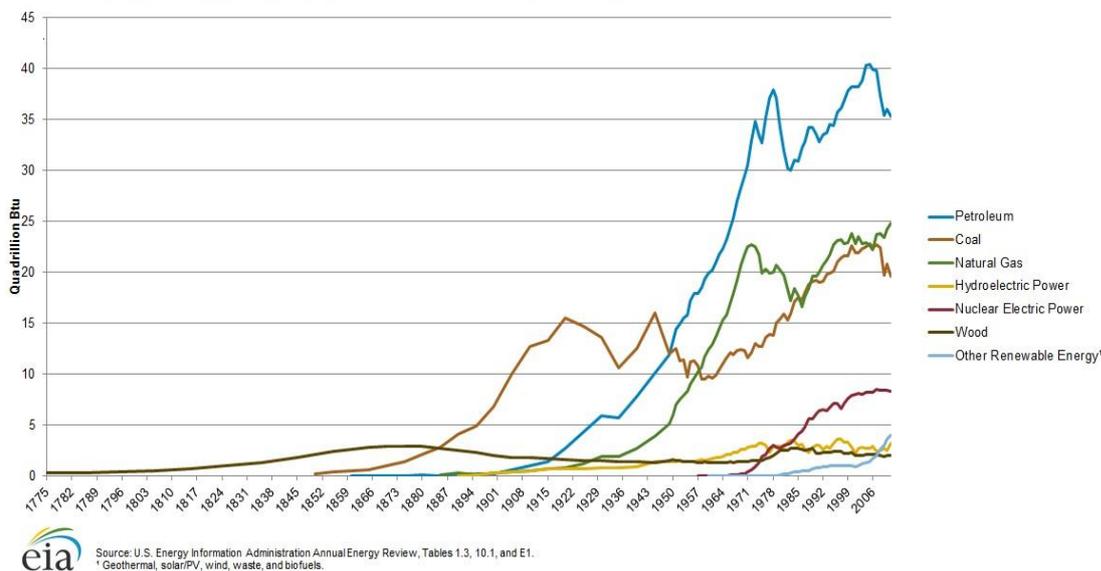
Figure 1.2: US energy consumption from 1775 to 2009 (U.S. Energy Information Administration, 2011).

Following this and prompted by the finite supply of coal, a more concerted effort was placed into the research of renewables. In particular, photovolataics. These were first defined in 1839 by Edmond Bequerel (Williams, 1960) and subsequently explored by Willoughby Smith who found selenium to be photo-conductive in 1873 (Smith, 1873).

It was not until the theory of peak oil was published by M. King Hubbert in the June of 1956 (Hubbert, 1956) that more attention was given to the potential future crisis due to finite fossil fuel reserves. Estimates based upon this theory expect oil production to begin imminent decline. Optimistic estimates show forecasts of this peak occurring around 2020 (Miller and Sorrell, 2014), with pessimistic forecasts showing this peak to have already occurred as early as 2007 (Aleklett et al., 2010).

Due to reliance on fossil fuels and the need for energy independence, the oil crisis of 1973 renewed fears of security of supply and the interest in renewable energy increased on a global scale (Van Gool, 1997). Further pressure to move to renewable energy sources occurred in the 1980's with the establishment of the Intergovernmental Panel on Climate Change (1988) (IPCC) and UK Prime Minister Margaret Thatcher calling for action to occur to deal with the issue of global warming (Thatcher, 1990).

1.2   Climate change

As the use of fossil fuels such as coal, natural gas and petroleum have increased, more carbon dioxide has become present within the atmosphere. Carbon dioxide (CO_2) is commonly referred to as a "greenhouse gas" – those gasses which absorb and emit infrared radiation – and as such, are closely correlated to global temperatures (Pachauri and Reisinger, 2007).

Figure 1.3 shows the correlation of CO_2 levels in parts per million (PPM) with the associated global temperature (Celsius) over the period of 1964 to 2013. Although there is variability across shorter time frames, a warming trend can be observed. Currently, the general consensus between climate scientists is that 97% of climate scientists agree that trends in warming of the climate over the past century are very likely due to human activity (Anderegg et al., 2010). This is echoed by the IPCC who in their 2013 report state that not only is CO_2 the largest driver of global warming but also that:

> *"Human influence has been detected in warming of the atmosphere and the ocean, in changes in the global water cycle, in reductions in snow and ice, in global mean sea level rise, and in changes in some climate extremes [...] It is extremely likely (95-100%) that human influence has been the dominant cause of the observed warming since the mid-20th century'.'* (Stocker et al., 2013).
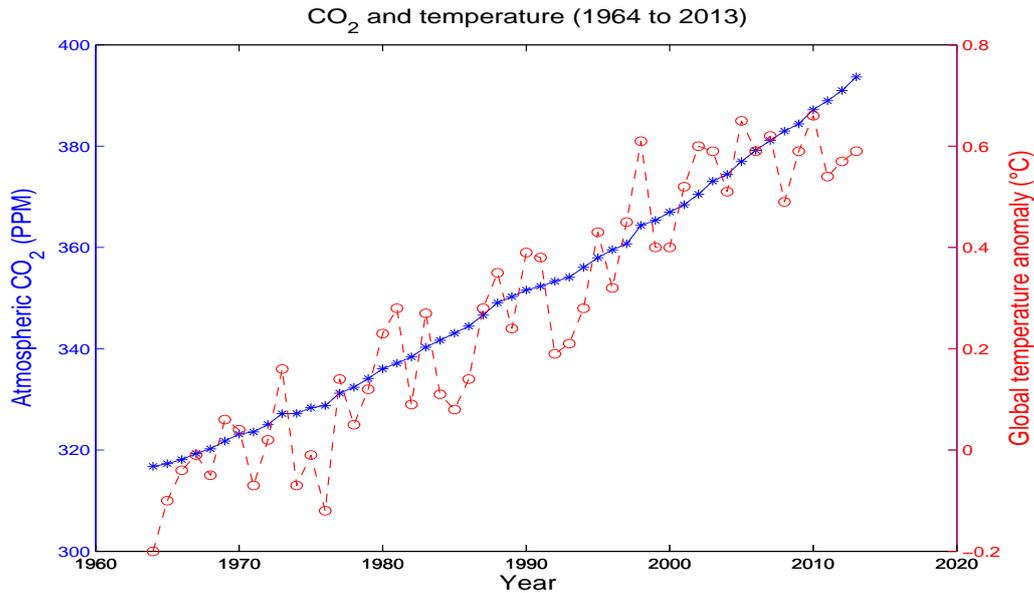
Figure 1.3: CO_2 (PPM) and temperature (Celsius) from 1964 to 2008.

Global warming is a serious issue for the entire population. There will be significant financial implications across all sectors. The most notable effect of global warming due to climate change will be on agricultural food production. Under current scenarios, staple crops such as Maize could see a yield decrease of up to 30% with decreases of similar staples in Southern Asia decreasing by up to 10% (Lobell et al., 2008). Warming of above 3 degrees Celsius in 2100 (relative to 2000 levels), is expected to lead to a net reduction in global food production (Battisti and Naylor, 2009). Given that the population is expected to double by this time, this may potentially lead to significant malnutrition in less developed countries (Pachauri and Reisinger, 2007).

Similarly, warming will cause an increased reduction of glacial mass leading to a rise in sea levels. Sea level rises of 200mm would displace 740,000 people in Nigeria alone, with approximately 100 million people living globally within three feet of sea level. Current estimates which have been established and independently verified place an expectation of an increase of 32 centimetres above the 1980 - 1999 period by 2050 (Onofeghara and Okali, 1990; Awosika et al., 1992).

1.3   Cost of energy

Due to CO_2 being the main contributor to global warming (Pachauri and Reisinger, 2007), it is essential to move away from carbon intensive energy production thereby

mitigating their contribution to climate change. As can be seen in figure 1.4, energy generation is the single largest contributor to global CO_2 levels. As such, increasing generation through clean energy sources has the potential to prevent substantial quantities of greenhouse gasses entering the atmosphere in the future.

However, although there has been considerable research, investment and development of renewable generation over the last 40 years (specifically in the last 10), these technologies cannot as of yet compete commercially against traditional sources of generation. Figure 1.5 depicts the levelised cost of energy (LCoE) estimates for new projects starting in 2013 as compiled by the DECC (Department of Energy and Climate Change, 2013). Traditional sources (such as coal) are not included as legislation prohibits the commissioning of such generation on environmental grounds. This figure shows that whilst nuclear energy is the cheapest carbon free source of energy (~£90/MWh) wind generation is not far behind (~£101-129/MWh). It should be noted that nuclear generation is often subsidised (indirectly) through agreed strike prices and as such, this is not an entirely accurate representation of the cost of nuclear energy. It is the governments intention that renewable energy should be able to compete with traditional sources, and has set a target for 2020 to reach of a LCoE of £100/MWh for offshore wind (Department of Energy and Climate Change, 2013; Estate, 2013). For completeness, table 1.1 shows the breakdown of costs across traditional and renewable sources as of 2010. These are further broken down in figure 1.6, and it can be observed that
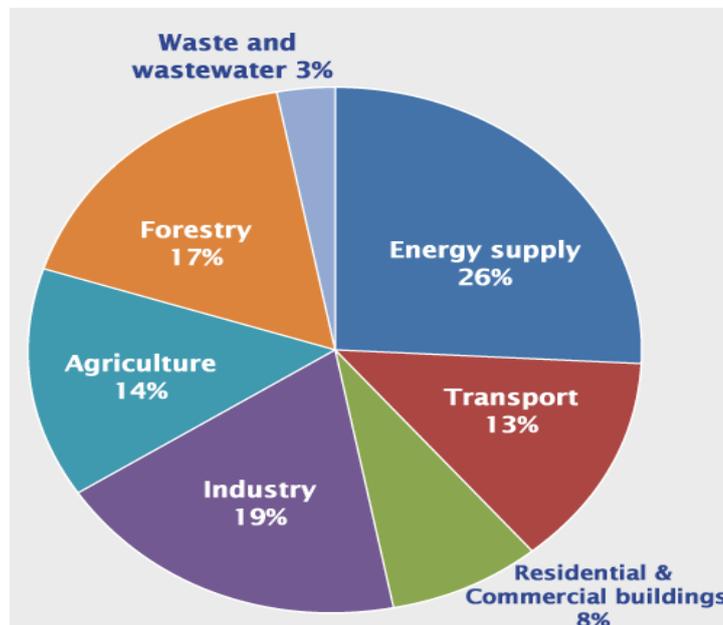
Figure 1.4: CO_2 Contribution by sector (Pachauri and Reisinger, 2007).

Table 1.1: Cost of energy in the UK (£/MWh), in 2010.

Technology	Cost (£/MWh)
Natural gas turbine, no CO_2 capture	55-110
Natural gas turbines with CO_2 capture	60-130
Biomass	60-120
New nuclear	80-105
Onshore wind	80-110
Coal with CO_2 capture	100-155
Solar farms	125-180
Offshore wind	150-210
Tidal power	155-390

wind generation (both onshore and offshore) has substantially higher relative O&M costs than traditional sources.

By increasing economies of scale, achieving a higher capacity factor through more sophisticated wind analysis and by the use of cutting edge O&M techniques to reduce both fixed and variable O&M costs, a reduction in the cost of wind energy to £100/MWh is believed to be readily attainable. DONG energy already believe this barrier to have been broken (Gosden, 2013).

Within the UK, the wind resource is substantially higher than the next leading low-carbon renewable source; solar (Department of Energy and Climate Change, 2013). The
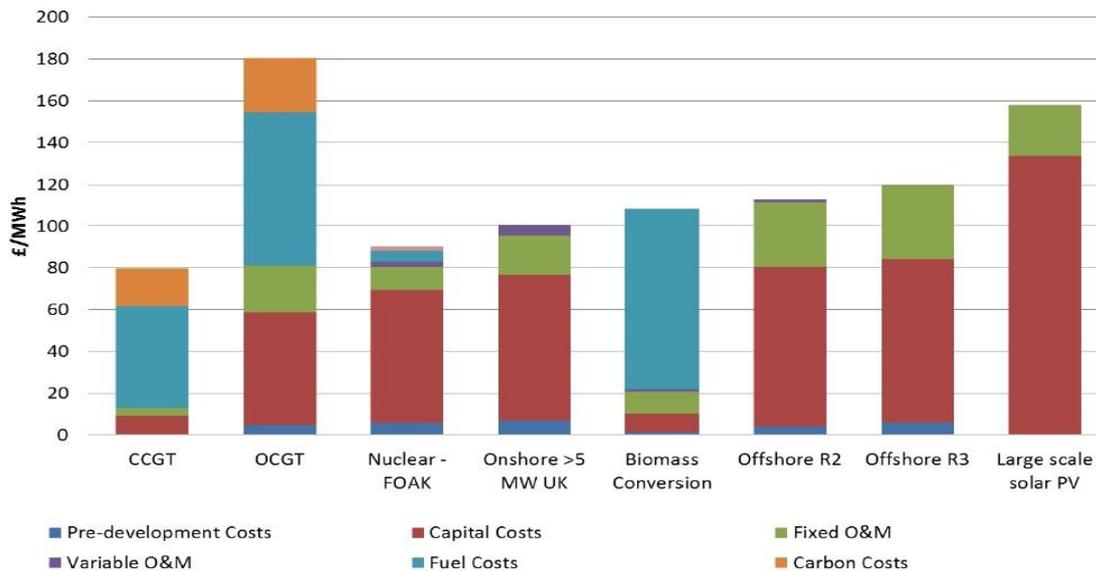
Figure 1.5: Energy cost by generation type (£/MWh) (Department of Energy and Climate Change, 2013).
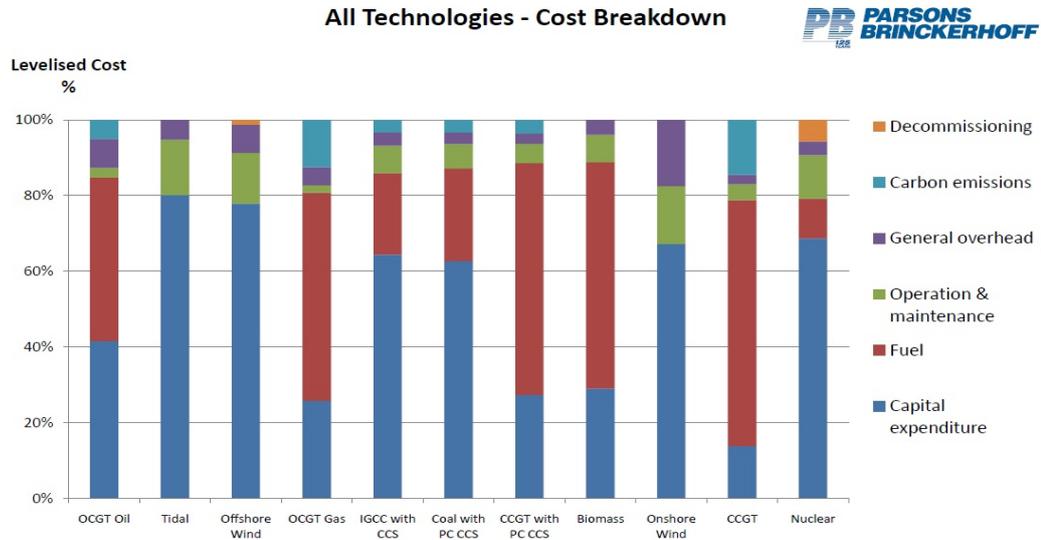
Figure 1.6: LCoE breakdown by generation source (£/MWh).

solar resource in the UK is limited, where the yearly sum of global horizontal irradiation measures approximately 700 to 1050 kWh per square metre (MacKay, 2008). This, coupled with the relatively low efficiency of current photo-voltaic technology – ∼8% for currently available silicon based technology (Gordon et al., 2007) – the cost of generation for solar energy is not as attractive commercially within the UK as in other countries. For instance, in the south of Spain, the yearly sum of global horizontal irradiation is approximately 1800 to 1900 kWh per square meter – over double that of the UK – double the generation can be achieved with the same investment costs.

The commercial attractiveness and the viability of each renewable resource can be implied through the installed capacity completed over the last 5 years. In the 2013 report by the DECC (Department of Energy and Climate Change, 2013), the growth of installed wind capacity can be seen to exceed that of the installed solar PV capacity. As solar PV technology is commonly available for residential generation, this provides confidence that wind is a more promising resource for future commercial renewable electricity generation within the UK. In 2013, approximately 2.7 GW of wind capacity was brought online, this is in comparison to an estimated 1 GW of PV installed over the same period (including residential installation) (RenewableUK, 2013).

1.4   Offshore wind

The rapid growth of wind energy within the UK is down to a variety of factors which include both government investment (including subsidies) (Department of Energy and Climate Change, 2013) and the UK wind resource (Crabtree, 2011). As can be seen in figure 1.7, the UK has access to substantial wind resources; both onshore, and offshore. As the distance from the UK increases into the Atlantic ocean, the average wind speed increases substantially. Onshore, the typical wind speed can be seen to be approximately 6 to 7 metres per second. Offshore, this rises to 9 to 11 metres per second (an increase of 50% to 57%). As the available energy within the wind is proportional to the cube of the wind speed (Patel, 2012), this increase equates to an increase of over 300% energy within the wind available for extraction. Due to this, offshore wind in the UK has attracted significant investment over the past 10 years, and is regarded as one of the best places in Europe for wind energy (Department of Energy and Climate
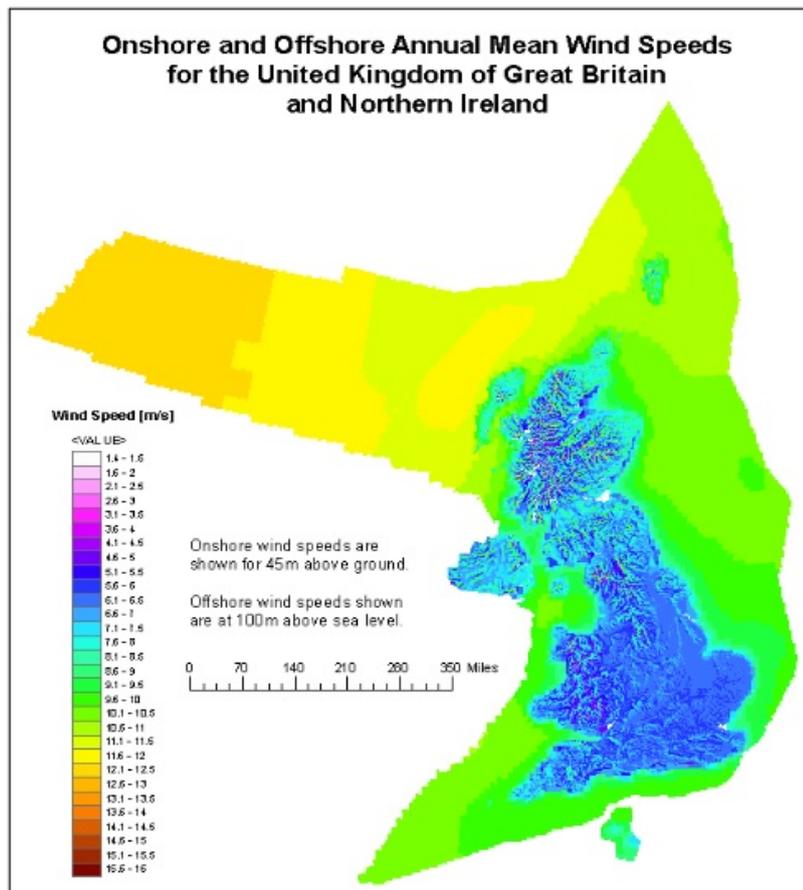
Figure 1.7: The UK wind resource (Crabtree, 2011).

Change, 2011).

Figure 1.8 presents the total installed capacity of wind generation in the UK from 1990 to 2013. As technology has matured and offshore costs have been reduced, the proportion of offshore capacity to onshore capacity has increased. This has been assisted in recent years due to the difficulty in gaining planning consent for onshore wind turbines; although public opinion supports wind energy and renewable sources of energy (European Wind Energy Association, 2003), new legislation has made it difficult to install capacity onshore (Smith, 2014).

The move offshore has enabled larger turbines to be designed and installed. Currently, 8 MW wind turbines have been built and are under testing for use offshore (Natarajan, 2014). This is possible due to the economies of scale which are achieved when working offshore. As the wind resource is much more stable and less turbulent, the associated risks and liabilities are reduced (Tavner, 2012). Due to the substantial costs involved when working offshore, installing a smaller number of larger turbines is substantially more cost effective. For instance, an 80 MW array consisting of ten 8 MW turbines has less installation costs (such as vessel hire, grid connection and foundation creation), less maintenance (servicing, and both corrective and preventive maintenance) and less spare inventory requirements than an array consisting of twenty 4 MW turbines. Similarly, due to the power generated by the wind turbine being determined by the swept area and also due to wake effects, turbines cannot be placed in
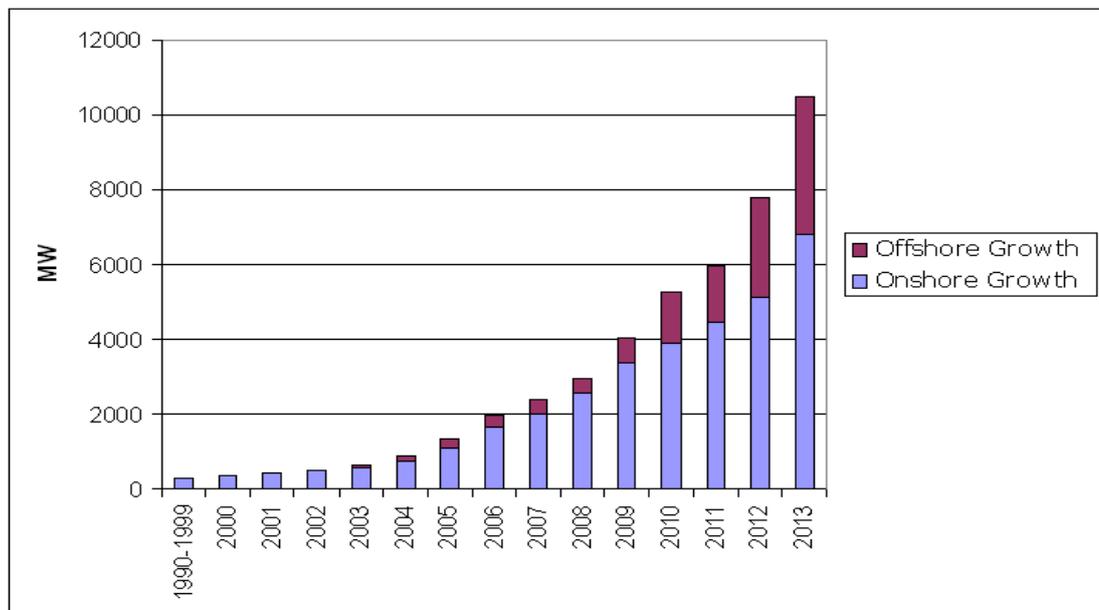
Figure 1.8: UK installed capacity (RenewableUK, 2014).

close proximity to one another. Typically, turbines will be spaced at 5 to 7 times their blade diameter apart. By employing larger turbines, smaller sites can be identified, and land rental costs can be reduced. As land rental can account for approximately 18% of O&M costs (European Wind Energy Association, 2005), a reduction in this expense can directly lead to a lower LCoE provided by the wind farm. Similarly, for the same area of land, larger generation capacity can potentially be installed, increasing the effective utilisation of this area.

However, moving larger wind turbines offshore into deeper water with intermittent access has caused various discrepancies between the reliability of wind turbines onshore and those offshore. Within the US, onshore availability is typically 95% to 98%, with offshore availability being reduced to 80% to 95% (Meadows, 2011). It should also be noted that O&M costs are approximately 21% of life cycle costs offshore, in comparison to approximately 11% of life cycle costs onshore. Due to the significant cost of wind turbines, it is essential when utilising smaller arrays of larger turbines to ensure high levels of availability are achieved.

1.5   Wind turbine reliability

Ensuring high availability for wind turbines is essential to increasing yield, reducing the LCoE and making wind energy a commercially viable and attractive alternative to fossil fuels. Figure 1.9 (Crabtree, 2011) presents the failure rates (denoted λ) and downtime (days per year) of two onshore wind farms, categorised by subsystem. This compromises over 20,000 turbine years of operation, over an operational period from 1993 to 2006.

As can be seen, the largest cause of downtime at a sub-assembly level is due to the gearbox. This represents over 14 days of downtime per failure. Although these figures are derived from onshore wind farms, due to the additional constraints imposed by offshore locations (such as intermittent accessibility) it is anticipated that downtime across all subsystems will increase when moving offshore.

Limited offshore reliability data is currently available in the public domain due to the commercially sensitive nature of the information. Research done by Dinwoodie et al. (2012) has identified two publicly available sources of offshore reliability data and performed similar analysis upon this data. Figure 1.10 presents the results of their analysis. This confirms anecdotal evidence that mean time to repair (MTTR) is higher offshore, and also that λ is higher offshore: in this case, twice as high. Their work presents a MTTR of 80 days; higher than the onshore analysis and also higher than the 40 days suggested in Tavner (2012). These figures will vary considerably
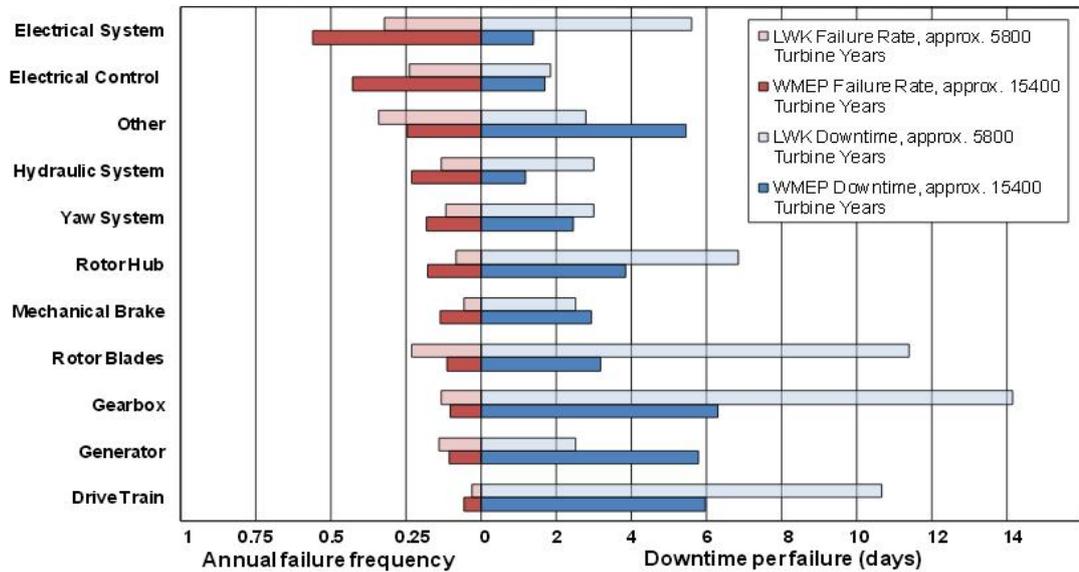
Figure 1.9: Wind turbine reliability by subsystem (onshore) (Crabtree, 2011).
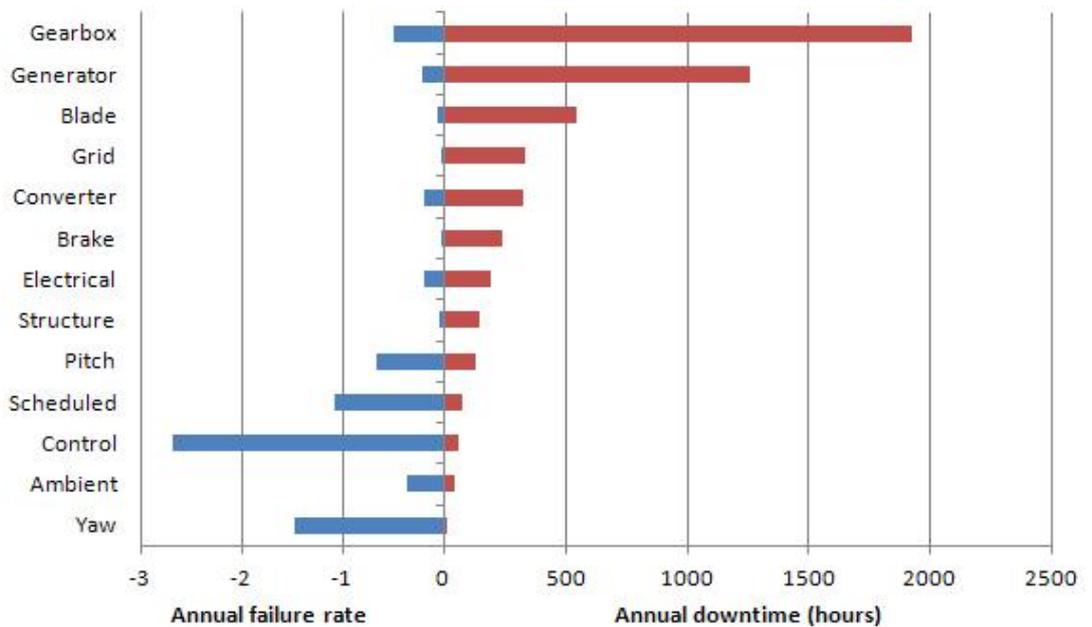
Figure 1.10: Wind turbine reliability by subsystem (offshore) (Dinwoodie et al., 2012).

dependent upon wind turbine size and location, the operational characteristics and various sub-assembly configurations (for instance, hydraulic pitch systems will have a different failure rate to electronic pitch systems).

The gearbox has the highest consequence of failure (probability of failure – λ –

multiplied by expected downtime), under traditional reliability analysis such as failure methods and effects criticality analysis (FMECA) (Dinwoodie et al., 2012). This would be regarded as the highest risk priority due to both the relatively high probability of failure (6th of 13 subsystems) and the largest cost of failure. As such, understanding the failure mechanisms associated with the gearbox is essential to increasing the availability of the wind turbine and thus reducing the LCoE.

1.5.1 Wind turbine gearbox reliability

As previously stated, the gearbox is responsible for the largest quantity of downtime both onshore and offshore. As such, understanding the failures modes of the wind turbine gearbox is essential to reducing downtime and increasing gearbox reliability. Presented in figure 1.11 is a breakdown of failure modes as compiled by the National Renewable Energy Laboratories (NREL) compiled by Sheng (2014). Their database contains 257 recorded instances of damage to wind turbine gearboxes. Of these failures, 180 (70%) were due to bearing failures and 68 (26%) were due to gear failures. The remaining 9 failures (4%) refer to internal shaft, the high speed shaft coupling and housing failures.

Of the bearing failures, one can identify the high speed shaft (HSS) bearing as being particularly susceptible to failure. Of the 257 failures recorded, 123 were due to the HSS bearing. This represents 48% of the failures, and as such, a substantial proportion
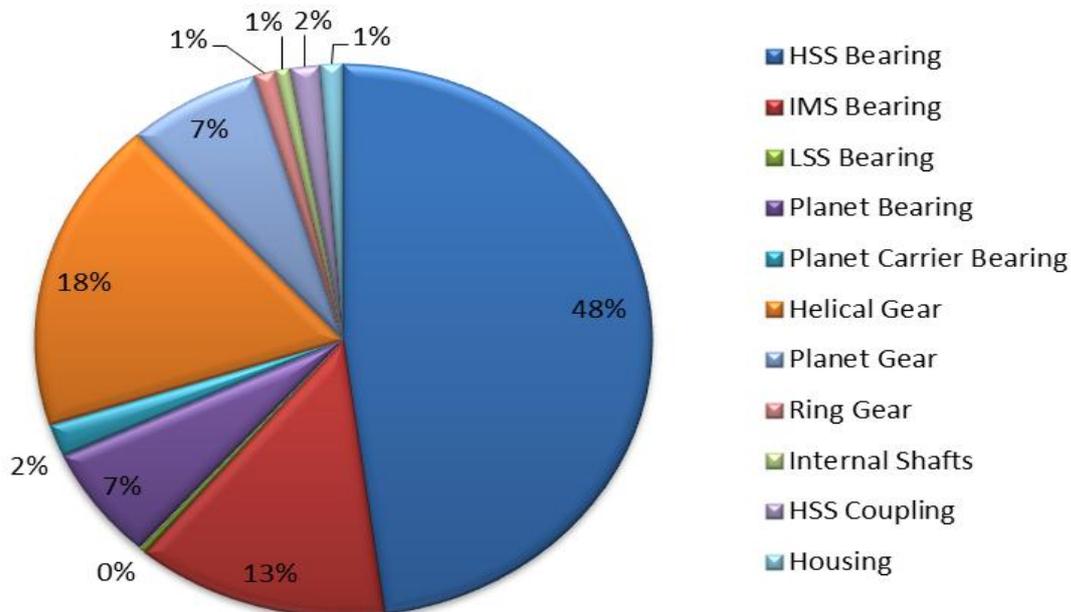
Figure 1.11: Wind turbine gearbox reliability by failure mode (Sheng, 2014).

of the failures which are likely to occur within the wind turbine gearbox. If a wind farm owner or operator was to receive advance warning of potential degradation or damage to the HSS bearing, inspection of the gearbox could be performed to confirm the early detection, perform diagnosis and have the component repaired or replaced at a reduced cost.

1.6   Challenges in operations and maintenance

Various challenges are presented when wind farms are moved offshore. Within the context of this thesis, two main issues exist.

Firstly, the issue of physical access to the wind turbine to perform maintenance. Due to the harsher weather conditions offshore, access to wind turbines is typically intermittent and heavily dependent upon the local weather conditions at the wind farm. As such, during winter months (and to some extent, the summer), limited opportunities to perform either corrective or preventive maintenance exist. Although this is well known within industry, down time is increased as failures cannot be as easily rectified as they can be onshore. It is estimated that 30% to 50% of downtime offshore is due to poor weather conditions (Meadows, 2011). Similarly, additional costs arise in the purchase or hire of vessels (typically a boat or helicopter) to transport maintenance personnel and components to the wind turbine.

Secondly, the issue of the exploitation of condition monitoring data for the purposes of remote diagnosis and prognosis. As physical access to the wind turbine is limited, it becomes essential to be able to perform remote fault identification, diagnosis and eventual prognosis through the use of data analytical techniques. It is often impractical (both financially and logistically) to perform inspections for all potential wind turbine faults which are identified through the on-board SCADA systems, and as such, the effective identification of potential faults and the prioritisation of maintenance actions are critical to ensuring high availability. However, this data often quickly becomes unmanageable and ineffective to exploit due to the overwhelming quantity of data recorded.

Both of these challenges can be addressed and their influence mitigated through the use of start of the art techniques employing data taken directly from the wind turbine. Non-linear techniques can be used to accurately forecast the likelihood of turbine accessibility to ensure that maintenance can be performed based upon wind farm environmental condition data (Wu and Hong, 2007), with similar techniques existing to autonomously predict the degradation and remaining useful life (RUL) of components (Jardine et al., 2006). As such, there is a strong synergistic link between the

exploitation of condition monitoring data, the availability of wind turbines and thus the LCoE.

1.7   Benefits of condition monitoring

Condition based maintenance (CBM) systems are referred to as the "holy-grail" of maintenance techniques due to their ability to increase availability, increase yield and simultaneously reduce maintenance expenditure (Coble and Hines, 2008). This is because condition monitoring systems provide a means to move away from traditional 'fail and fix' maintenance methodologies towards 'predict and prevent' maintenance strategies (Levrat et al., 2008).

It is widely agreed that preventive (proactive) maintenance actions are significantly and substantially more cost effective than corrective (reactive) maintenance action (Wu and Clements-Croome, 2005; Association, 2011). Hameed et al. (2009) identified five key motivators for condition monitoring systems with regards to wind turbines. These are:

- *Avoidance of premature breakdown:* The most important aspect of incipient fault detection is to prevent catastrophic failures and secondary defects. For example, late detection of a rotor-bearing fault may in the worst case imply complete destruction of the wind turbine.

- *Reduction of maintenance costs:* With on-line monitoring, inspection intervals can be safely increased. Replacement of intact parts is avoided by condition-based maintenance.

- *Supervision at remote sites, remote diagnosis:* Large wind turbines are usually built at remote sites where, in contrast to industrial machinery, optical or acoustical changes may not be observed for a quite long time. On-line monitoring systems can detect such changes at an early stage and, if equipped with a modem, send a warning and diagnostic details to the maintenance staff.

- *Improvement of the capacity factor:* With early warning of impending failures, repair actions need not be taken immediately. They can be carried out in a period of low wind speed when the wind turbine is off-line, without affecting capacity factor.

- *Support for further development of a wind turbine:* A condition monitoring system yields detailed information on the dynamic behaviour of a wind turbine over long periods of time that may help optimizing future wind turbine design.

Recent literature has begun to quantify the benefits of moving away from traditional maintenance techniques towards proactive strategies. The period for return on investment of condition monitoring systems and associated benefit is a relatively recent debate which is beyond the scope of this thesis, however, there are many documented cases within the literature of condition monitoring systems effectively reducing costs and being insisted upon for warranty purposes. For instance:

- In May et al. (2014), a vibration condition monitoring system is shown to be able to achieve an estimated £4.2 million in savings over traditional preventive maintenance alone (over the entire turbine life cycle).

- In May et al. (2014), a completed condition monitoing and structural health monitoring system for the wind turbine foundations is shown to be able to achieve an estimated £7.1 million in savings over traditional preventive maintenance alone (over the entire turbine life cycle).

- SKF's WindCon condition monitoring system has supposedly saved operators over a million pounds due to early detection of failure (Winny, 2013).

- Similarly, SKF's WindCon system saved £57k due to the early detection of a cracked gear tooth within the gearbox (Winny, 2013).

- A saving of over $300,000 (£200,140; correct as of 25th January 2015) could be achieved with a condition monitoring as per Rehmet (2006).

- GL-Garrad Hassan estimate that condition monitoring systems may save up to $315,000 (£210,147; correct as of 25th January 2015) per gearbox failure if damage can be isolated to a single state (LeBlanc and Graves, 2011).

- Cost savings of up to 44% are shown for wind turbines using condition based maintenance in Tian et al. (2011).

1.8 Original contributions of the thesis

This thesis extends the state of the art in various ways. Of the contributions, 6 major contributions, and 2 minor contributions are identified. The major contributions of this thesis have been peer reviewed and published in either internationally recognised conference proceedings or within internationally recognised journals as articles. The minor contributions of this thesis have been published as a peer reviewed book chapter. The major contributions of this thesis are:

1. A bearing condition model based upon statistical time domain feature selection (based on robust, EWMA statistical process control) for use in robust multivariate statistical analytical techniques. **Section 4.2.11**.

2. An extended physics of failure model for the analysis of wind turbine gearboxes, taking turbine loading, transient states and both curtailment and de-rating into consideration. **Section 4.3**.

3. An elitist memetic algorithm for the identification of normal operational asset behaviour based upon historical data with potentially quicker convergence than traditional heuristics methods. **Section 5.3**.

4. A general methodology for the automated labeling of historical data to enable novel artefact extraction in the multivariate domain and to enable traditional data mining techniques where labels are not necessarily available (as applied in this thesis to wind turbine gearbox rule extraction). **Section 6.2**.

5. A general methodology for fault diagnosis based upon robust multivariate statistical techniques (as applied in this thesis to a bearing). **Section 6.3**.

6. A general methodology to predict the remaining useful life of components (as applied in this thesis to a bearing) through exploiting robust multivariate statistical techniques and artificial neural networks. **Section 6.4**.

Furthermore, the various minor contributions to the state of the art of this thesis include:

7. An extensive analysis of time domain features for the use in bearing fault detection, diagnosis and prognosis in the multivariate domain (as explored in this thesis in relation to bearings). **Section 4.2**.

8. An extensive analysis of multivariate distance metrics for use in assisting with the early detection, diagnosis and prognosis of components (as explored in this thesis in relation to bearings). **Section 5.2**.

1.9   Structure of the thesis

This thesis is structured in a series of chapters, compromising the research undertaken by the author at the University of Durham from April 2011 until November 2014. Each chapter contains a coherent piece of research which builds and follows upon the work undertaken in the former chapter.

In this chapter, we have introduced the context for the research. The effects of climate change and the necessity to reduce the cost of energy in order to move to a low carbon economy have been discussed. Wind energy was shown to be a viable candidate for the UK in particular to enable this. Following this, both the historical and recent offshore developments of wind energy in the UK and ensuing issues of reliability were presented with the mitigation that state-of-the-art maintenance techniques techniques (namely condition based maintenance) can provide. Finally, the major and minor contributions of this thesis were detailed, and the structure of the thesis presented.

In chapter 2, the relevant literature is discussed within the context of wind energy. Both high frequency and low frequency techniques are explored for both bearings and gearboxes across the time, frequency and time-frequency domains. The concept of "big data" is introduced, and the application of these techniques within both the wind energy and the prognostics and health management (PHM) literature are presented.

Chapter 3 details the research undertaken in this thesis. The overall thesis question, the three research objectives, and the six research questions this thesis aims to answer are given. The methodologies utilised for their exploration are presented along with the descriptions of the three distinct datasets employed in this thesis.

Chapter 4 introduces the novel models developed in this thesis for the condition assessment of bearings and gearboxes. A variety of time-domain statistical signatures are explored for their potential in degradation assessment, with feature selection then being performed. A full analysis of these features, and their ability to detect degradation on the bearing is performed, with normalisation against internal and external factors detailed.

In chapter 5, we build upon the work in chapter 4 to extend the analysis into the multivariate domain. Multivariate distance metrics for condition assessment are explored in detail, with a focus on robust techniques which can be employed in practice due to the practical constraints of real-world data. Following this, a novel methodology employing AI search techniques for the selection of a robust subset to calculate attribute covariance is detailed.

Chapter 6 exploits the derived condition index for the purposes of diagnosis, prognosis and knowledge codification. Methodologies are presented to label data based upon their potential representation of degradation, to identify fault types based upon ensembles of metrics and also to determine the remaining useful life of components.

Finally, Chapter 7 provides the conclusions of the thesis, discusses both the advantages and the limitations of the research, and provides insight into future work which is a natural extension to this thesis.

Chapter 2

Techniques Enabling Fault Detection, Diagnosis and Prognosis

In this chapter, the body of literature relevant to this thesis is presented and discussed. Topics including wind turbine condition monitoring, bearing and gearbox early detection, diagnosis and prognosis with machine learning (including decision support) finally considered. The state of the art is discussed, with the advantages of each approach given. Current challenges and limitations are analysed and explored with the context of this thesis.
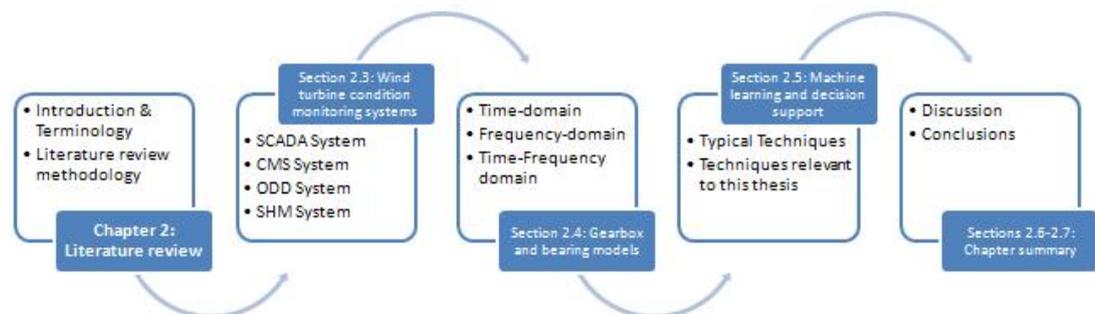
Figure 2.1: Overview of the structure of chapter 2.

2.1   Introduction & terminology

In order to understand the current state of the art in early detection, diagnosis and prognosis, it is essential that an extensive review of the current literature is performed. For the purposes of this thesis, publicly available grey literature was included, in addition to peer-reviewed research from internationally recognised conference proceedings and journal articles. This review will assist in understanding the challenges faced moving forward with regards to the early detection, diagnosis and prognosis of both the gearbox and bearings within a wind turbine. For the purposes of this thesis, "Danish concept" wind turbines are considered; three-bladed variable pitch horizontal wind turbines (Hansen and Hansen, 2007).

Due to the ambiguous nature of many terms in this area, and for clarity for the purposes of this thesis, early detection, diagnosis and prognosis are defined as follows:

- **_Fault Detection:_** The detection (and potential quantification) of degradation on a specified component through the analysis of high (or low) frequency data – typically through automated techniques – with the aim to provide actionable intelligence into an enterprise for the purposes of maintenance, prior to what would ordinarily be attained through traditional maintenance practices.

- **_Diagnosis:_** The identification of the most likely current (or expected) failure mode of a specified component, prior to, during or after a failure, through automated techniques, for the purposes of maintenance.

- **_Prognosis:_** A forecast (or prediction) of the degradation of a specified component – in terms of remaining useful life, survival probability or predicted future condition – at a future point in time, allowing for the subsequent exploitation of this knowledge for the purposes of maintenance.

Prognosis can refer to a forecast (a probabilistic statement, typically over a longer time period) or a prediction (a specific statement, definitive in nature). The context of which will be determined by the application. For clarity, where otherwise stated, this thesis will refer to prognosis as a prediction, rather than as a forecast.

Prognosis can be further broken down into three distinct categories as defined by Coble and Hines (2008) depending upon the available historical and operational data:

- **_Reliability data based (Type I):_** In type I prognosis, historical reliability data is employed for the analysis. This data is utilised to determine average life of a component operating under historically average conditions. This is a population based prognostic approach, and is often performed in traditional reliability analysis, such as Weibull analysis Leemis (1995).

- ***Stress-based (Type II):*** Type II techniques complement type I techniques by incorporating environmental stresses (vibrations, temperature, load etc.) and as such, are employed to estimate the average life of a population of components given the operational conditions. This leads to a more accurate and precise life estimation. The proportional hazards model (Breslow, 1975) is an example of a type II prognostic.

- ***Effects-based (Type III):*** A type III prognostic specifically considers individual components, and how these individual components reacts to specific usage. As this is component specific, type III prognostics have the possibility to provide the most accurate and precise life estimation for a given component. The general path model (GPM) is an example of a type III prognostic (Coble and Hines, 2009).

This chapter is structured in a manner akin to that of a condition based maintenance program, which typically compromises three distinct parts as discussed in Jardine et al. (2006). These are referred to as data acquisition (what data is collected and available for analysis), data processing (what can be done with the data, such as feature extraction and principle component selection) and maintenance decision making (for instance, to replace, repair or service an asset or component). This process is shown graphically in figure 2.2.

Initially, the methodology of this review is presented in section 2.2. Next, this review provides context to the problem by detailing the availability of data on wind turbines through the individual sub-systems. This is detailed in section 2.3. Following this, an extensive review of the available techniques to process the acquired data for the purposes of early detection, diagnosis and prognosis is undertaken in section 2.4. Finally, relevant machine learning and decision support techniques to enable maintenance and enhance capabilities are detailed and critically discussed in section 2.5.

Figure 2.2: Condition based maintenance (CBM) process.

2.2 Literature search methodology

In order to survey the current practices, techniques and state of the art, a systematic literature review was conducted. This effectively enabled the identification of relevant literature for inclusion into the thesis, whilst providing a means to exclude irrelevant literature.

For the purposes of this thesis, publicly available grey literature was included. This refers to academic work which has not been formally published in commercial publications. This includes work such as pre-prints, technical reports (for example, from government think-tanks, or research groups), patents and white papers. Due to the commercial benefits which can potentially be attained from PHM systems, many of these documents are not widely circulated within the public or academic domains. It has been estimated that more grey literature exists within various engineering sciences than any other subject area; approximated at 39% to 42% (Schöpfel, 2008).

A search of the literature was performed on all articles published between January 1980 and July 2014 was performed. Initial searches were performed through 10 on-line repositories, databases and websites. Specifically utilising: Science direct, IEEE Xplore, Wiley online, ACM library, Taylor & Francis, google scholar, ISI web of knowledge, JSTOR, ERIC and EThOS. In addition to this, manual screening of the most relevant publications in this area was performed in order to assist in the identification of further literature which was not identified in the preliminary search. Specifically, the proceedings and volumes of the following publications were explored: Mechanical Systems & Signal Processing (MSSP), Reliability Engineering & Systems Safety, IEEE transactions on reliability, the proceedings of all conferences sponsored by the IEEE reliability society (including RAMS and the IEEE PHM conference), COMADEM, all publications produced by the PHM Society, conferences hosted by the EWEA, AWEA or the GWEA, the proceedings of the IEEE Aerospace conference and conferences sponsored or hosted by the MFPT.

Identification of keywords was performed, utilising "AND" and "OR" operators to assist in refining the search criterion. The search criteria was: (Bearing OR Bearings OR Roller element OR ball OR gearbox OR gearboxes OR planetary OR gear OR stage) AND (signal OR signals OR diagnosis OR diagnostics OR prognostics OR prognosis OR RUL OR remaining useful life OR prediction OR health OR data mining OR machine learning OR condition OR failure OR estimation OR fault OR monitoring OR SCADA OR feature OR extraction OR optimisation OR exploiting OR early OR detection OR assessment) AND (wind OR turbine OR offshore OR onshore OR energy).

2.3 Wind turbine condition monitoring systems

Initially, to identify both the opportunities which exist and challenges which are faced when performing early detection, diagnosis and prognosis through the condition monitoring systems on a wind turbine, we must first identify the available systems which collect and collate data for analysis. Monitoring of the wind turbine is essential for various reasons. Typically, wind farm owners will wish to utilise the data collected to ensure that their turbines are working correctly and generating income, operators will wish to exploit the data to reduce maintenance costs and original equipment manufacturers (OEMs) will typically use the data for warranty purposes (Crabtree, 2011). As such, various different condition monitoring systems exist on a wind turbine.

Figure 2.3 identifies the four main data acquisition and analysis systems typically in place on a wind turbine. As can be seen, different systems will collect data at different frequencies dependent upon their intended use. These are described in detail below. For context, the supervisory control and data acquisition (SCADA) system is discussed in section 2.3.1, the high frequency condition monitoring system (HFCMS) in section 2.3.2, the on-demand diagnostics (ODD) system in section 2.3.3 and the structural health monitoring (SHM) system in section 2.3.4.
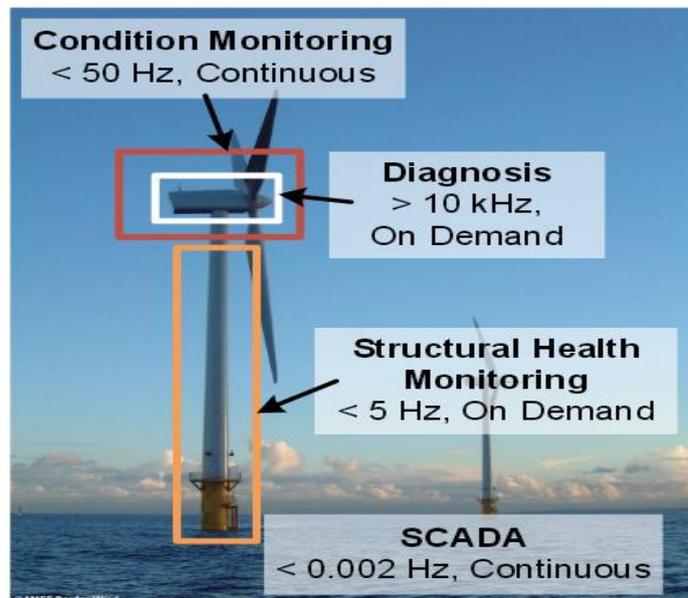
Figure 2.3: Condition monitoring systems on a wind turbine (Crabtree, 2011).

2.3.1 Supervisory control and data acquisition (SCADA) system

The supervisory control and data acquisition (SCADA) is the most common condition monitoring system in place on a wind turbine. Data is collected at a low frequency. Typically this is at one sample per 5 minutes or per 10 minutes – as in Godwin and Matthews (2014b) – although systems employing SCADA systems at 1 sample per 10 seconds are mentioned within the literature (Kusiak and Verma, 2012a).

SCADA data typically contains data pertaining to the general operation of the turbine, and often collates and integrates signals from various controllers so that reporting can be done holistically from one system coherently. As various sub-assemblies are connected to the SCADA system, many different signals are collected. For instance, Sainz et al. (2009) reports that meteorological information (such as wind speed, wind direction, humidity and ambient temperature) are collected along with turbine operating information (such as power generated and rotor speed) and various sub-assembly information (such as gearbox temperature, pitch angle, yaw angle, etc.).

Due to various different sample rates across the various collated subsystems, different forms of averaging and aggregation are performed by the SCADA system. This enables various statistics of highly variable information – such as maximum, minimum and average wind speeds – to be reported over a specified time period (i.e. 10 minutes) along with temperature data. Whilst this does enable various forms of analysis, due to this sampling and aggregation, information in inherently lost. An example subset of SCADA records as attained from a wind turbine are presented in table 2.1.

Current wind turbine SCADA systems employ approximately 200 channels of data to provide a "snapshot" of current operational behaviour. It is anticipated that next generation wind turbine systems may potentially employ as many as 800 channels for the same purpose. This is of interest as currently only single measurements are typically collected from each specific point of interest (i.e. gearbox temperature). It may be the case that various levels of redundancy within these systems becomes the de-facto

Table 2.1: Example of SCADA records from a wind turbine.

Time	Wind speed (m/s)	Rotor speed (RPM)	Power (kW)	Ambient temperature	Oil temperature (Celsius)
18/5/11 14:20	8.92	12.65	113.05	15.44	46.44
18/5/11 14:30	10.76	13.9	147.05	14.03	46.81
18/5/11 14:40	7.21	11.1	69.8	15.47	47.19
18/5/11 14:50	7.44	11.6	82.2	19.68	46.51
18/5/11 15:00	6.55	10.25	81.85	17.58	45.63
18/5/11 15:10	11.07	13.35	127	16.97	46.32

standard should it be deemed commercially viable. Similarly, if multiple measurements are taken from the same component (i.e. 3 gear temperatures from various gears within the gearbox), this can be exploited in ways which are yet to be explored fully, such as through multivariate techniques (Rencher, 2003).

It should be noted however, that SCADA systems are not perfect. Within the literature, there are consistent reports of various issues which have occurred. These have been explored by Sainz et al. (2009), which identified the following issues which occur within SCADA systems:

- **Missing data:** The SCADA system fails to collect sensor readings, resulting in no sample being recorded (such as no ambient temperature being recorded).

- **Implausible data:** The SCADA system incorrectly reads the sensor reading, resulting in impossible values being recorded (such as ambient temperatures below absolute zero, or power production being orders of magnitude higher than the turbine is capable of).

- **Duplicate data:** The SCADA system duplicates the previous sensor reading, resulting in continuous acquisition of the same value (such as ambient temperatures not changing during day and night cycles).

- **Erroneous data:** The SCADA system incorrectly reads the sensor reading, resulting in plausible – but incorrect – data being recorded (such as ambient temperatures being incorrect by 3 to 4 degrees Celsius).

As such, there is a distinct need to validate the readings from the SCADA system. Due to the large sensor arrays employed, there is also a need to utilise both statistical and also data-intensive techniques for the analysis of this data. Work within the literature has shown that SCADA data is potentially useful in maintenance optimisation as the SCADA systems cover all main turbine sub-assemblies as detailed in figure 1.9 (Spinato et al., 2009; Crabtree, 2011).

Along with the ability to transmit SCADA data at predefined intervals, the SCADA system also has the capability to generate alarms based upon the data received. These SCADA alarms are often based upon static thresholds (Qiu et al., 2012) and are often naïve to the general health of the wind turbine. For instance, an alarm may be triggered when the gearbox temperature reaches 45 degrees Celsius, causing the turbine to shut down. However, if this is due to a relatively high ambient temperature, rather than damage to the gearbox (which was the intention of the rule), this would be the cause of unnecessary down-time and result in a loss of production and revenue. Within the

literature, their are numerous stories regarding the "information overload" caused by quantity of SCADA alarms presented to maintenance operators on a hourly, daily or weekly basis (Chen et al., 2011a; Qiu et al., 2012; Godwin and Matthews, 2014b).

2.3.2   High frequency condition monitoring system

The second data acquisition system on the wind turbine moves away from the low frequency SCADA data. The high frequency condition monitoring system employed on wind turbines – as the name suggests – collects data at a much higher frequency than that of the SCADA system. As such, this system is employed in situations when SCADA data cannot provide the granularity of analysis required. They are often utilised in conjunction with the on-demand high frequency diagnostic system to enable effective decision making to take place. Two distinct high frequency systems exist; first, condition monitoring systems which typically record data at 50 Hz, and secondly on-demand diagnostics which are available at up to 50 kHz.

Initially, high frequency systems were encouraged by insurers of large wind turbines to reduce the liability of failure by providing early detection of failures which could be remedied at a reduced cost (Becker and Poste, 2009). Various insurers of wind turbines now have certification schemes, with many condition monitoring systems available commercially which are certified by these insurers. For instance, "WindCon 3.0" (developed by SKF) is GL certified by insurer Allianz. Reports by Allianz have shown that even a minimal condition monitoring system (CMS) applied to a wind turbine had assisted in risk management (Gellermann, 2013).

Although the high frequency system has a higher data acquisition frequency than the SCADA system, it is less than that of the on-demand diagnostics system. This enables data at multiple levels of granularity to be assessed. This is essential on critical assets such as wind turbines, where the management of the asset is performed remotely. By having data available at various frequencies, many techniques can be employed in order to validate and verify potential degradation and fault modes. Typically, the SCADA system will alert the operator that a potential issue has arisen, this will then be explored through the use of the high frequency condition monitoring system (HFCMS). If the potential fault cannot be proven or discredited, the on-demand diagnostics system is utilised.

2.3.3   On demand diagnostic system

The third data acquisition system on the wind turbine is that of the on-demand diagnostics system. Due to the high frequencies employed by the diagnostic system (up

to 50 kHz), this data is available on demand and not stored offline for retrospective analysis. This is due to the high bandwidth and storage requirements imposed by the frequency of the data. The relatively lower frequency condition monitoring data (50 Hz) typically is collected and stored for analysis. This poses many challenges. As this data is collected on a sub-second basis, this data quickly becomes unmanageable and difficult to process. One hour of SCADA data will contain 6 datapoints, whereas over the same time period, 180,000 data points of 50 Hz condition monitoring data will have been recorded. This makes identifying the development and propagation of faults difficult due to the computational complexity of the various algorithms employed for analysis. For comparison, over the same time period 180,000,000 data points of 5 kHz diagnostic data would have been recorded. A high frequency wind turbine condition monitoring system recording data at 50 Hz requires approximately 2 TB of storage per month (Ferguson and Catterson, 2014). As such, performing analysis simultaneously across an entire farm of turbines requires substantial quantities of bandwidth, computationally efficient algorithms and both a reliable and actionable end result. A typical example of the high frequency data collected from roller element bearing is presented in figure 2.4.

The high frequency nature of the data acquisition on both the high frequency condition monitoring system and the diagnostic system enables analysis which is not attainable through SCADA data. For instance, if accelerometer or acoustic emission data is available for bearing analysis, using this data it would be possible to perform diagnosis of specific fault types, and identify both the location and severity of the fault using the various techniques discussed in sections 2.4.3 and 2.4.4.

There are many interactions between the three distinctly separate systems which
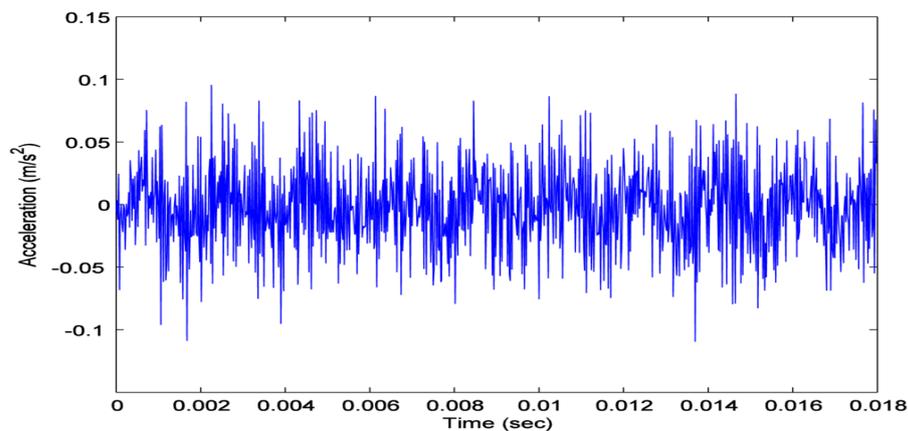
Figure 2.4: Typical high frequency roller element bearing data (Kwak et al., 2014).

have already been presented. The SCADA system will continually monitor the wind turbine and present alarms to the maintenance operator. If this alarm is confirmed by the high frequency condition monitoring system, it establishes a basis to perform analysis through the on-demand diagnostics to either confirm or reject the potential issue. This persistence checking and layering of the various frequencies of data acquisition helps reduce the quantity of data presented to the operator, whilst providing them with the level of detail required in order to to perform their analysis. It should be noted that due to both the bandwidth and infrastructure requirements of the high frequency systems, both the diagnostic and high frequency condition monitoring systems require a substantial investment. According to Yang and Jiang (2011), this is considerably more than the low frequency SCADA systems.

2.3.4 Structural health monitoring

The fourth wind turbine monitoring system is that of the structural health monitoring system. Offshore wind turbines, due to their size, typically employ condition monitoring techniques to ensure the structural integrity of the wind turbine. Most commonly, the structural health monitoring system is utilised to assist in determining faults with the wind turbine tower structure or foundation, such as those caused by wind gusts and wave slam (Chen, 2014).

Due to the slow development and propagation of these faults – Paris' law can be used to determine and predict fracture size growth (Paris and Erdogan, 1963) – continuous monitoring is not required, and as such, typically these systems are available on-demand with data not stored for retrospective analysis. Data is typically acquired at 5 Hz (Crabtree, 2011).

Structural health monitoring systems can be utilised to provide condition monitoring of the blades of the wind turbine through various techniques. Most commonly, fibre optics are employed to determine the quantity of light which is sent and received along each blade (Takeda, 2002). When utilising plastic fibre optics, the quantity of light received is proportional to the strain on the plastic fibre. As such, damage to any blade with this technology attached will exhibit different responses due to a differing load profile (Takeda, 2002). Similar techniques exist utilising acoustic emissions (AE) sensors; for a review of applicable techniques for structural health monitoring, please refer to Ciang et al. (2008). These techniques are typically performed on-demand due to their sensitivity to operational behaviour.

2.3.5 Commercially available systems and other techniques

Due to the benefits which can be attained when utilising data collected condition monitoring equipment, various commercially available solutions exist to provide this functionality. A recent survey conducted on behalf of the SuperGen Wind consortium (Wilkinson et al., 2010) performed an extensive review of commercially available high frequency CMS systems and found by Crabtree et al. (2010) that:

- One system exists for condition monitoring of wind turbine blades based upon vibration analysis.

- Two systems exist for fibre optic strain measurement of wind turbine blades.

- Three systems exist for oil debris monitoring.

- Fourteen systems exist for drive train monitoring based upon vibration analysis.

The authors then specify that the majority (14 of 20) systems are based upon traditional techniques taken from rotating machinery industries. Furthermore, the authors conclude that the commercial products are beginning to adapt to the demands of the wind market – which typically generate non-stationary signals due to their variable speed – and are moving away from products used in traditional rotating machinery industries. However, it is also noted that although traditional techniques can assist in the detection of faults, it takes experienced condition monitoring engineers for successful analysis.

Whilst the survey undertaken by Crabtree et al. (2010) provides context for the condition monitoring systems in place on a wind turbine, the SCADA systems used on wind turbines were beyond the scope of their work. A recent survey undertaken by Chen et al. (2014) provides an extensive list of commercially available SCADA systems for use on wind turbines. In the report by Chen et al. (2014), it was found that:

- One product was developed by a wind turbine operating company.

- One product is a demonstration platform developed by IMS.

- Three products were developed by electrical equipment providers.

- Five products were developed through renewable energy consultancies.

- Six of the products were developed by wind turbine manufacturers.

- Ten products were developed by industrial software companies.

The authors note that the techniques employed varied across different platforms, but covered both statistical and artificially intelligent techniques, with some platforms being able to perform component level diagnosis. Although the report details success of SCADA systems, little information is provided regarding the assessment or prognosis of wind turbine health through SCADA data analysis.

2.3.6   Summary and conclusions

The four data acquisition systems present on a wind turbine enable the analysis of substantial quantities of data at varying levels of granularity. All critical elements of the wind turbine are monitored (typically for insurance purposes) ensuring high availability is achieved. It should be noted that traditionally in wind energy, many of the data analytical techniques are taken from the domain of traditional rotating machinery (such as traditional gas turbines, for instance). Vibration analysis is the most common technique employed for drive train condition monitoring, compromising almost 75% of the commercially available systems.

As vibration analysis is heavily employed within industry, SCADA data is often overlooked (potentially due to its low frequency acquisition rate). Given that the SCADA system is the most commonly available wind turbine data acquisition system, the development of techniques which employed SCADA data which are as powerful as those based upon vibration analysis would assist in mitigating the necessity for high frequency data. As SCADA data has a lower acquisition rate, the potential for forecasting is also of interest due to the possibility of forecasting further ahead than with high frequency condition monitoring data.

2.4   Gearbox and bearing condition assessment

2.4.1   Introduction

As found in section 2.3, various low and high frequency data acquisition systems are typically present on a wind turbine, monitoring the bearings and the gearbox. This data can be explored and processed in order to understand the health of the wind turbine so that more accurate decisions can be made regarding the operation of the turbine. In accordance with figure 2.2, this section discusses the data processing techniques typically employed for condition assessment utilising high frequency (over 5 kHz) data. For completeness, various techniques which have shown promise on low-frequency data – specifically with regards to SCADA data – are included here, and discussed separately in section 2.4.5.

Historically, due to the development of many techniques utilised for the condition assessment of wind turbines being employed from traditional rotating electrical machinery, there is a substantial body of work in this domain. As such, many of the presented techniques are well suited to both bearings and gears. Furthermore, due to generalisability of many of these techniques, a substantial body of work is applicable to provide early detection, diagnosis and prognosis for both bearings and gearboxes. As such, techniques which are have been proven within the literature to be capable of being utilising in both applications are presented together for consistency.

Although data processing for condition assessment can take various forms, four main categories exist and are of interest to this thesis. These are:

- ***Data-driven models:*** Where data-repositories (typically component specific) are "mined" for features which can be exploited for condition assessment and the eventual early warning, diagnosis and detection of faults (Kai Goebel, 2008).

- ***Physics of failure models:*** Employing derived physical models of degradation – such as Paris' law (Paris and Erdogan, 1963) – to provide condition assessment.

- ***Statistical models:*** Where a model (an approximation of the underlying system) is employed to define assumptions regarding the generation of the data. Simply put, a statistical model is a probability distribution constructed to enable inferences to be drawn or decisions made from data (such as hypothesis testing) (Davison, 2003).

- ***Hybrid approaches:*** These consist of two or more of the above models combined together in the expectation of improved performance. Such as in Chen (2014).

Similarly, due to the high rate of data acquisition available for data collected from a wind turbine (typically at 5 kHz or higher), many avenues of analysis can be explored. Three distinct domains are commonly utilised for the exploration and analysis of the high frequency data. Briefly, these are:

- ***Time domain:*** The analysis of high frequency data with respect to time.

- ***Frequency domain:*** The analysis of high frequency data with with respect to frequency, rather than time. No notion of time is present within the frequency domain.

- ***Time-frequency domain:*** The analysis of high frequency data with respect to both time and frequency data. This often (but not always) utilises a time-frequency distribution to represent the data, often created through the use of a transformation function.

As such, for the purposes of the review of techniques for data processing, the techniques identified and presented within the literature will be grouped according to the domain in which their analysis is performed. Time domain techniques will be discussed in section 2.4.2, frequency domain techniques in section 2.4.3 and finally, time-frequency domain techniques will be discussed in section 2.4.4. Due to the inherent differences between low and high frequency data, the techniques applicable to low-frequency data – specifically SCADA data – are discussed separately in section 2.4.5. Following this, a discussion of the literature with the challenges and limitations of the current state of the art is given in section 2.6. A summary of these techniques, their applicability to this thesis, the current state of the art, challenges and limitations is then presented in section 2.7.

2.4.2   Time domain techniques

Time domain analysis covers a board spectrum of condition monitoring techniques. Jardine et al. (2006) defines time-domain analysis as any technique which:

> "... is directly based on the time waveform itself. Traditional time-domain analysis calculates characteristic features from time waveform signals as descriptive statistics such as mean, peak, peak-to-peak interval, standard deviation, crest factor, high-order statistics: root mean square, skewness, kurtosis, etc." (Jardine et al., 2006).

These techniques for analysis of data within the time domain are explored in detail below. In total, 7 broad techniques for the analysis of condition monitoring data with respect to time (in the time-domain) are presented.

Time domain analysis

Time domain analysis – as the name suggests – is employed to compare the raw time series data collected from an asset. It is perhaps the most primitive time domain technique, and as such, is not often utilised in practice. This is because both frequency features and domain knowledge cannot be extracted from the signal utilising this technique, however, it does provide a computationally inexpensive means of comparing the required signals directly (Box et al., 2008).

Statistical features

Statistical features have been explored in time domain analysis for condition assessment for the last 60 years (Davis, 1952). As such, these features are well understood and have

been utilised extensively in many different studies. Statistical time-domain features which have been explored for condition assessment include (but are not limited to):

- **Mean** *(Yang et al., 2011):* This quantifies the central point of the signal over time and can be utilised to identify changes in behaviour.

- **Standard deviation (and variance)** *(Yang et al., 2011):* This quantifies the variability of the signal over time and can be used to assess changes in the underlying distribution of data.

- **Skewness** *(Yang et al., 2011):* This quantifies the asymmetry of the underlying distribution and can assist in characterising changes of the underlying data.

- **Kurtosis (and other higher order moments)** *(Yang et al., 2011):* This quantifies the "peakedness" of the underlying distribution and can assist in characterising changes of the underlying data.

- **Interquartile range** *(Yang et al., 2011):* This quantifies the range between the 25th and 75th percentile and can be explored as a measure of variability of the underlying data.

- **Percentiles** *(Yang et al., 2011):* These quantify various positions of the underlying distribution can be employed to assist in the characterisation of extreme or anomalous points.

- **Root mean square (RMS)** *(Yang et al., 2011):* This is the quadratic mean and can also be utilised to quantify the central point of a distribution over time.

- **Crest factor** *(Yang et al., 2011):* This is a ratio of the peaks to the average value of a waveform, quantifying how extreme peaks within the data are and can be used to assist in characterising the peakedness of the underlying signal.

- **Peak** *(Eric Bechhoefer, 2009):* This denotes the maximal (or minimal) peak identified within a signal, and can be utilised to assist in characterising extreme values in the underlying signal.

- **Peak-to-peak interval** *(Eric Bechhoefer, 2009):* This quantifies the time between peaks within a signal and can be used to assist in quantifying deviations from periodicity of an asset or component.

Typically, the computation of these techniques can be performed in real-time due to their simple nature and efficient techniques for their computation (Godwin and

Matthews, 2014c). Recent work has begun to employ ensembles of these features which are incorporated into a univariate condition metric. This is done as individually some features may not fully encapsulate the current operational behaviour of the component which is being monitored.

Time synchronous averaging

Time synchronous averaging (TSA) is a more recent technique which has appeared in many recent publications within the PHM community. TSA involves the use of a tachometer to track the revolution of the specified component – such as a bearing or gear – in order to deconstruct the signal into what could effectively be described as a "revolution" (or angular) domain (Eric Bechhoefer, 2009). Each complete revolution of the specified component is averaged together; this assists in the removal of noise and enhances the quality of the waveform. In its simplest form, the TSA of a signal is given by the re-sampling:

$$\overline{TSA}(t) = \frac{1}{N} \sum_{n=1}^{N-1} s(t + nT), \quad 0 \le t < T \tag{2.1}$$

Where $\overline{TSA}(t)$ denotes the TSA of signal t, T the averaging period and N the number of samples for averaging.

For the purposes of this thesis, the TSA re-sampling would be beneficial into the angular domain. With respect to a gearbox, this can be performed as per Combet and Gelman (2007) and first proposed by McFadden (1986)):

$$\overline{TSA_{angular}}(t) = \sum_{k=1}^{K} X_K (1 + a_k(t)) \cos(2\pi k f_m t + \phi_k(t) + \Phi_k) + b(t) \tag{2.2}$$

Where Combet and Gelman (2007) define X_k as the amplitude of the k-th mesh harmonic, f_m is the average gear mesh frequency, $b(t)$ is background noise, Φ_k is the initial phase harmonic of k, and $a_k(t)$ and $\phi_k(t)$ are the amplitude and phase modulation functions of mesh harmonic k.

Although the implementation of the TSA algorithm utilises a tachometer to track the current revolution of the component, various "tachometer-less" TSA algorithms have been proposed within the literature. Typically, these use one of two techniques; time slicing or re-sampling (Eric Bechhoefer, 2009).

Time sampling assumes a stationary signal of constant revolution. As such, if a given revolution is known to take a specified length of time, the waveform can be partitioned into these lengths for the purposes of averaging.

Re-sampling is performed by identifying peaks within the waveform as an identifier as to the approximation of the current revolution of the component; gear mesh harmonics and bearing inner or outer race harmonics are typically used for this purpose (Johnson, 2011). Once these are known, waveform partitioning can be performed in a similar manner to time slicing, utilising the more accurate approximation of the current revolution (however, this is not as accurate as the use of a tachometer).

There are many advantages and disadvantages to the use of TSA as a data processing technique. Interested readers should refer to the reviews by both Eric Bechhoefer (2009) and Dalpiaz et al. (2000) which detail both tachometer and tachometer-less algorithms, their advantages and their disadvantages.

Practical applications of the use of the TSA algorithm for the purposes of condition assessment are numerous throughout the literature. The works listed below details a selection of applications of this technique which may be of interest to readers of this thesis:

- The application of two tachometer-less TSA algorithms to planetary gearboxes is performed by D'Elia et al. (2013) for the purposes diagnosis.

- In work by Combet and Gelman (2007), the application of tachometer-less TSA is employed for condition assessment of a two-stage helical gearbox.

- Pump condition assessment as per patent US6681634 B2 (Sabini et al., 2004) is performed via TSA.

- The application of TSA for bearing condition assessment is undertaken by Luo et al. (2010).

Autoregressive moving average

Autoregressive (AR) techniques are another commonly employed time domain waveform analytical technique employed for condition assessment. This is a linear technique utilised in random processes to define an output variable based upon its own previous values. An $AR(p)$ model (an auto-regressive model of order p) can be defined as per Pandit and Wu (1983):

$$AR(p) = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t \tag{2.3}$$

where c is a constant, ϵ_t is white noise and $\varphi_1 \cdots \varphi_p$ are parameters of the model. Due to the linearity of the approach, often moving average (MA) models are employed in their place. These models require non-linear fitting models rather than the traditional

least squares technique employed in auto-regressive models. A $MA(q)$ model (a moving average model of order q) can be defined as per Pandit and Wu (1983):

$$MA(q) = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{2.4}$$

where μ is the expectation of X_t, $\epsilon_t, \epsilon_{t-1} \cdots$ are white noise and $\theta_1 \cdots \theta_q$ are the parameters of the model.

Due to auto-regressive techniques utilising the lagged dependent variable and moving average techniques utilising lagged error terms, often these two techniques are employed together in an auto-regressive moving average (ARMA) model. Within ARMA models, two polynomials are defined to encapsulate the behaviour of the signal; the first representing the auto-regressive element, and the second representing the moving average. An $ARMA(p, q)$ model (an auto-regressive moving average model with p auto-regressive terms and q moving average terms) can be defined by Pandit and Wu (1983) as:

$$ARMA(p, q) = c + \epsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{2.5}$$

This technique is used to model and predict the future values of the dependent variable. This modelling can be performed with minimal historical data, and can be utilised within linear time-invariant systems, however, their application to non-linear or dynamic systems is limited (Lee et al., 2014).

ARMA models have been employed extensively over the last 30 years for condition assessment of both bearings and gearboxes. For instance, the works presented below present practical applications of the ARMA technique which may be of interest to readers of this thesis:

- Higher order ARMA models were utilised by Li et al. (2007) for the fault detection of bearings.

- Auto-regressive modelling was exploited by Junsheng et al. (2006) for bearing fault diagnosis.

- ARMA modelling for gearbox condition assessment was performed by Chen et al. (2011b).

- In Kang et al. (2012) gearbox condition assessment for non-stationary signals is performed after re-sampling into the angular domain.

Various extensions to the AR, MA and ARMA models exist, most commonly the auto-regressive integrated moving average (ARIMA) model. This model has also extensively been used for condition assessment (Ahmad and Kamaruddin, 2012; Kosasih et al., 2014). Similarly, the ARAR model which is an adaption of the ARARMA model (Newton and Parzen, 1983). Also, models such as seasonal ARIMA (SARIMA) have also been proposed (Brockwell and Davis, 2002).

Singular spectrum analysis

Singular spectrum analysis (SSA) is a non-parametric spectral estimation technique often employed in the field of chaos theory in order to assist in the understanding of chaotic and non-linear behaviour. This technique can be exploited to forecast the future condition of components of interest and can provide various ancillary features, as described by Hassani (2007), which include:

- Finding trends of different resolution.

- Smoothing.

- Extraction of seasonality components.

- Simultaneous extraction of cycles with small and large periods.

- Extraction of periodicities with varying amplitudes.

- Simultaneous extraction of complex trends and periodicities.

- Finding structure in short time series.

- Change-point detection.

The application of singular spectrum analysis consists of four steps performed in two stages as follows (Hassani, 2007):

1. ***Deconstruction (embedding):*** In this stage, the one-dimensional time series is embedded into a multi-dimensional space with a fixed window size and lag length. This creates a Hankel matrix (the trajectory matrix). This is a common operation in time-series analysis.

2. ***Deconstruction (singular value decomposition):*** In this step, singular value decomposition (SVD) is performed upon the trajectory matrix, representing it as a sum of rank-one, bi-orthogonal elementary matrices.

3. **_Reconstruction (grouping):_** The elementary matrices from the previous step are split into several groups and summed. This is done by eigentriple grouping.

4. **_Reconstruction (diagonal averaging):_** Hankelisation of the matrix is performed (the matrix diagonals are averaged) allowing for a reconstruction of the time series.

Whilst the technical details of this technique are beyond the scope of this thesis, interested readers should refer to Elsner and Tsonis (1996) or Hassani (2007) for more in depth details of the implementation of this technique in practice. Due to the quality of prediction obtained and the non-linear nature of this technique, it has been utilised in many cases for condition assessment of both bearings and gearboxes. Below are practical examples of this use of singular spectrum analysis which may be of interest to readers of this thesis:

- Bearing inner race fault diagnosis (Muruganatham et al., 2010) and roller element fault diagnosis (Muruganatham et al., 2013) were performed with the assistance of SSA.

- Early detection of faults using SSA was performed in Bovic Kilundu and Dehombreux (2011).

- SSA was employed in Wang et al. (2001) in addition to the pseudo-phase portrait and correlation dimension for condition assessment.

- Component wear was assessed and predicted using SSA by Salgado and Alonso (2006) based upon high frequency data (10 kHz).

Lyapunov exponent

The Lyapunov exponent (also known as the Lyapunov characteristic exponent) is a metric which quantifies the separation of two trajectories in phase space. The rate of divergence of two given trajectories in phase space can be given as per Massimo Cencini and Vulpiani (2009):

$$|\delta\mathbf{Z}(t)| \approx e^{\hat{\lambda}t}|\delta\mathbf{Z}_0| \tag{2.6}$$

Where $\hat{\lambda}$ (in this case) is the Lyapunov exponent, $\mathbf{Z}(t)$ is the phase state vector at time t and $\mathbf{Z}_0$ is the initial phase state separation. As the rate of separation will differ intrinsically dependent upon the initial orientation of a vector (i.e. initial asset health), there is a distribution of Lyapunov exponents. For the purposes of condition assessment,

typically the largest Lyapunov exponent within this distribution is of interest; this is referred to as the maximal Lyapunov exponent (MLE).

The maximal Lyapunov exponent is of interest within the realm of condition monitoring as it is often used to represent the predictability of a dynamic system (such as a bearing or gearbox). Thus, being able to quantify the predictability of the component is inherently beneficial for the purposes of condition assessment or predictive (proactive) maintenance. The MLE can be defined for a continuous system as per Massimo Cencini and Vulpiani (2009):

$$\hat{\lambda} = \lim_{t \to \infty} \lim_{\delta \mathbf{Z}_0 \to 0} \frac{1}{t} ln \frac{|\delta \mathbf{Z}(t)|}{|\delta \mathbf{Z}_0|} \tag{2.7}$$

Similarly, for a discrete system (Massimo Cencini and Vulpiani, 2009):

$$\lambda(\hat{x_0}) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} ln|f'(x_i)| \tag{2.8}$$

For the purposes of condition assessment, both global and local maximal Lyapunov exponents are of interest. Global Lyapunov exponents refer to the predictability of the overall (entire) system, whereas local Lyapunov exponents refer to the predictability of a given point in phase space. For instance, the behaviour of a bearing or gearbox operating under normal operational condition may behave in a predictable manner, however, degradation is often referred to as a a chaotic phenomena (Lu et al., 2013), and as such, determining how predictable the current system is may provide insights into the degradation that is present.

As this technique is able to quantify the chaotic nature of signals (and infer asset or component condition from this), this technique has also been explored extensively. The works given below present practical applications of exploiting the Lyapunov exponent for condition assessment which may be of interest to readers of this thesis:

- Bearing fault severity was assessed with the assistance of the Lyapunov exponent in work done by Ghafari et al. (2008).

- Lyapunov exponents were used to assess the chaotic nature of bearing data in Harsha et al. (2003).

- Nonlinear behaviour of a geared rotor was characterised by Gao and Zhang (2014) utilising Lyapunov exponents.

- The identification of undesirable operational modes for gears was performed by Chang-Jian and Chang (2011) utilising SSA in order to minimise degradation by avoiding such conditions.

Principal component analysis

Principal component analysis (PCA) is a statistical technique which decomposes a set of data (such as a a set of condition monitoring waveforms) into a new set of variables (principal components) which represent the maximum variability within the data, with the minimal loss of information. For this reason, it is popular in areas such as data compression, however, it can be used for variable classification, detection of anomalies (outliers) and also early indication of abnormality in a given data structure (Baydar et al., 2001).

In order to perform principal component analysis, multivariate data is needed. Due to the reduction of economic constraints on data storage and data acquisition technologies, typically condition monitoring data is multivariate in nature. For instance, three accelerometers monitoring a bearing or gearbox is an example of a multivariate problem which can be explored using principal component analysis. Similarly, extracted features from a univariate signal (such as skewness, kurtosis and standard deviation as discussed previously) would likewise be a multivariate problem which could be also explored within the framework of principal component analysis.

Principal component analysis is an orthogonal transformation of the data which takes a set of potentially correlated variables and creates a set of linearly uncorrelated variables. Given a data matrix X, the transformation is defined by a $p-$dimensional vector of weights $w_k = (w_1 \cdots w_p)$ that map the vector x_i of X to a vector of scores $t_i = (t_1, \cdots t_p)_i)$ where $t_{k_i} = x_i \cdot w_k$. The first weight vector has to satisfy (Jolliffe, 2002):

$$w_{(1)} = \arg \max \left\{ \frac{w^T X^T X w}{w^T w} \right\} \tag{2.9}$$

Following the selection of this principal component, further components can be selected as per Jolliffe (2002):

$$w_{(k)} = \arg \max \left\{ \frac{w^T \hat{X}_{k-1}^T \hat{X}_{k-1} w}{w^T w} \right\} \tag{2.10}$$

Where $\hat{X}_{k-1}$ is defined by subtracting the first $k-1$ principal components from X (Jolliffe, 2002):

$$\hat{X}_{k-1} = X - \sum_{s=1}^{k-1} X w_s w_s^T \tag{2.11}$$

Principal component analysis can be performed in many ways, and is typically done in the stages as specified by Abdi and Williams (2010):

1. Organisation of the dataset.

2. Calculation of empirical mean.

3. Calculation of deviation from the empirical mean.

4. Calculation of covariance matrix.

5. Calculation of eigenvector and eigenvalues from covariance matrix.

6. Rearrangement of eigenvectors and eigenvalues.

7. Computation of eigenvector energy content.

8. Eigenvector subset selection.

For convenience, the details of these calculations and computations are not discussed in depth in this thesis. Readers interested in this technique should refer to work of andW. Wu et al. (2002), Jolliffe (2002) and Abdi and Williams (2010).

Due to the high dimensionality of data which is often available within the realm of condition assessment and health management, ensuring the correct features are extracted from the data is essential to reduce noise and increase the sensitivity of condition assessment techniques. As such, principal component analysis is often performed to assist in the early detection, diagnosis and prognosis of faults on bearings and gearboxes. The works listed below employ principal component analysis for practical applications which perform condition assessment, which may be of interest to readers of this thesis:

- Work done by Zvokelj et al. (2011) use principal component analysis to assist in the de-noising for bearing fault detection and identification.

- In work done by Malhi and Gao (2004), principal component analysis is performed without fault data to extract useful signatures for bearing fault identification.

- Non-linear feature extraction is performed on gearboxes in He et al. (2007) by exploiting principal component analysis for condition monitoring.

- Similarly, principal component analysis is utilised in Shao et al. (2014) to find two principal components for gearbox fault classification.

Time-domain technique conclusions

As we have explored, many techniques exist within the time-domain for the purposes of extracting features which can be used for condition assessment. Many of the presented

techniques are readily available in many commercial off the shelf systems. Due to the use of the time domain and the simplicity of the techniques, often these can be computed in real time. This enables real-time analysis of data, and can (if necessary) be utilised in mission-critical components.

Furthermore, due to the nature of the time domain, it is often possible (when utilising dominant features) to identify the moment at which a change to the time series occurs. Due to the often temporal dependency of degradation within bearings and gears, it is of interest to identify deviations from previous conditions at the earliest possible moment. This enables the potential for proactive maintenance to be undertaken, realising reductions in maintenance budgets and potentially increasing availability. This would not be possible in other domains where no notion of time exists (such as in the frequency-domain, for instance).

2.4.3 Frequency domain techniques

Having explored the analysis of data with respect to time, it follows that a further, complementary analysis can be performed in the frequency domain in order to overcome the limitations identified by the review of time domain techniques. As time-domain techniques have no notion of frequency, various meta-data cannot be included in the analysis (for instance, if we know the configuration of a gearbox, this can be utilised as meta-data within frequency-domain analysis).

Frequency domain analysis covers algorithms which typically employ a transform of the data into the frequency domain (typically from the time domain). In some instances, this may have an advantage over traditional time domain analysis in that it is easy to isolate various frequency features within the data (Jardine et al., 2006). Typically, within the realm of condition assessment, the notion of the frequency domain is closely coupled with that of the Fourier transform (Rahman, 2011), however, there are a range of transforms which can represent frequency data; for instance, the Laplace transform (Widder, 1941). For the purposes of this thesis, the frequency domain will assume the Fourier transformation unless stated otherwise.

Fourier transform

The Fourier transform (Rahman, 2011) is a mathematical transformation that converts a signal based from the time domain to that of the frequency domain. The Fourier transform is a generalisation of the complex Fourier series and can be defined as (Rahman, 2011):

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i x \xi} dx \quad \text{for any real number } \xi \tag{2.12}$$

Where x represents time and ξ represents the frequency component. In this case, $f(x)$ is determined by the inverse transformation (Rahman, 2011):

$$f(\xi) = \int_{-\infty}^{\infty} \hat{f}(\xi)e^{2\pi i \xi x} d\xi \quad \text{for any real number } x \tag{2.13}$$

Together, these equations are referred to as the Fourier transform pair (Rahman, 2011). Due to the representation of the time signal in the frequency domain, various artefacts can be observed in the data which would otherwise be difficult to analyse. For instance, given a bearing, it is possible to determine specific harmonic frequencies for the inner race, outer race, train frequency and ball spin. These harmonics are referred to as the ball pass frequency outer (BPFO), ball pass frequency inner (BPFI), fundamental train frequency (FTF) and ball spin frequencies (BSF). For completeness, these are detailed in Eq. 2.14 through Eq. 2.17 as per McInerny and Dai (2003).

$$BPFO(Hz) = S\frac{N}{2}(1 - \frac{B}{P}\cos\phi) \tag{2.14}$$

$$BPFI(Hz) = S\frac{N}{2}(1 + \frac{B}{P}\cos\phi) \tag{2.15}$$

$$FTF(Hz) = S\frac{1}{2}(1 - \frac{B}{P}\cos\phi) \tag{2.16}$$

$$BSF(Hz) = S\frac{P}{2B}(1 - \frac{B^2}{P^2}\cos^2\phi) \tag{2.17}$$

Where S is the shaft rotation (in Hz), P is the pitch diameter, B is the ball diameter, ϕ is the contact angle between the inner and outer races and N is the number of balls (McInerny and Dai, 2003).

As such, by employing the Fourier transform, changes associated with these specific faults may be indicative of faults associated with the component under scrutiny. Similarly, harmonics can be defined for gears within a gearbox to allow for frequency domain analysis. Their are five fundamental gear frequencies of interest (Niola et al., 2009): gear (F_{rg}), pinion (F_{rp}), tooth mesh (F_m), assembly phase (F_a) and tooth repeat (F_{tr}) – or hunting tooth – frequencies. These are defined below for completeness (Niola et al., 2009):

$$F_{rg} = \frac{R_g}{60} \tag{2.18}$$

$$F_{rp} = \frac{R_p}{60} \tag{2.19}$$

$$F_m = F_{rp} \cdot N_p \quad \text{or} \quad F_{rg} \cdot N_g \tag{2.20}$$

$$F_a = \frac{F_m}{N_a} \tag{2.21}$$

$$F_{tr} = \frac{F_m \cdot n_a}{N_g \cdot N_p} \tag{2.22}$$

Where R_g, R_p represents the gear and pinion speed in RPM (respectively), and N_g, N_p represents the number of teeth on the gear and pinion (respectively).

To illustrate the Fourier transform and provide a practical example of the technique, figure 2.5 presents the Fourier transform of bearing data with both the inner and outer race harmonics highlighted. Typically, the computation of the Fourier transform is performed through the use of the fast Fourier transform (FFT) algorithm (Cooley and Tukey, 1965). This reduces the computational intensity of the calculation, and makes it viable to perform Fourier analysis on high frequency condition monitoring data.

The fast Fourier transform utilises the discreet Fourier transform, defined as per
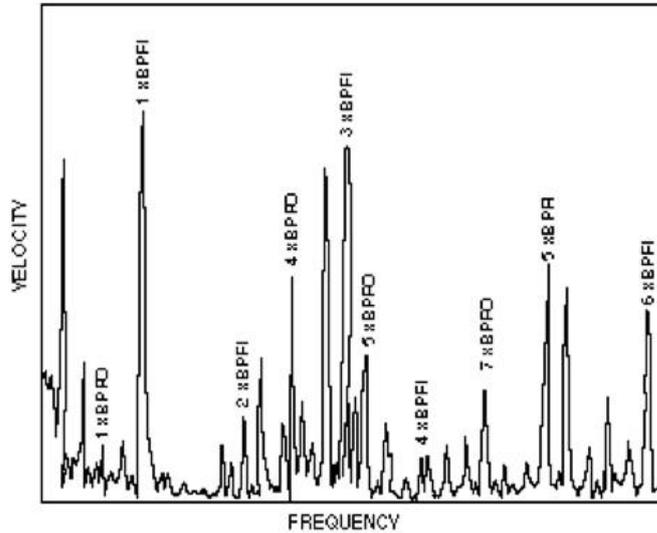
Figure 2.5: Example Fourier analysis of bearing with highlighted inner and outer race harmonics (Buscarello, 1994).

Rahman (2011):

$$FFT(x) = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, \cdots, N-1 \qquad (2.23)$$

Due to the efficiency in the calculation and the insights gained from analysis of specific harmonics, Fourier analysis is regarded as the most commonly utilised frequency domain technique in condition assessment. The works below utilise Fourier analysis for bearing and gearbox condition assessment which may be of interest to readers of this thesis:

- In María et al. (2000) various Fourier transforms are compared for their effectiveness on vibration data.

- Fourier analysis is employed by Rafsanjani et al. (2009) to validate their proposed bearing diagnostic methodology.

- In work done by Rai and Mohanty (2007), Fourier analysis is explored and then extended for bearing fault diagnosis.

- Trade off analysis when using Fourier analysis is explored for aircraft bearings in work done by Hai Qiu and Eklund (2009).

Cepstrum analysis

Cepstrum analysis is a frequency domain technique which extends the Fourier transform in order to detect harmonics and sidebands in the power spectrum (Jardine et al., 2006). Cepstrum analysis can employ various different spectra; typically within the realm of condition assessment, the power Cepstrum is utilised. This is often defined as *"the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal."* (Bogert et al., 1963), or formally:

$$PC(x) = |f(\log(|f(\hat{f}(x))|^2))|^2 \qquad (2.24)$$

Where (previously) $f(x)$ represents the inverse Fourier transformation and $\hat{f(x)}$ represents the Fourier transformation. Other Cepstrums exist, such as phase, complex and real Cepstrum. The complex Cepstrum is defined as (Oppenheim, 1965):

$$CC(x) = \hat{f}(\log(\hat{f}(x)) + j2\pi m) \qquad (2.25)$$

Where m is the integer required to properly unwrap the imaginary part of the complex log function.

Although the power Cepstrum has a notion of time – the independent variable is effectively time (as it is the inverse Fourier transformation of a spectrum) – it is not a time domain technique. The independent variable is often referred to as "quefrency" (Bogert et al., 1963) as it is not a true representation of time, but rather represents the periodic harmonics within a signal.

Due to the concept of quefrency arising in Cepstrum analysis, along with the ability to identify harmonics and sidebands, it has been utilised in condition assessment. Below details works employing Cepstrum analysis which may be of interest to readers of this thesis:

- In work undertaken by Park et al. (2013), Cepstrum analysis is used to denoise the signal and perform early detection of bearing faults.

- Bearing fault diagnosis is performed by Fang et al. (2012) by utilising the power cepstrum.

- Cepstrum analysis was employed in collaboration with other techniques for non-stationary signals by Sawalhi et al. (2014) to separate and enhance gearbox and bearing signals.

- Work done by Ziaran and Darula (2013) has shown the benefit of Cepstrum analysis in conjunction with traditional Fourier analysis for gear condition assessment.

Sideband Analysis

Sideband analysis is a natural extension to the Fourier transform and is closely related to Cepstrum analysis. With this technique, not only are specific known harmonics targeted, but frequency peaks associated with the harmonics are included in the analysis for additional context. This ensures that interactions between harmonics are taken into consideration. By looking at the power spectrum of the Fourier transform, and with the assistance of techniques such as Cepstrum analysis or the manual identification of harmonics, further insight into asset condition can be achieved.

Sideband analysis can be exploited for the assessment of non-stationary signals (Zappalá et al., 2012). As the harmonics of interest will change with respect to the loading of rotating machinery such as gears and bearings, by tracking these and the interactions with the sidebands, non-stationary signals can be analysed effectively. This is somewhat similar in nature to the re-sampling of time-domain signals into the angular (or revolution) domain, however, in this instance, in the frequency domain. Figure 2.6 (Zappalá et al., 2012) provides an contrasting example of normal operational behaviour of the gearbox compared to a missing tooth.
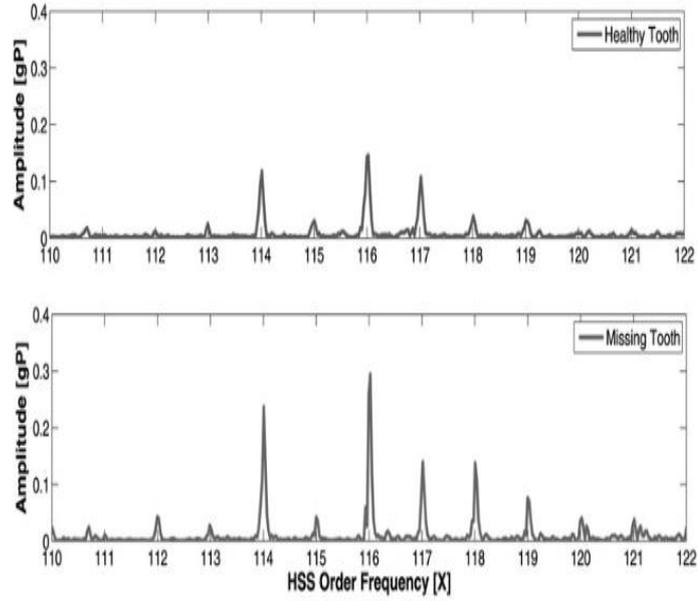
Figure 2.6: Sideband analysis as performed by Zappalá et al. (2012), showing healthy and missing gear tooth fault.

Quantification of the sidebands in order to provide a metric by which to measure asset condition can be performed. This can further be correlated against known fault modes to assist in early detection and diagnosis. For instance, focusing on the second harmonic of the gear tooth mesh, the sideband power factor (SBPF) metric can be written as (Zappalá et al., 2014):

$$SBPF = PSA(2xf_{mesh,HS} + \sum_{i=-5}^{i=5} PSA(SB_i)$$

(2.26)

Where $PSA(k)$ represents the power spectrum amplitude of k, with $2xf_{mesh,HS}$ representing the second harmonic of the meshing frequency of the high speed shaft, and SB_i representing the ith sideband of this frequency, defined as (Zappalá et al., 2014):

$$SB_i = (2xf_{mesh,HS} \pm i)X$$

(2.27)

Where in this case, X represents the normalised order spectra (X) which is defined by (Zappalá et al., 2012) as by normalising the frequency axis against the high speed shaft rotational frequency of the gearbox.

Due to the ability to perform condition assessment and diagnosis on non-stationary signals, this technique has been employed in wind turbine gearbox health analysis. The

works presented below provide practical examples of sideband analysis and are related to this thesis:

- In work by Zappalá et al. (2012), the SBPF algorithm is defined, and is shown to successfully detect and diagnose gearbox faults.

- Building on Zappalá et al. (2012), work done by Zappalá et al. (2014) presents a metric by which to quantify current degradation of the gearbox.

- Bearing condition assessment is performed by Reuben and Mba (2014), and exploits sideband analysis to assist their analysis.

- Angular re-sampling is performed without a rotational reference signal by Villa et al. (2011) to enable sideband analysis on a variable gearbox.

Envelope Analysis

Envelope analysis builds upon traditional Fourier analysis by pre-filtering signals for analysis in order to remove noise from the resulting Fourier spectrum. Three basic steps are performed (Hochmann and Bechhoefer, 2005):

1. ***Signal frequency shift (heterodyne):*** In this stage, a heterodyne operation is performed to effectively shift the current waveform frequencies into a new frequency range (window selection)

2. ***Low pass filter:*** In this stage, high frequencies within a specific range are removed from the waveform.

3. ***Magnitude calculation:*** In this stage, the squared magnitude of the complex signal is calculated though multiplication with the complex conjugate.

This procedure is defined formally in the work of Hochmann and Bechhoefer (2005). As the signal is filtered, noise is removed and further enhances the analytical capabilities already explored by the Fourier transform. As can be seen in figure 2.7, envelope analysis can assist in the identification of harmonic frequencies which are not explicitly clear under traditional Fourier analysis.

Traditionally, envelope analysis is a parametric technique which requires careful selection of both the window and low pass filter selection. Alternatively to this process, the Hilbert-Huang transform can be utilised for envelope analysis (Hochmann and Bechhoefer, 2005). This produces the same result as shifting the signal and performing low pass filtering. The formalities of this technique are detailed within Hochmann and Randall (2000).
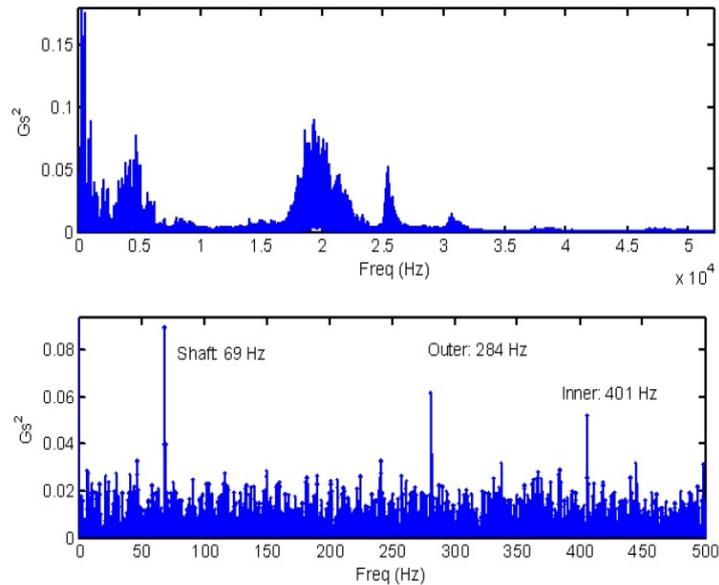
Figure 2.7: Fourier analysis (top) with envelope analysis (below) as performed by Hochmann and Bechhoefer (2005).

Due to the ability to de-noise and assist in condition assessment, envelope analysis has been used extensively in the literature and is available on commercial condition monitoring systems, such as those available by Brüel & Kjær Vibro. As such, there is substantial literature available based upon the use of envelope analysis for condition assessment. Below details some practical examples of envelope analysis which are relevant to this thesis:

- In work undertaken by Zhang and Ai (2008) envelope analysis is used in conjunction with empirical mode decomposition and the intrinsic mode function to diagnose bearing faults.

- Work done by Bechhoefer and Menon (2009) shows the application of envelope analysis to bearing condition assessment and provide a means to select the correct window.

- Gearbox fault diagnosis is performed by Liu et al. (2006) using Hilbert spectrum analysis.

- Similarly, gearbox faults are identified in work performed by Lin and Williams (2013) with pre-processing using the Hilbert transform.

Frequency-domain technique conclusions

As shown, the frequency domain provides a complementary (and often additional) avenue for analysis for high frequency data. Within the frequency domain, the use of meta-data can be employed to assist with the early detection and diagnosis of both bearings and gear faults (if this is known a priori). This is traditionally not possible within the time domain, and as such, in many cases enables a deeper insight to be achieved. It is for this reason that techniques such as the Fourier transform have been utilised by practitioners for many decades.

However, within the frequency domain, no notion of time exists. As such, it is not possible to tell when a specific change occurred, just that a change has occurred. Furthermore, many techniques within the frequency domain are incapable of handling non-stationary signals. Given the non-stationary nature of wind turbine gearbox data, this immediately limits the use of these techniques within this thesis.

Furthermore, analysis within the frequency domain requires high frequency data. As the reconstruction of a signal (at a given frequency) requires data to be collected at twice this frequency (due to Nyquist's theorem), these techniques cannot practically be applied to the low frequency SCADA data collected from a wind turbine.

2.4.4   Time-frequency domain techniques

As shown, both the time-domain and frequency-domain have limitations. In the case of the time-domain, frequency elements of the signal are ignored and meta-data cannot be utilised in analysis. Similarly, with regards to the frequency domain, as no notion of time exists, it is not possible to determine the time at which degradation occurred (only that degradation has, or has not, occurred at the time of the analysis).

Time frequency domain techniques incorporate elements from time domain analysis and frequency domain analysis in order to mitigate the shortcomings of both the time-domain and the frequency-domain to provide both of these elements (at some degree of resolution) to assist in the condition assessment of components. Due to frequency-domain techniques traditionally not being capable of handing non-stationary signals, time-frequency techniques have been developed to handle the non-stationary signals which typically occur when components fail (Jardine et al., 2006).

Short time Fourier transform

The short time Fourier transform (STFT) is a transform in the time-frequency domain, which extends traditional frequency-domain Fourier analysis to express time as well as frequency. Simply put, it is a fixed size window which is moved over a snap shot of

time, from which Fourier analysis is performed. In the continuous case, it is defined as (Cohen, 1995a):

$$STFT\{x(t)\}(\tau, \omega) \equiv X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}dt \qquad (2.28)$$

Where $x(t)$ is the raw signal and $w(t)$ is the window function. Given the fixed nature of the window, selection of an appropriate window size for the Fourier analysis is of crucial importance. This fixed window size provides fixed resolution for the resulting analysis. This causes a trade off between the resolution of time and the resolution of frequency which is presented. The window width defines if there is good time resolution or good frequency resolution. Increased time resolution allows easier identification of the point in time at which frequencies change, whereas increased frequency resolution allows different frequency components close together to be separated. As such, the wider the window, the better the frequency resolution, but lower the time resolution. This is demonstrated in figure 2.8.

Due to the simplicity of the short time Fourier transform, and the use of time-frequency techniques for the analysis of non-stationary signals, the technique has shown promise for condition assessment. The works presented below employ the short time Fourier transform for the practical application of condition assessment, and which may be of interest to readers of this thesis:

- The short time Fourier transform was employed by Cocconcelli et al. (2012a) for bearing fault diagnosis.

- In work undertaken by Cocconcelli et al. (2012b), fault data is extracted from the
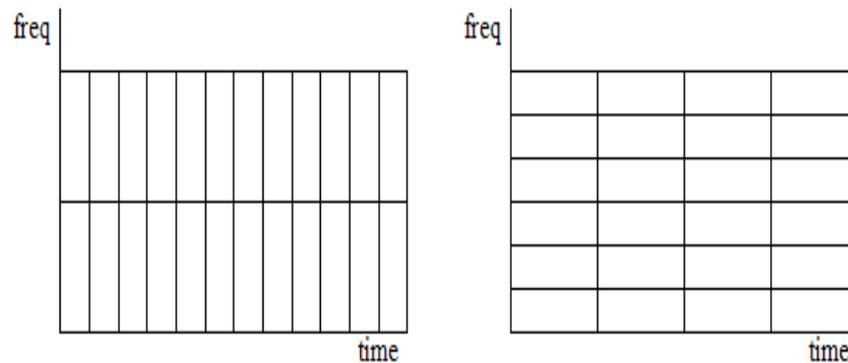
Figure 2.8: Short time Fourier transform showing superior time resolution (left) and superior frequency resolution (right).

short time Fourier transform.

- Gearbox condition assessment is performed with the assistance of the short time Fourier transform by Kar and Mohanty (2006).

- A notion similar to the short time Fourier transform is explored by Wang and McFadden (1993) in the angular domain.

Wavelet analysis

Perhaps the most well known time-frequency technique for condition assessment is that of Wavelet analysis. A wavelet is simply an oscillation – like one would observe on a heart monitor – which starts at zero, increases, then decreases, and finally returns to zero. As such, various different wavelets exist. The wavelet transform utilises a wavelet to transform the signal from the time domain to the frequency domain. This differs from the short time Fourier transform, which employs sinusoidal functions. The continuous wavelet transform (CWT) is defined as per Rafiee et al. (2010):

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi \star \left(\frac{t-b}{a}\right) dt \qquad (2.29)$$

Where in this case, $x(t)$ is the signal, a is a scale parameter, b is the time parameter, ψ is a wavelet function and $\star$ denotes the complex conjugate. A discrete wavelet transform also exists. Various different wavelet functions exist, and are known as mother wavelets. Over 15 different mother wavelet functions exist (Rafiee et al., 2010). For an extensive comparison of over 300 wavelets across 15 families, the application and review by Rafiee et al. (2010) discusses the use of these wavelets in gear diagnosis.

The Morlet wavelet function (as shown in figure 2.9) is commonly employed for condition assessment; for instance, as employed in both Nikolaou and Antoniadis (2002) and Saravanan et al. (2008). Due to the impulse like nature of this particular wavelet function, it has higher similarity to vibration data than other wavelet functions, as expressed in Rafiee et al. (2010). As such, it is employed for impulse isolation (Lin and Zuo, 2003). For completeness, the mother Morlet wavelet is defined by Lin and Zuo (2003) as:

$$\psi(t) = \exp(-\beta^2 t^2/2)\cos(\pi t) \qquad (2.30)$$

Similarly, a daughter Morlet wavelet can be obtained by scale dilation and time translation of the mother Morlet wavelet. This was defined by Lin and Zuo (2003) as:
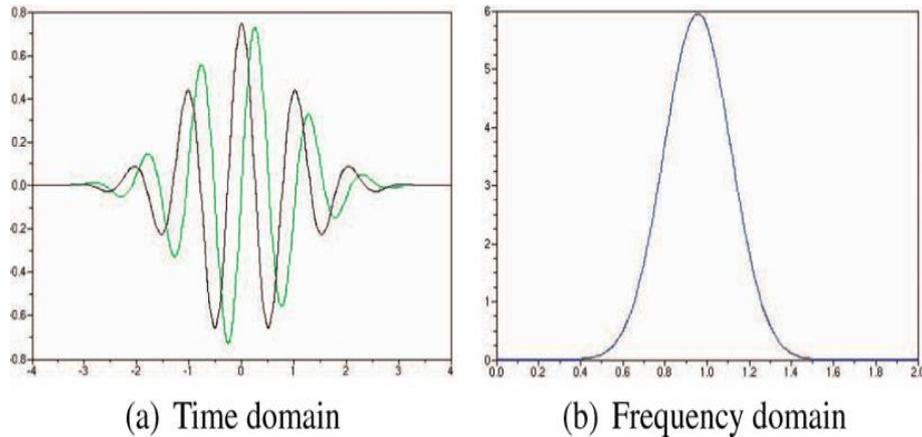
Figure 2.9: Morlet wavelet, showing time-domain (left) and frequency-domain (right) components (Janicke et al., 2009).

$$\psi_{a,b}(t) = \psi \left(\frac{t-b}{a} \right) = exp \left[-\frac{\beta^2 (t-b)^2}{2a^2} \right] \cos \left[\frac{\pi(t-b)}{a} \right] \qquad (2.31)$$

As can be seen in Eq. 2.30 and Eq. 2.31, two parameters must be set; a scaling parameter a, and shape parameter β. In order to identify the impulses in the signal (which may indicate potential degradation) the location and shape of the band of frequencies must be determined. As such, accurate identification of these parameters is essential for effective condition assessment. Typically, for condition assessment, the daughter wavelet parameters are varied and the maximal kurtosis of the resultant time series is utilised as a metric for the setting of the parameters. This is performed as kurtosis is sensitive to sharp impulses (as this measures the peakedness of a distribution), and is explored in depth in Rafiee et al. (2010).

Wavelet analysis is a multi-resolution technique. Unlike the short time Fourier transform (which has fixed resolution), the resolution in both the time and frequency domains is independent of the window function. The use of the continuous wavelet function allows for a enhanced time resolution at high frequencies (thus, reduced frequency resolution at these high frequencies) whilst providing enhanced frequency resolution at low frequencies (thus, reduced time resolution at these low frequencies). This is shown in figure 2.10.

As wavelet analysis has the capability to show both time domain and frequency domain information, it is used extensively in condition assessment. The works presented below provide practical illustrations of this technique on both gearboxes and bearings, which may be of interest to readers of this thesis:

- In work done by Su et al. (2010), optimal Morlet wavelets are employed for bearing
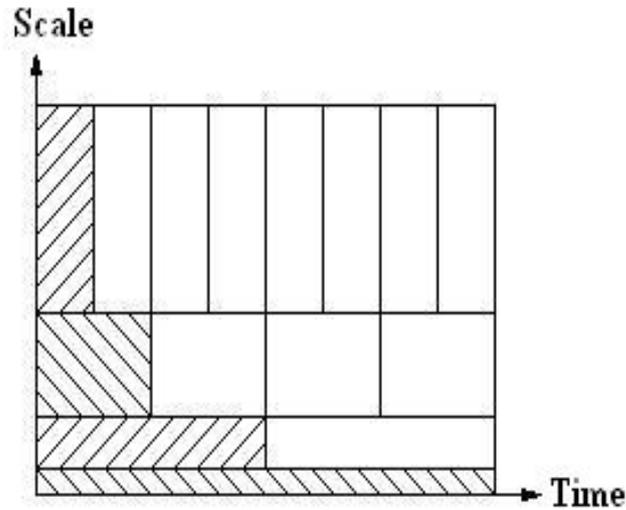
Figure 2.10: Resolution of the wavelet transform (Addison, 2002).

fault diagnosis. A genetic search is employed to derive the wavelet parameters.

- Gear tooth defects are de-noised and identified with the use of wavelet analysis in Omar and Gaouda (2012).

- Non-stationary analysis of the gearbox is performed successfully with wavelets in Wang et al. (2010).

- Bearing diagnosis is performed with the assistance of wavelets in Li et al. (2012) and presented in morphological slices.

Other Time-frequency distributions

Due to the trade-off which exists in linear techniques such as the short time Fourier transform and wavelet transform, other non-linear techniques have been presented within the literature. One such technique is the Wigner-Ville distribution (WVD), which is a bilinear time-frequency distribution which has a higher time-frequency resolution than both the short time Fourier transform and wavelet transform, but suffers from cross-term interference as a result (Feng et al., 2013). Cross term interference produces false signals between every component of the signal (Dong and Cui, 2012), and as such, has limited the application of the Wigner-Ville distribution to the problem of condition assessment. The Wigner-Ville distribution can be defined as Dong and Cui (2012):

$$WVD_x(t, f) = \int_{-\infty}^{\infty} x\left(t - \frac{\tau}{2}\right) x \star \left(t - \frac{\tau}{2}\right) \exp(-j2\pi f\tau) d\tau \tag{2.32}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A_x(\tau, \gamma) \exp[-j2\pi(\gamma t + f\tau)] d\gamma d\tau$$

Where $x(t)$ represents the signal, and A_x is the ambiguity function, defined as:

$$A_x(\tau, \gamma) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x\left(t + \frac{\tau}{2}\right) x \star \left(t - \frac{\tau}{2}\right) \exp(-j2\pi\gamma t) dt \tag{2.33}$$

Due to the limitations of the Wigner-Ville distribution, various other distributions have been proposed such as Bessel, Born-Jordan, Choi-Williams, Page, Rihaczek and Zhao-Atlas-Marks. These are all classes of Cohen class distributions which aim to exploit the high resolution of the Wigner-Ville distribution and remove noise and cross-term interference by means of a kernel function, denoted as $\Phi(t, f)$. These can be defined generally – denoted time frequency representations (TFR) – as per Cohen (1989):

$$TFR_x(t, f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(t - t', f - f') WVD_x(t', f') dt' df' \tag{2.34}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(\tau, \gamma) A_x(\tau, \gamma) \exp[j2\pi(\gamma t - ft)] d\tau d\gamma$$

Where $\phi(t, f)$ is defined as:

$$\phi(t, f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(\tau, \gamma) \exp[j2\pi(\gamma t - f\tau)] d\tau d\gamma \tag{2.35}$$

The kernel functions for commonly utilised time-frequency representations are (Feng et al., 2013):

- *Bessel:* $\Phi(\tau, \gamma) = J_1(2\pi\alpha\gamma\tau)/(\pi\alpha\gamma\tau)$ where J_1 is a Bessel function (of the first kind) of order one, with $\alpha > 0$.

- *Born–Jordan:* $\Phi(\tau, \gamma) = \sin(\pi\gamma\tau)/(\pi\gamma\tau)$.

- *Choi–Williams:* $\Phi(\tau, \gamma) = \exp[2(2\pi\gamma\tau/\sigma)^2]$ where σ is a scaling factor.

- *Page:* $\Phi(\tau, \gamma) = \exp(j\pi\gamma|\tau|)$.

- *Rihaczek:* $\Phi(\tau, \gamma) = \exp(j\pi\gamma\tau)$.

- Zhao–Atlas–Marks: $\Phi(\tau, \gamma) = \exp(4\pi^2\gamma^2\tau^2/\sigma^2)\cos(2\pi\beta\tau)$, where σ and β are parameters.

Other time frequency representations exist, such as affine class distributions, adaptive optimal kernel functions, reassigned bilinear time-frequency distributions, time-variant higher order spectra, adaptive parametric (and non-parametric) techniques and time-frequency ARMA models. For more in-depth details of these techniques and for an extensive review as applicable to this thesis, please refer to (Feng et al., 2013).

Time-frequency domain technique conclusions

As we have seen, the use of the time-frequency domain overcomes some of the limitations of the frequency domain. In particular, the notion of time is included within the analysis (to some degree of resolution). This enables the power of frequency analysis to be combined in a practical way with the insights gained through traditional time domain analysis. Furthermore, due to this, some techniques are capable of handling non-stationary signals (which was not possible in the frequency domain).

There is often a trade off within the time-frequency domain as to the resolution of each component (as in the STFT). Although some techniques (such as employing bi-linear distributions) have stronger resolution in both domains, there is a trade off. In this case, it is against cross-term interference and computational complexity.

Wavelet analysis is perhaps the most popular tool within the time-frequency domain for condition assessment of components. However, as with all frequency domain techniques, requires the use of high frequency data. As such, these techniques are not possible for the analysis of low-frequency SCADA data.

2.4.5   Wind turbine SCADA data analysis

Due to the low frequency nature of SCADA data, frequency-domain and time-frequency domain techniques are typically not well suited to the analysis of this data. As such, various time-domain techniques have been developed in order to assist in this process. As SCADA data is typically of a different nature to that of high frequency data (for instance, that of bearing accelerometer data) different techniques are often employed. SCADA data can potentially be more beneficial for long term trend analysis over lower frequency data due to less information overload and the lower frequency aiding predictive techniques. SCADA data is expected to trend over longer time frames which would be difficult to observe with high frequency data. For instance, it is unlikely (but not impossible) that a gearbox would catastrophically fail without any degradation or artefact within the data; typically, this process takes a number of weeks or months, and

is potentially easier to analyse with lower frequency data than higher frequency data. As with the high frequency techniques, typically either physical-modelling, data-driven or statistical techniques are employed for the analysis.

Physical models

Physical modelling of phenomena is typical when utilising SCADA data. In Feng et al. (2011), a physical model of the operation of a wind turbine gearbox is presented. This model is then normalised against the efficiency of the gear, allowing for the degradation of gearbox components to be assessed through SCADA data analysis alone. In this model, gearbox temperature and oil temperature are utilised. The paper also identifies oil pressure as an attribute within the SCADA data which can be exploited for gearbox condition assessment.

Similarly, work done by Long et al. (2012) exploits SCADA data in conjunction with a physical model of the gearbox for the analysis of overloading. Their visualisation of the torque histogram allows for comparison of operational behaviour to design standards, to determine if operational conditions are excessively fatiguing components.

The work of Yang et al. (2013) goes one step further and provides insight into variables potentially correlated with failure for ten subsystems of the wind turbine. In this, a physical model is developed to label SCADA data in order to employ statistical techniques whilst allowing for the future use of data-driven techniques. Least squares optimisation is performed to determine correlations of variables, which are then exploited for early detection and potential diagnosis of faults.

Physical failure models are comparatively evaluated in Wilkinson et al. (2013) to both statistical techniques (trending) and artificial neural networks (in their case, self organising maps), and found to be perform the most favourably. Although the authors admit physical models require the most effort in initial setup, these models were able to detect and identify failure better than the other techniques evaluated.

Data-driven techniques

Due to the large quantity of subsystems on the wind turbine, it would be difficult to create physical models of each subsystem and their interactions (hence the motivation of Yang et al. (2013) as discussed above to discover correlations of behaviour based upon known physical models). As such, many data-driven techniques that exploit the large quantity of SCADA data available exist. Amongst the most common of these techniques is the application of artificial neural networks (Priddy and Keller, 2005) for their non-linear mapping capabilities. These are discussed in detail in section 2.5.7. In Kim et al.

(2011), an auto-associative neural network is employed to assist in non-linear principal component analysis, to empirically define normal operational behaviour and suspected fault data taking into account a priori knowledge (failure histories). Clustering is then performed using a self organising map (SOM) as in Kim et al. (2011) to assist in the detection of gearbox faults.

Further non-linear techniques were employed in Wang and Infield (2012) where non-linear state estimation (NSET) techniques were exploited to estimate and predict gearbox temperature. The technique is shown to successfully estimate the temperature under normal operational behaviour. In cases of degradation or damage, the model residuals deviate, and can be used as a means to identify abnormal conditions.

Due to the availability of ancillary information to increase the information content available within SCADA data – namely, SCADA alarm data and also maintenance records – the analysis of these data sources in conjunction with SCADA data has successfully assisted in the early detection of wind turbine faults (Qiu et al., 2012). This is taken further by Chen et al. (2011a), who utilised artificial neural networks to provide non-linear mapping from patterns of SCADA alarm signals to pitch system faults. Ancillary data is also taken into consideration in Godwin and Matthews (2013) for the labelling of historical SCADA data to enable encapsulation of fault models for the purposes of rule extraction.

Data mining approaches have also been considered. Work by Kusiak and Verma (2012b) employed neural networks to predict bearing temperatures as acquired through a higher-frequency SCADA system (1 sample per 10 seconds). As in Wang and Infield (2012) residual analysis is also exploited in (Kusiak and Verma, 2012b) for the detection of faults. In Kusiak and Verma (2012a) a selection of five data mining algorithms are employed on higher-frequency SCADA data (namely; support vector machines, Chi-square automatic interaction detector, boosting tree algorithm, neural networks and random forest algorithm). These are comparatively evaluated for their ability to perform fault detection and identification.

Statistical techniques

Research undertaken by Butler et al. (2013) utilise Gaussian processes for regression and exploit residual analysis for condition assessment of the wind turbine bearing.

Frequency-domain analysis is performed in Zhang et al. (2012) on ten second SCADA data and is shown to be able to detect changes when a support vector machine (SVM) is exploited for state estimation.

Statistical modelling of behaviour is performed in Kusiak and Verma (2013) by the use of the Mahalanobis distance (De Maesschalck et al., 2000) and statistical process

control charts based upon low frequency (10 minute) SCADA data. These are employed on various wind turbine performance curves then exploited to derive a health metric for the wind turbine. Further statistical modelling of performance charts is performed by Uluyol et al. (2011) and is shown to enable early detection of gearbox faults.

Other techniques, such as the use of Bayesian networks (Gelman et al., 2014) have likewise been exploited for wind turbine fault detection utilising SCADA data. In Chen et al. (2012), Bayesian networks are exploited for root cause analysis of wind turbine faults to allow for inherent uncertainty.

SCADA techniques conclusions

Given the often low frequency acquisition rate of SCADA systems on a wind turbine, analysis within the frequency and time-frequency domains is limited. As such, physical models, data-driven techniques and also statistical techniques employing time-domain analysis are often employed.

There is a general consensus that physical models are the most accurate techniques to employ. However, the development of these models takes longer than either data-driven or statistical techniques. Within the use of data-driven techniques, neural networks are the most commonly employed. Whilst these techniques are black boxes, their ability to approximate a non-linear mapping of any function enables strong assessment of condition. Statistical techniques are perhaps the least commonly explored of the three techniques, with performance charts often employed to assist in quantifying the condition of the wind turbine gearbox.

In practice, it is unlikely that a single technique will prove dominant for the analysis of SCADA data, so it is likely that issues with any subsequently developed models will be mitigated using multiple techniques. For instance, it may be that statistical techniques can quantify degradation accurately, but cannot perform prognosis without the assistance of a neural network.

2.4.6 Summary and conclusions

In this section, the techniques which can be employed to derive features to assist with condition assessment from data have been presented. Techniques covering the time domain, frequency domain, time-frequency domain and also SCADA data analysis have been covered.

Time domain techniques were often found to be computationally simple and enable the analysis of both high and low frequency data in real time. Furthermore, it was shown that time domain techniques are able to be utilised in conjunction with non-

stationary signals.

Frequency domain techniques were found to provide an additional level of analysis over time domain techniques, which could include meta-data in order to perform more in depth analysis. However, these techniques often require both stationary signals and also high frequency data.

Time-frequency domain techniques were found to be a favourable alternative to time and frequency domain analysis, overcoming limitations of the frequency domain (such as allowing the analysis of non-stationary signals). However, due to this, a trade-off between time and frequency resolutions exists (as in the case of the STFT or wavelet analysis) or the techniques are too computationally intensive to be performed in real-time (as is the case with many bilinear distributions).

Due to the limitation of both frequency and time frequency techniques for the analysis of SCADA data, alternative techniques were explored. The use of physical models, data-driven and statistical techniques were analysed. Physical models were found to be the most accurate (however, the most difficult and time consuming to develop), with data-driven techniques being the most common within the literature. Often these are utilised to model normal operational behaviour so that residual analysis can be performed to identify deviation from known conditions.

2.5   Machine learning and decision support for early detection, diagnosis and prognosis of asset condition

Having explored in detail the techniques which are available to quantify the current degradation on an asset or component, it is necessary to explore how this data can be utilised in practice in order to drive maintenance decision making and also decision support. Merely quantifying the current behaviour of an asset is not enough to reduce maintenance costs or increase availability. Specific, targeted action based upon the condition of the asset (or component) must be taken in order to realise these benefits. As such, it is of interest to explore the use of machine learning and decision support technologies for the purposes of early detection, diagnosis and prognosis.

In addition to the techniques described in sections 2.4, various hybrid methodologies have been explored. These aim to exploit the strengths of each technique in order to create a stronger condition index or prognostic technique. This is typically synergistic and can provide stronger results than otherwise may be achieved. Many of these techniques are based upon artificial intelligence, however, some techniques simply exploit various characteristics of widely applied algorithms to increase their sensitivity and accuracy.

The techniques presented below provide an overview of the techniques commonly utilised to enhance the development of condition indices and prognostic metrics. Typically, the techniques detailed in section 2.4 are employed as inputs to these techniques, often (but not necessarily) with the addition of contextual information – such as failure histories, maintenance actions or other condition data – to further enhance the derived metric.

For context, a general overview of applicable techniques is presented to the reader in section 2.5.1, with references relevant to the practical application of the research performed in this thesis. Following this, specific techniques and notions of vital importance to the research undertaken in this thesis are presented in detail from section 2.5.2 to section 2.5.6.

2.5.1 Typical techniques

Presented below is a brief overview of associated techniques applicable to the research within this thesis, but which are not directly utilised. They are included for completeness, as their application complements the research undertaken in this thesis and provide an alternative avenue for the exploratory analysis performed in this thesis.

Statistical techniques

Machine learning and decision support within condition assessment have relied extensively on the development of statistical techniques which can be exploited to assist in the justification of maintenance actions or identification of fault or failure modes. It is somewhat intuitive to ask if the current operational behaviour of a component differs from normal operation. Statistical techniques provide a framework to perform this assessment whilst quantifying certainty (or uncertainty) and the significance of these changes (if any exist).

Statistical techniques often employ the use of a null hypothesis (representing a given, or estimated, state or condition) in order to determine if any noticeable deviation from this has occurred. This enables the quantification of changes to be formally measured from the known state whilst utilising the substantial body of research within the realm of statistics in order to demonstrate the validity of the result.

- ***Bayesian networks:*** Bayesian networks are probabilistic models employing Bayes theorem (Nielsen and Jensen, 2009) which allow for the modelling of conditional interdependency and associated probability distributions in a domain (Lee et al., 2014). These employ prior knowledge of the domain to weigh new

observations and update the probability of a given state (such as a failure) being present. Full details of Bayesian networks can be found in Gelman et al. (2014), with Bayesian techniques as practically applied to condition assessment in Gebraeel et al. (2005) and Chen and Hao (2012).

- **Hidden Markov models:** In a hidden Markov model (HMM), a process is assumed to be a Markov process with the addition of unknown state parameters (Lee et al., 2014). As such, this enables the technique to be applied to both stationary and non-stationary signals for fault detection and identification (Dong and Peng, 2011). Specifically, hidden Markov models are a form of Bayesian network (Gelman et al., 2014). Full details of hidden Markov models can be found in Zucchini and Macdonald (2009), with this technique applied to condition assessment in both Dong and Peng (2011) and Chen and Yang (2012).

- **Statistical pattern recognition:** Statistical pattern recognition provides a formal framework for quantifying substantial and significant differences between a known operational state (such as normal behaviour) and a current state (which may be degradation) (Lee et al., 2014). Although typically applied to Gaussian data, robust non-parametric techniques for this also exist (Godwin and Matthews, 2014a). Full details of these procedures can be found in Sheskin (2003), with this technique applied for condition assessment in Cong et al. (2010).

- **Logistic regression:** Logistic regression is similar to classical linear regression, however, it is a classification technique to determine probabilistic membership of a current observation to one of many potential classes. For instance, the probability of belonging to normal behaviour or degraded behaviour. This is performed by finding the model which most accurately describes the relationship between the input variables and output variable (Lee et al., 2014). This requires failure data in conjunction with normal operational behaviour. Full details of Logistic regression can be found in Hosmer Jr. et al. (2013), with the practical application of this technique to condition assessment in both Liao et al. (2006) and Caesarendra et al. (2010).

- **Statistical process control:** SPC is a technique which provides a framework for monitoring future process measurements and for identifying new data that are inconsistent with past data (Lee et al., 2014). Given a mean behaviour and the variance of this behaviour, it is possible to create control charts which identify inconsistent data points. These inconsistencies can be thought of to represent data which are inconsistent of normal behaviour, and thus, represents potentially

degraded conditions. Full details of statistical process control can be found in Oakland (2003), with practical applications of this for condition assessment in Fugate et al. (2001).

- ***Fishers linear discriminant:*** Fishers linear discriminant is a technique which aims to reduce the dimensionality of data whilst retaining the characteristics which are the most efficient at discriminating conditions (Lee et al., 2014). This is a linear technique, and as such is not suited to the analysis of non-stationary and non-linear signals. Full details of Fishers' linear discriminant can be found in McLachlan (2004), with practical application of this technique as applied to condition assessment in Moosavian et al. (2012).

General Models

General models provide methodologies for condition assessment which are domain insensitive, typically non-parametric and often robust. These techniques often associate events over time to a condition indicator, and analyse the effect of these events on the overall system. Typically, these techniques are not as formal as statistical techniques, however, due to their nature can often be employed in situations where the underlying data is not well suited to standard statistical analytical techniques.

- ***Proportional hazards model and survival analysis:*** Survival analysis is a natural paradigm for the condition assessment of both gearboxes and bearings. By associating events which occur to time passing, the current probability of a component surviving for a designated period of time can be determined (Hosmer Jr. et al., 2011). This technique is extensively used in the medical domain for patient life-expectancy calculations (Hosmer Jr. et al., 2011). Typically, a Weibull hazard function is employed; this is beneficial within the realm of condition assessment as Weibull analysis is often performed in reliability engineering. As such many components life-cycles are known to conform to a Weibull model. Non-parametric estimation of the survivor function can be performed using the Kaplan-Meier estimator (Borgan, 2014), with Cox's proportional hazards model (Cox, 1972) employed for analysis of the influence of factors (known as covariates) on the survivor function. These techniques have been applied to condition assessment in Widodo and Yang (2011b).

- ***General path model:*** In Lu and Meeker's seminal paper (Lu and Meeker, May, 1993), the general path model (GPM) was proposed. This aims to move away from "time to failure" and move to "process of failure" (Lu and Meeker, May,

1993). Three basic assumptions are made about the degradation process in the general path model. Firstly, condition histories are selected from a population and the measurement errors of these are independent. Secondly, each condition history is derived from a homogeneous environment (i.e. the same operational and usage conditions). Finally, data is sampled at known, specified time points or intervals (which may or may not be equal) (Lu and Meeker, May, 1993). This model requires a function η which is capable of modelling degradation. This can be a physical model, or derived through the data (Coble and Hines, 2009). A-priori knowledge about the degradation process (domain specific expertise) can be incorporated into the model through the use of Bayesian updating, allowing for degradation paths to be expressed in the absence of data. The general path model has been applied to condition assessment in Coble and Hines (2009).

Logic

Logic provides a natural decision making paradigm for the justification of maintenance actions based upon condition health. These techniques allow inference rules to be generated from the data; for instance *"IF-THEN"* rules such as *"IF* the health metric is < 0.2 *THEN* maintain the asset this week". Logic has been studied extensively across many domains, including philosophy, economics and reliability. Two common logics as explored for condition assessment are:

- *Fuzzy logic:* Fuzzy logic (Zadeh, 1965) provides a framework which allows the quantification of qualitative variables (such as linguistic variables used by mainte- nance operators). Three steps are performed. First, "fuzzification" is performed, which quantifies qualitative variables. Secondly, fuzzy inference is performed (linking input variables to output variables). Finally, "defuzzification" is per- formed (the inverse of fuzzification) (Chekkal et al., 2014). Membership to a fuzzy set ranges from 0 to 1, allowing for probability of membership to be ex- pressed (such as the belief in the membership to the degraded state set). Often, fuzzy logic is used in control strategies, such as in Mesemanolis et al. (2012) and Salomao et al. (2012). This is due to the ability of the technique to handle uncer- tainty (Welte and Wang, 2014). Further detail on the application of fuzzy logic can be found in Ross (2009).

- *Rough sets:* Rough sets differ from fuzzy logic in that class membership is binary (i.e. membership is true or false). This makes rough sets useful for feature selection and for rule extraction (or induction) – and in this context – fault diagnosis (Shen et al., 2000). Fault diagnosis is intuitive under this paradigm,

and has been explored extensively. For instance, the work of Shen et al. (2000) and Sakthivel et al. (2010) both explore the use of rough set theory for fault diagnosis. Further details regarding rough sets can be found in Polkowski (2013).

2.5.2 Suspension histories

Due to the use of failure histories for the purposes of machine learning and decision support, there is substantial interest in exploring the use of suspension histories in order to increase the accuracy of techniques whilst minimizing the error. Suspension histories – also referred to as right censored data – are common in many area of reliability engineering. These are collections of data from an asset or component being brought online until preventive maintenance is performed, and as such, differ inherently from their counterpart; failure histories. Suspension histories are prevalent in industry – with often an order of magnitude more suspension histories being available than failure histories (Heng, 2009; Tian et al., 2010; Widodo and Yang, 2011a) – but are often overlooked for their benefit from a reliability and prognostics perspective.

Traditionally, failure histories are exploited with the assistance of statistical or artificially intelligent techniques to increase metric quality. However, due to the criticality of many assets in practice, preventive maintenance is used to ensure that these critical assets do not fail unexpectedly. Whilst this is more economically effective than allowing the asset to fail and performing corrective maintenance, it is also inefficient as preventive maintenance must be performed more often than is optimally required.

The benefits gained from utilising suspension histories are numerous; due to the quantity of data available for analysis, the issue of "data scarcity" is mitigated and increased statistical power or lower training error can be achieved. However, perhaps the most beneficial characteristic which can be exploited through suspension histories is their assistance in preventing the underestimation of remaining useful life (Widodo and Yang, 2011b).

The rich information contained within suspension histories is of a two fold nature. Firstly, the raw sensor readings which compromise the suspension history, and secondly, the enterprise meta-data which exists as a result. This meta-data can be of various forms:

- **Domain Knowledge:** The suspension history exists due to a preventive maintenance action, and as such, the meta-knowledge derived from the suspension history can be exploited to assist in the development and optimisation of maintenance strategies such as reliability centred maintenance (RCM).

- **External Factors:** Due to preventive maintenance being performed, if this ac-

tion differs from pre-defined business operating procedures, it provides motivation to explore the external factors which necessitated this action (such as overheating due to high ambient temperatures).

- ***Tacit Knowledge Codification:*** If no standard business operating procedure (SOP) is defined for an asset, meta-knowledge extracted from suspension histories can be codified to assist in the development of this process. If assets are preventively maintained (for instance) every 14 days, or 5000 cycles, or when operational temperatures exceed a specified threshold, this knowledge can be extracted and codified into a standard business operating procedure.

Utilising suspension histories in conjunction with artificial intelligence techniques is a relatively new approach and can build upon the use of the previously explored models such as the proportional hazards models (Hosmer Jr. et al., 2011) in traditional reliability analysis. Previously, only the suspension time and quantity of suspensions were considered in this analysis. More recently, various techniques have exploited this rich data source for the benefit of early detection, diagnosis and also prognosis.

Heng (2009) initially explored the concepts of the proportional hazards model in conjunction with an artificial neural network. In this case, suspension histories were utilised to provide a more accurate survival probability. This probability can then be predicted through the use of artificial neural networks. Kaplan-Meier survival probabilities are employed (Borgan, 2014) as this is able to take into consideration right-censoring (such as suspension histories). The technique is non-parametric and well suited to reliability analysis. Lower training error is found when employing both suspended and failure histories.

Tian et al. (2010) builds upon the work of Heng (2009) by utilising artificial neural networks to establish failure histories from suspension histories. Initially failure histories are employed to train the neural network to predict failure times. Using this trained network, suspension histories can then be passed to the network to predict their optimal estimated failure time. This works as the underlying degradation is the same regardless of whether or not the history is from suspension or failure. As the neural network will universally approximate this underlying function, this can effectively be employed to synthetically create failure histories from suspension histories. These synthetic failure histories can then be exploited to further enhance the quality of the remaining useful life prediction. Again, lower errors are found when utilising suspension histories rather than relying on failure histories.

Widodo and Yang (2011a) and Widodo and Yang (2011b) employ a similar approach to Heng (2009) in that Kaplan-Meier survival probabilities are employed for their ability

to utilise right-censored data. In this case, the analysis is performed on high frequency bearing data and exploits the characteristics of support vector machines rather than artificial neural networks. Again, strong early detection was achieved, with lower errors than simply using failure histories alone.

2.5.3 Data balancing

Within the realm of machine learning, data balancing is essential to increasing the performance of derived techniques. Typically, due to the disparity between data scientists and reliability engineers, the use of data-balancing techniques is often neglected when performing machine learning for PHM even though substantial research has been undertaken in this area. As such, these techniques are presented here for completeness.

Data from wind turbines will inherently be biased. Due to the operational requirement of high availability and the resultant preventive maintenance regimes, wind turbines will typically exhibit normal behaviour. As such, the data available for learning degraded or damaged states is limited. For the most common fault on a wind turbine – a pitch system fault (Chen, 2014) – the imbalance can be as high as 1000 normal observations for every 1 observation relating to a pitch system fault (Verma and Kusiak, 2011). This is referred to as "majority class bias" (Batista et al., 2004) and interferes with traditional machine learning techniques.

For instance, an expert system developed to classify SCADA pitch data could achieve an estimated classification accuracy of 99.99% by performing no learning and reporting the default class of "No fault found" (NFF). This would be correct 99.99% of the time (due to the majority class bias), however, it would have failed to have performed the intended function, as it would not be capable of identifying the intended fault. Data balancing ensures this bias is removed and effective learning can take place. The benefits gained through data balancing as specified by Batista et al. (2004) include:

- ***Reduction (or removal) of majority class bias:*** Due to the balancing, no advantage is gained through default classification. As such, the learning which takes place more accurately represents the underlying mechanisms which are being sought after.

- ***Improved decision regions:*** By identifying and removing noisy, erroneous and anomalous observations, the decision regions created by the learner algorithm can potentially more accurately encapsulate the desired behaviour.

- ***Reduced computational complexity:*** By removing noisy, erroneous and anomalous observations from the dataset, the computational requirements of algorithms are reduced.

- ***Higher accuracy and sensitivity:*** By more precisely defining the data for learning, higher accuracy and sensitivity can be achieved.

Various different techniques for data balancing exist. These include both heuristic and non-heuristic methods, with both performing a form of either under-sampling or over-sampling. Random under-sampling and random over-sampling are non-heuristic methods and were defined by Batista et al. (2004) as:

- ***Random over-sampling:*** this is a non-heuristic method which balances the distribution of the classes by randomly replicating observations from the minority class (i.e. degraded data).

- ***Random under-sampling:*** this is a non-heuristic method which balances the distribution of the classes by randomly eliminating observations from the majority class (i.e. normal data).

Whilst these are computationally effective, over-sampling is known to increase the likelihood of over-fitting (Kotsiantis and Pintelas, 2003) and under-sampling has the potential to remove useful information from the dataset (Chawla et al., 2011). Due to this, a variety of heuristic data-balancing techniques have been proposed. These are summarised below:

- ***Tomek links:*** Tomek links (Tomek, 1976) identify data points which are either borderline cases or noise. This is done by defining the distance to the nearest data point: If the nearest data point is the same class as this data point, no action is taken. If the nearest data point is a differing class, then either one sample is noise or both samples are on the decision border between the classes. As such, the technique can be used to balance data (by identifying and removing those points in the majority class which are identified as Tomek links) and to clean data (where any point identified as belonging to a Tomek link is removed).

- ***Condensed nearest neighbour rule (CNN):*** This technique (Hart, 1968) is a form of k-nearest neighbour (k-NN) classifier. CNN finds a consistent subset of points (consistent being that the closest neighbour of all points within the dataset belong to the same class, i.e. can be classified by the 1-nearest neighbour algorithm). Initially, all minority data points and a single majority class data point are selected as a new subset, the 1-nearest neighbour algorithm is then used to classify the remainder of the dataset, with all misclassified data points moved to the new subset.

- **One sided selection (OSS):** OSS (Kubat and Matwin, 1997) is a hybrid methodology utilising both Tomek links and CNN. Initially, Tomek links are employed as an under-sampling technique to identify both noisy and borderline cases. This removes potentially unsafe cases where a small quantity of noise within the data would cause the data point to fall within a different class. The CNN algorithm is then applied to further reduce the imbalance of the dataset by removing the majority class data points which are distant from the decision region (which has been increased due to the use of Tomek links). This is performed to enable machine learning techniques to effectively identify the border region without noise, borderline or extreme cases.

- **Neighbourhood cleaning rule (NCL):** Similar to CNN, NCL (Laurikkala, 2001) uses a nearest neighbour rule to remove majority class examples. In this case, NCL identifies the three nearest neighbours of a given data point. If a given data point is from the majority class, and each the three nearest neighbours are from a different class, the point is removed. Similarly, if a given data point is from the minority class, and each of the three nearest neighbours are from a different class, the three nearest neighbours are removed. This is used mainly for data cleansing.

- **SMOTE:** The SMOTE algorithm (Chawla et al., 2011) (Synthetic minority oversampling technique) creates new data points by sampling between current minority class examples. The idea behind the algorithm is to reduce the imbalance of data by utilising data from the minority class to synthetically calculate data attributes based upon a heuristic. These attributes can then make new data points which populate the dataset and are within expected parameters for the minority class. Various heuristics can be employed to determine the attribute value for the synthetic data point. Some commonly utilised SMOTE algorithms are:

  - *SMOTE* (Chawla et al., 2011) 1 of k-nearest neighbours are selected, with attributes for the new data point being randomly generated in the region between attributes from the selected data point and the current data point.

  - *FSMOTE* (Zhang et al., 2011): This extends SMOTE by employing fractal interpolation theory to generate new synthetic examples. Data points are selected as per the SMOTE algorithm, but attributes of the new data point are exactly half way between the attributes of the seeding data points.

  - *MSMOTE* (Hu et al., 2009): This categorises data into three groups: borderline cases, safe cases and noise cases. No new noise instances are created.

For borderline cases, only the nearest neighbour is employed, and for safe data the k-nearest neighbours are employed. The algorithm then proceeds as per the original SMOTE algorithm

- *B-SMOTE* (Han et al., 2005): Borderline-SMOTE proceeds similar to SMOTE by only synthesising new data points on the decision boundary, as these are more likely to be misclassified.

It should be noted that all of the algorithms discussed for the balancing of datasets require labelled data which must be known a-priori. In practice, with regards to data collected for condition assessment, this information is typically not available or known. This may be because in many cases it is desirable to regress data to predict degradation phenomena occurring, rather than classifying current operational behaviour into a predefined category.

The comprehensive review by Batista et al. (2004) on this subject concludes that in datasets with few positive (minority) examples, they found that utilising the original SMOTE algorithm in conjunction with either Tomek links or the NCL algorithm produced the most consistent results. Interestingly, they also found that random over-sampling produced better results than had been anticipated, and state that when there are large quantities of the minority class points, this technique may produce meaningful results.

Perhaps the most interesting aspect of data balancing to assist in machine learning is the novel use of distance or similarity metrics. All of the techniques which have been discussed here require the identification of an appropriate means to measure the distance between two arbitrary points in a (potentially) high dimensional space. It is clear that different distance or similarity metrics will produce different results when used to assist in the balancing of a dataset. As such, the choice of distance metric and how this distance metric is calculated is essential. Given that the distance or similarity between two points in a high dimension space can be quantified, this also leads to the notion of clustering. As certain subsets of data may be more tightly distributed in specific areas, the identification of the underlying operational behaviour of these areas is of interest.

2.5.4 Cluster analysis and case based reasoning

Cluster analysis is the process of identifying and grouping similar components together into coherent groups (clusters) which are anticipated to share similar characteristics. This is similar to the notion of case based reasoning (CBR), where new problems are solved based upon the solutions derived from previously solved problems which

exhibited similar characteristics.

Clustering of data can take many forms, and many algorithms have been suggested to perform this for condition assessment. Relationships between attributes, as well as attributes between clusters can be taken into consideration for this task. For instance, hard clustering specifies if a data point belongs to a given cluster, whereas soft clustering specifies a degree of membership to a cluster and is more akin to "fuzzy" techniques (Tan et al., 2005). Three main types of clustering exist, these are defined as per Tan et al. (2005):

- ***Strict partitioning:*** In this, each data point belongs to exactly one cluster. This can be extended to allow outliers to be identified. For instance, classifying condition assessment data as "healthy" or "damaged" would be a case of strict partitioning.

- ***Hierarchical clustering:*** In this, each data point belongs to a parent cluster, but may have a subset which specifies a similar (yet distinct) "child" cluster. For instance, race faults may represent a parent cluster, with "inner race" and "outer race" clusters within.

- ***Overlapping clustering:*** In this, objects can be assigned multiple clusters if required. For instance, a bearing may exhibit multiple degradation modes simultaneously; this paradigm would allow for this to be represented.

Clustering techniques can be further defined by their use of chosen similarity metric. Utilising a distance metric typically lends itself well to hierarchical clustering techniques (Kaufman and Rousseeuw, 1990), whereas identifying typical behaviour of a cluster may be performed using centroid models, such as the k-means algorithm (MacQueen, 1967). Other techniques such as distribution models (Xu et al., 1998) and density based models (Ester et al., 1996) also exist.

Within the realm of condition monitoring, density based clustering has typically out performed other clustering techniques. DBSCAN (Ester et al., 1996) (Density-based spatial clustering of applications with noise) assumes different clusters have similar densities (which is often not the case) whereas OPTICS (Ankerst et al., 1999) (Ordering points to identify the clustering structure) allows these different density structures to be clustered effectively.

Clustering provide a means for an implementation of a case based reasoning strategy, and as such, has been explored for fault diagnosis extensively.

- In Kerroumi et al. (2013) RMS was used as an attribute to determine bearing health based upon the DBSCAN algorithm.

- Work done by Lee et al. (2010) shows the practicality of utilising DBSCAN for data-cleansing for PHM.

- Wind turbine gearbox fault identification and detection is enhanced by cluster analysis in Kuo et al. (2012).

2.5.5 Outlier analysis

"Outliers" represent an ambiguous idea which can represent many things. For the purposes of this thesis an outlier can be thought of as defined by Beckman and Cook (May, 1983) in which they state that an outlier is:

> *"data which is inconsistent with the remainder of the dataset"*

or by Hawkins (1980) in that an outlier is:

> *"an observation which deviates so much from the other observations as to arouse suspicions that it was generated by different mechanisms".*

This is intuitive in the domain of condition assessment as it can be thought of that degradation is a different mechanism by which some data – which is outlying – is generated. As such, identification of these outliers may assist in the identification of degradation mechanisms and phenomena.

As seen in section 2.5.4, the notion of similarity is of critical importance to the development of metrics for condition assessment. By defining a metric for similarity, we are able to quantify the distance from a known state (such as normal operational behaviour) to the current state (which may be similar to, or deviate from normal behaviour or degraded behaviour). Typically within the literature, Minkowski distances are employed. The Minkowski distance in an N-dimensional space between points $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$ is defined as per Godwin and Matthews (2014c):

$$D_{Minkowski}(X, Y, p) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \qquad (2.36)$$

Wherein this case n represents the dimensionality of the data and p represents the order of the Minkowski distance. Typically, p is either 2; the Euclidean distance (Godwin and Matthews, 2014c), or 1; the Manhattan distance (Godwin and Matthews, 2014c). The use of multivariate distance metrics is essential within the realm of condition assessment due to the potentially large dimensionality of data (and meta-data) which is often acquired for this purpose.

It can be argued that Minkowski distances are not well suited in the development of condition assessment metrics, as will be critically discussed in section 5.2 of this thesis. It is crucial that any distance metric employed in the multivariate domain is able to take into consideration the variance which inherently exists due to noise in normal operational behaviour, and the interaction between the attributes in a high-dimensional space.

Minkowski distance metrics are not scale invariant. As such weighting of attributes is not equal and is directly dependent upon attribute scale. In order to ensure that equal weighting is given to each attribute, normalisation of the data must be performed. Although many techniques exist as explored in Priddy and Keller (2005), two normalisation forms are common. Namely, min-max normalisation, and z-score normalisation (Kumar and Bhandare, 2011). Min-max normalisation can be defined formally as per Kumar and Bhandare (2011):

$$\hat{\mathbf{x}}_i = \left(\left(\frac{\mathbf{x}_i - \mathbf{X}_{min}}{\mathbf{X}_{max} - \mathbf{X}_{min}} \right) \cdot (\mathbf{K}_{max} - \mathbf{K}_{min}) + \mathbf{K}_{min} \right) \tag{2.37}$$

Where $\mathbf{x}_i$ represents the current observation, $\mathbf{X}_{min}, \mathbf{X}_{max}$ represent the minimum and maximum values of vector $\mathbf{X}$, and $\mathbf{K}_{min}, \mathbf{K}_{max}$ represent constant values to bound the new maximum and minimum values by. This differs from z-score normalisation, which can be defined as per Kumar and Bhandare (2011):

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \bar{\mathbf{X}}_{\mu}}{\bar{\mathbf{X}}_{\sigma}} \tag{2.38}$$

Where in this case $\mathbf{X}_{\mu}$ represents the mean of vector $\mathbf{X}$ and $\mathbf{X}_{\sigma}$ represents the standard deviation of $\mathbf{X}$. This normalises the data around zero.

The identification of outliers enables noise, erroneous data and potential contamination to be removed from the data. As a motivating example, if a suspension or failure history is collected from a bearing or gearbox and outlier analysis is not performed then both healthy and potential degradation (the motivation for the maintenance or cause of the failure) will be present in the data. Subsequently, any technique employed will not be sensitive to the level of degradation present in this data, should this data be utilised for training purposes in data driven models. In order to increase the sensitivity of the technique, it would be essential to ensure no degradation data is present by removing them through the application of techniques developed for outlier analysis.

Further information regarding outlier analysis can be found in Aggarwal (2013) and Rousseeuw (2005).

2.5.6 Rule extraction

Rule extraction is the process of defining a set of rules which represent the underlying data as completely as possible. This may represent a holistic model of the data, or specific local patterns within the data which are of interest. Many algorithms to perform this task have been proposed, such as PRISM (Cendrowska, 1987), C4.5 (Quinlan, 1993) and RIPPER (Cohen, 1995b).

In order to extract meaningful rules, classification labels for the data are required. These are often unavailable, ambiguous or erroneous in practice due to inadequate business standard operating procedures (SOPs), business culture or the the nature of the data itself. As such, without an automatic methodology for labelling data (such as through the use of clustering, outlier analysis, an independent domain expert, or otherwise) the accuracy of this technique will be limited. This is critically discussed in section 2.5.6 of this thesis.

If data classification labels are available, the use of rule extraction techniques can effectively be employed for a variety of tasks, such as:

- **_Expert system development:_** If labels are available with failure data, rules could be extracted which represent various failure modes. This would enable expert system diagnosis to be performed. Similarly, if labels are derived from suspended data, rules explaining early degradation behaviour could be extracted, which would be capable of the early detection of faults.

- **_Tacit knowledge codification:_** If a domain expert provides labels for historical data, rules representing the behaviour of this expert can be extracted. This provides a means to codify tacit knowledge held by the expert; allowing an expert system to mimic this behaviour, and for further in-depth analysis of the behaviour of the expert to be undertaken.

The representation of knowledge can be tacit or explicit. These are defined by Nyame-Asiamah (June, 2009) as:

> *"Tacit knowledge exists in the form of mental models, beliefs, values, assumptions and other know-how of individuals which are not easily conveyed. On the other hand explicit knowledge resides in various forms of artefacts including procedures, texts, reports, memos and books." (Nyame-Asiamah, June, 2009).*

As such, rule extraction provides a means to convert tacit knowledge from an expert – human or otherwise – into explicit knowledge in the form of a set of rules which can be

independently verified and validated. This enables training of individuals within an organisation through codified expert behaviour, and ensures specific skills and techniques are available once an expert leaves the organisation.

As rule extraction provides a means to characterise both normal behaviour and multiple fault modes in a clear, transparent way, it has been used to extract rules from bearing and gearbox data. For instance, in Godwin and Matthews (2014b) rules are autonomously extracted which represent three distinct wind turbine gearbox maintenance levels (normal behaviour, recommend inspection and potential damage).

2.5.7 Artificial neural networks

An artificial neural network (ANN) is a computational models inspired by biological systems which provide a mapping from an input feature space to an output feature space; this mapping is non-linear, and can approximate any function (Priddy and Keller, 2005). Artificial neural networks are employed within the literature for many purposes. This is due to their capability to perform non-linear prediction as well as non-linear mapping. Artificial neural networks are capable of this due to the universal approximation theorem (Cybenko, 1989). Essentially, this states that neural networks can represent a large variety of functions. Mathematically, it is formulated as (Hornik et al., 1989):

$$F'(x) = \sum_{i=1}^{N} \varphi_i \sigma(w_i^T x + b_i) \tag{2.39}$$

Where φ is a bounded, non-constant, monotonically increasing continuous function (such as a sigmoid), α_i, b_i are real constants, N is an integer and w_i an m dimensional real vector. Given I_m denoting the m dimensional hypercube $[0,1]^m$, then given any function f' which is an element of $C(I_m)$ (where C represents the space of continuous functions), then equation 2.39 is an approximation of the realised function f' (where f' is independent of φ) (Hornik et al., 1989). That is;

$$|F'(x) - f'(x)| < \epsilon \tag{2.40}$$

Thus, artificial neural networks can provide an approximate realisation (with the error ϵ) for many functions. If we assume that there exists a function f' that represents the true remaining useful life of a component, we can utilise neural networks to estimate this function and provide insight (diagnostic, prognostic, or otherwise). It should be noted however that whilst this theorem states it is possible to approximate this function, it does not state the characterisation of the network structure required or the capacity

of the network to learn the function. As such, artificial neural networks are capable of performing a vast array of tasks, and have been used extensively in condition assessment for the early identification, classification and prediction of both faults and failures.

Within an artificial neural network, inputs are passed through a single layer (or multiple layers) of neurons which compromise a set of adaptable weights and biases. Training of the network modifies the weights and biases in order to minimise the training error (the error between the expected output and received output). Typically, neural networks are fully connected between layers.

Due to the wide range of capabilities offered by artificial neural networks, many distinct categorisations of neural networks exist. The full categorisation of these network topologies is beyond the scope of this thesis, however for completeness, five common artificial neural network types are presented here:

- **Feed forward back propagation network:** These represent the simplest (and perhaps the most common) neural networks. Information is passed forward, with errors propagated back through the network to minimise error (Priddy and Keller, 2005).

- **Radial Basis function:** This is conceptually similar to a feed forward network, however, traditional activation functions are replaced by radial basis functions (Broomhead, 1988).

- **Self organising map:** Kohonen's self organising maps (SOM) (Kohonen, 1982) utilise the concept of artificial neural networks for the purpose of dimensionality reduction (typically into a two dimensional plane). As high dimensional data is mapped to lower dimensions, this aids the visualisation of data. As high dimensional data is represented in a two dimensional plane, distinct features are placed at distinct locations on this plane. As such, self organising maps lend themselves well to the visualisation of asset condition as in Wong et al. (2006) and Liao et al. (2004).

- **Cascade neural network:** Cascade neural networks (CNN) differ greatly from traditional neural networks. Rather than propagating errors through the network for learning in a fixed topology, the cascade network automatically adds neurons and layers to determine its own structure. Weights are frozen once new neurons are added, and new weights are manipulated to reduce error before proceeding to add additional neurons. This is explained and practically applied to fault diagnosis in Ghate and Dudul (2011).

- **_Auto-associative neural network:_** Auto-associative neural networks (AANN) are similar to feed forward back propagation neural networks, with the additional of a "bottleneck" layer which compresses the data and as such, is often used for feature extraction (Kramer, 1992).

2.5.8 Summary and conclusions

In this section, the core elements and techniques which can be utilised to exploit features for condition assessment were presented. Initially, techniques covering statistical techniques, general models and logics were explored for capability to quantify asset condition when performing condition assessment.

Following this, the use of suspension histories was explored. The use of suspension histories (right-censored data) was identified as being of value in increasing the quality of derived PHM solutions. Within the literature, suspension histories are employed to assist in reducing the training error of supervised techniques such as artificial neural networks. By presenting more data pertaining to the life of a component, a richer data source can be exploited in the training process.

Next, the use of data balancing techniques was explored. As condition monitoring data by its nature is inherently unbalanced (as substantial quantities of data refer to normal operational behaviour), balancing of the data for use in data mining and machine learning is of critical importance. This ensures that no default class bias is present and assists in ensuring that the learning undertaken is valid.

After this, the notions of clustering, case based reasoning and outlier analysis were presented. Each of these techniques utilise the notion of distance (or similarity) in order to classify data according to the rest of the dataset. Clustering is intuitive in many areas of PHM due to a limited number of failure modes, and a large variability in the notion of "normal operational behaviour". By defining regions of behaviours (rather than points), it becomes possible to identify current behaviour based upon known past behaviours (case based reasoning). Similarly, the notion of an outlier is explored with regards to normal behaviour, and the concept of degradation being an outlying behaviour is presented with techniques for identifying these presented.

Finally, both rule extraction and artificial neural networks were discussed with regards to their capabilities to enable PHM. Rules were explored as a means to extract expert knowledge in order to mimic historical operator behaviour. This enables the analysis of expert behaviour for the purposes of either codification (to ensure this knowledge is retained) or auditing (to ensure that operators actions are correct). The uses and common structures of neural networks were explored and shown to be widely applicable within the field of PHM.

2.6   Discussion

This systematic literature review has provided a means to collate currently employed techniques and data sources for the condition assessment of both gearboxes and bearings. This has enabled the identification of many factors which should be considered when moving the state of art forward. The review undertaken provides a basis for the identification of current challenges and limitations of current research whilst guiding the reader to practical applications of the discussed techniques within the context of this thesis.

Firstly, the four data acquisition systems which are typically installed on a wind turbine were presented. The high frequency condition monitoring system, on-demand diagnostic system, structural health monitoring system and SCADA system were discussed in detail with regards to their typical uses, the channels (streams) of data available and the frequency of their acquisition.

Following this, typical techniques for processing and analysing these sources of data were presented. In total, three domains were explored for their potential to quantify degradation. Techniques in the time, frequency and time-frequency domains were explored and analysed. Due to the nature of SCADA data in particular, it was identified that both frequency and time-frequency techniques would not be applicable to this data due to the low frequency nature of the acquisition.

Finally, the use of machine learning for decision support was explored. Techniques including outlier analysis, clustering and distance metrics were explored as a means to identify asset behaviour and determine deviation from known states. Although the quantification of asset condition is essential for a condition based maintenance strategy, this quantification alone is not enough to reduce maintenance costs or increase asset availability. Decisions need to be made – with strong evidence – in order to realise these benefits.

In total, four main points are observed throughout the literature with regards to the current state of the art for condition monitoring. Explicitly, these are:

The use statistical techniques for condition assessment

The use of statistical techniques for condition assessment should be given preference over data-driven techniques. Due to the large imbalance of data in favour of normal operational behaviour, and against degradation (or failure) data, exploiting data mining techniques to learn failure mechanisms severely limits the quantity of data available for analysis as was discussed in section 2.5.3. By exploiting all of the available data for the purposes of defining normal operational behaviour, more accurate encapsulations of

this behaviour can be made. This would allow easier identification of deviation from a known state (such as normal operational behaviour), thus aiding condition assessment by providing early warning.

The use of time domain features for condition assessment

Time-domain features are preferable for the analysis of condition monitoring data in the case of this thesis for two reasons.

Firstly, As detailed in section 2.3, the only two data sources available continuously from the wind turbine are the SCADA system, and the high frequency condition monitoring system. Due to the condition monitoring system reporting data at 50 Hz, the use of frequency-domain (and also time-frequency domain) techniques are limited as the data resolution typically employed for the early detection, diagnosis and prognosis of high speed components is not available. However, it should be noted that time-domain techniques can be applied effectively to SCADA data (as discussed in 2.4.5), and also high frequency condition monitoring data.

Secondly, as frequency-domain techniques potentially may not be capable of non-stationary signal analysis, and time-frequency-domain techniques have either an inherent trade off between resolution in the time and frequency domain or the cross-term interference; time-domain techniques present a potential avenue for analysis without these limitations.

The use of robust techniques for analysis in the presence of noisy data

Robust techniques are essential to the development of health metrics for condition assessment. These techniques require fewer assumptions, may have stronger statistical power (in some cases), and are capable of working effectively on contaminated (noisy) data (Huber, 2011). Within the realm of condition assessment, data will inherently be noisy due to varying operational behaviour, external conditions and degradation of the asset or component itself. By employing a statistically sound, robust methodology for identifying, removing or exploiting these data points, more accurate techniques for condition assessment may be derived.

The use of transparent techniques to quantify asset condition

Transparent techniques for condition assessment may be regarded by some as essential for within the domain of condition monitoring and health assessment. Typically, "black-box" techniques are unaccountable and unproven due to their nature. It is stated by Nusser (2009) that:

> *"common machine learning approaches are regarded with suspicion by do-main experts [...] it is often infeasible to sufficiently interpret and validate the learned solutions. [...] in a safety-critical system it is imperative to guarantee that the learned solution is correct and fulfils all given require-ments".*

When the consequences of an incorrect decision are large – such as a gearbox or bearing failure – it is imperative that decisions can be audited effectively. Work by Bostrom et al. (2013) identifies responsibility, transparency, audit-ability, incorruptibility and predictability as the key issues to consider when deciding upon an artificially intelligent system which is capable of replacing human judgement.

2.7   Conclusions

In this chapter, literature concerning the collection, processing and ultimate exploitation of both high and low frequency data for condition assessment was systematically presented. Analysis of the current literature identified the need to use transparent, time-domain features employing robust statistical techniques for condition assessment.

Although techniques in the time, frequency and time-frequency domain are explored, only those in the time-domain are suitable to both SCADA and condition monitoring data due to the low resolution which would ultimately be attained. Similarly, whilst data-driven techniques such as artificial neural networks have been shown to be capable of accurately modelling non-linear processes (such as degradation), their lack of transparency motivates the necessity for alternative techniques with similar capabilities.

As such, the research undertaken in this thesis will explore these characteristics in particular. The scope of the research undertaken throughout this thesis will now be defined in chapter 3.

Chapter 3

Research Methods

This chapter describes the overall thesis question, research objectives, research questions and the methodologies utilised to undertake the research presented in this thesis. Section 3.1 presents a brief overview of the research conducted for the purposes of this thesis, as well as the overall question that this thesis aims to answer. In section 3.2, the three distinct research objectives of this thesis are discussed in detail with an overview of the motivation for the research and the methodologies utilised. Section 3.3 presents the seven research questions addressed within this thesis which are employed to ensure that the three research objectives presented within this thesis are satisfied. Finally, in section 3.4, the three datasets which are employed throughout this thesis are described in detail, and are provided with sources (where applicable) to enable future continuation of this research.
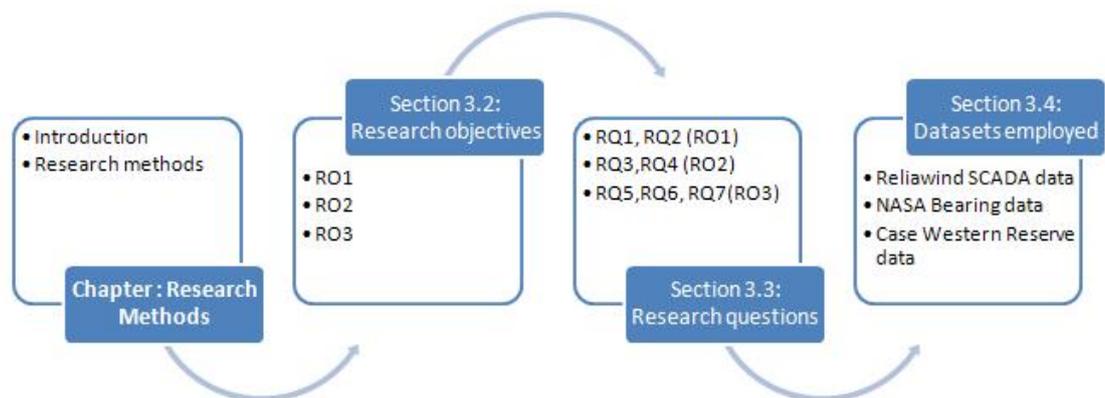
Figure 3.1: Overview of the structure of chapter 3.

3.1 Research undertaken in this thesis

Based upon the extensive and systematic review of the literature undertaken in chapter 2, the necessity to utilise transparent statistical techniques which could be exploited as a metric for condition assessment was presented. As such, given the current use of distance metrics within the literature, various facets of this notion can be explored. Simply put, with regards to multivariate distance metrics for the purpose of condition assessment, little to no exploratory analysis has been performed to determine which features can (or should) be employed within multivariate models, which multivariate distance metrics can (or should) be employed, or how best to exploit the derived condition index. As such, this thesis investigates the suitability of robustly defined multivariate distance metrics and their exploitation for three distinct condition assessment tasks on a wind turbine, namely; early detection, diagnosis and prognosis. Therefore, the overall research question to be explored throughout this thesis can be defined as:

> *To what extent can the use of robust multivariate statistical methodologies and data-driven techniques be employed and exploited as a metric for condition assessment to assist in the early detection, diagnosis and prognosis of both bearings and gearboxes with respect to wind turbines?*

Given the wide scope of this overall thesis question, it is necessary to explore the individual elements of the components which compromise this thesis question. As there are three key elements which must be considered when utilising distance metrics for the purposes of condition assessment (namely, the features which are utilised within the multivariate model, the distance metric employed and also the exploitation of the created condition metric), this leads to three research objectives which are to be explored throughout this thesis. These can be defined as:

> **RO1:** *To explore if the use of time-domain features are appropriate as leading indicators of degradation with regards to both bearings and gearboxes for either low or high frequency data, utilising either statistical, data-driven or physical models.*

> **RO2:** *To explore the use of traditional and robust robust multivariate statistical distance metrics to quantify asset condition.*

> **RO3:** *To exploit any suitably derived condition metric for the early detection, diagnosis and prognosis of both bearings and gearboxes.*

Each of these research objectives are explored in detail in section 3.2. In order to satisfy the three research objectives of this thesis, seven research questions were identified. These research questions enabled the creation of seven distinct experiments which were to be undertaken in this thesis. Each experiment assisted in enabling the research questions to be answered, thus allowing each research objective to be satisfied and the overall thesis question to be answered. The seven research questions identified and answered by thesis thesis are presented and defined in detail in section 3.3.

In the first experiment, the quantification of numerous time-domain features which were beneficial to the early detection of degradation of a bearing based upon high frequency condition monitoring data was performed. Based upon time-domain features identified from the literature, each feature was comparatively evaluated for their effectiveness. In total 28 features were evaluated for their effectiveness at early detection, enabling the creation of a multivariate model which accurately encapsulates the current condition of the bearing.

Following this, exploration was undertaken to develop a physical model of the wind turbine gearbox based upon low frequency data. This complements the research undertaken in the first experiment by providing an additional methodology for an alternative data source and builds upon the work identified within the literature which identified the superiority of physical models for the analysis of low frequency SCADA data. The current limitations of this model were identified and explored within the context of the issues in this thesis, with the identified limitations of this model being mitigated by extending the model to incorporate transient states. An extendible multivariate gearbox model was developed, extending the state of the art. Building upon a validated model from the literature, features specific to the dataset explored within the context of this thesis were taken into consideration; namely, turbine de-rating and curtailment. Next, the model was further extended to take into account transient states.

The third experiment builds upon the multivariate models described and developed in the previous two experiments. In this experiment, an exploration of numerous distance metrics to quantify degradation in a univariate manner is performed. Various different aspects of the metrics are explored to assess their suitability for condition assessment, and to determine if any metric (or category of metrics) was dominant for this purpose. In total, 8 multivariate distance metrics and 1 similarity metric were explored for their potential with regards to condition assessment.

Building upon this, the fourth experiment explores the limitations of employing distance metrics for condition assessment as are explored in the previous (third) experiment. Due to the inherent noise within condition monitoring data, a novel technique was developed which extended the state-of-the-art for encapsulating asset or compo-

nent behaviour. This enabled the autonomous identification of degradation within large datasets, enabling the removal of this data for the purposes of identification of normal behaviour to then be utilised in both supervised and unsupervised learning techniques. Artificial search techniques are explored and comparatively evaluated to heuristic methods and exact methods.

The final three experiments employ the developed methodology which has been refined throughout the prior four experiments. In the fifth experiment, low frequency SCADA data from a wind turbine gearbox is analysed for degradation and quantified by the developed models. Following this, using sound statistical techniques, labels for various levels of degradation (normal operation, recommend inspection and potential damage) are determined based upon well known statistical distributions. Exploiting these autonomously derived labels, data-mining is performed to extract tacit knowledge describing normal operational behaviour, behaviour which warrants inspection and finally behaviour which represents damage to the wind turbine gearbox. The performance of three data mining algorithms are compared and contrasted for this purpose, namely; C4.5, RIPPER and decision tables.

Experiments six exploits the quantified degradation from the methodology derived in the first four experiments. In this experiment, the use of ensembles of robust Mahalanobis distance functions are explored for the purposes of diagnosis. In the final experiment, an alternative non-linear mapping is performed by careful preparation of the data to enable the estimation of remaining useful life (RUL) for the roller element (bearing) data based upon current operational behaviour of the bearing.

3.2   Research objectives

By performing each of the seven experiment sequentially and exploring the results in-depth, a means to answer the three research objectives of this thesis is presented. As stated previously, the work conducted within this thesis has three specific objectives. These objectives have been defined in section 3.1, and will be explored in greater detail here. It should be noted that each of the seven research questions presented in this thesis directly correspond to a specific research objective. These are:

- **RO1:** This is answered by *RQ1 (section 3.3.1)*, *RQ2 (section 3.3.2)*.

- **RO2:** This is answered by *RQ3 (section 3.3.3)*, *RQ4 (section 3.3.4)*.

- **RO3:** This is answered by *RQ5 (section 3.3.5)*, *RQ6 (section 3.3.6)*, *RQ7 (section 3.3.7)*.

Each of these research questions are presented in detail in section 3.3. In order to understand the motivation behind each research question and how the identified experiments are able to assist in answering the overall thesis question, it is necessary to explore the objectives of this thesis in detail. The three objectives of this thesis are explored in sections 3.2.1, 3.2.2 and 3.2.3.

3.2.1 Empirical asset condition assessment models and feature selection

RO1: To explore if the use of time-domain features are appropriate as leading indicators of degradation with regards to both bearings and gearboxes for either low or high frequency data, utilising either statistical, data-driven or physical models.

The first objective of this thesis is to explore in depth the time-domain features which are appropriate for the analysis of both high frequency data (such as diagnostic system data) and low frequency data (such as SCADA data). As explored in section 2.4, frequency and time-frequency domain techniques are not well suited to the analysis of low frequency data. As such, by exploring the capabilities of time-domain features alone, it is hoped that a holistic methodology for the development of a condition assessment metric can be derived for the analysis of both high and low frequency data.

In this thesis, these time-domain features will be explored by analysing high frequency data (20 kHz) containing both healthy and degraded information from roller element bearings. As in this case mechanical systems are explored (as opposed to electrical systems) degradation phenomena should be apparent prior to failure; the features which are consistently able to perform early detection of degradation will then be employed to derive a model from which bearing health can be assessed. If time-domain features are unable to determine significant differences between normal operational behaviour of the bearing and degraded behaviour, this would suggest that time-domain features are not suited for the analysis of high frequency data in this context.

Should time-domain features prove beneficial for the condition assessment of roller element bearings, the exploration of time-domain features will be extended; The use of time-domain features will be explored within the context of low frequency SCADA data collected from a three stage planetary gearbox. If potential degradation can be detected through the use of time-domain features on low frequency SCADA data in the multivariate domain, then multivariate techniques can be explored to assist in the refinement of the derived models. It may be the case that multivariate techniques can increase the quality of the underlying models, thus providing more accurate condition assessment or a means to perform early detection, diagnosis and prognosis.

3.2.2 Robust multivariate distance metrics for asset condition assessment

RO2: To explore the use of traditional and robust robust multivariate statistical distance metrics to quantify asset condition.

The second objective of this thesis is to explore in detail the applicability of an extensive array of multivariate distance (or similarity) metrics for the purposes of condition assessment. Building upon the models and analysis derived from section 3.2.1, it may be the case that multiple time-domain features are required to accurately encapsulate the current operational behaviour of the bearing or gearbox. Due to this – and condition monitoring data representing a multivariate problem (as was concluded in section 2.6) – the multivariate domain provides a natural paradigm for the exploration of condition assessment metrics.

In this thesis, eight multivariate distance metrics and one multivariate similarity metric were explored with context provided by a practical application to both the bearing and gearbox data in order to quantitatively and qualitatively assess their capabilities, limitations and practicality for the purposes of condition assessment. This will verify and validate techniques currently within the literature, whilst exploring previously unverified multivariate distance and similarity metrics. It is hoped that this will enable the quantification of current operational behaviour and the condition of the asset (or component), thus potentially enabling the early detection of degradation.

The exploitation of any multivariate distance or similarity metric for the purpose of condition assessment requires two points in a given space from which to measure the distance between. One of these points will represent current operational behaviour, with the second representing a known state of normal operation. Due to variance in the manufacturing process, initial wear of components and use, "normal operational behaviour" will inherently contain both noise and degradation. Due to this, this thesis explores techniques to cleanse data in order to ensure an accurate encapsulation of normal operational behaviour is achieved. This would increase the sensitivity of the derived condition metric and assist in the reduction of false-negative indications of potential degradation.

3.2.3 Exploitation of multivariate distance metrics for early detection, diagnosis and prognosis

RO3: To exploit any suitably derived condition metric for the early detection, diagnosis and prognosis of both bearings and gearboxes.

The final research objective of this thesis is to exploit the identified multivariate distance or similarity metric in order to provide a holistic methodology for the early detection, diagnosis and prognosis of asset condition. As a multivariate distance metric may be capable of autonomously providing quantifiable degradation information for both high and low frequency data, it may be the case that it can be exploited for many purposes. Although a multivariate distance metric quantifies the current behaviour of an asset or component, without further analysis, the early detection, diagnosis or prognosis of an asset or component is not possible. As such, this research objective aims to exploit the derived metric for condition assessment for these purposes.

Due to the transparent and objective quantification of potential degradation as explored by the previous objective, it becomes possible to apply traditional data-mining techniques to previously unexplored repositories of data. As such, one possible exploitation of the use of multivariate distance metrics is as an 'enabler' of data-mining techniques for the purposes of tacit knowledge codification, early detection, diagnosis and prognosis. Although it is envisaged that the early detection, diagnosis and prognosis are possible under this paradigm, it would be beneficial to exploit the techniques previously discussed in section 2.4 in order to maintain consistency with prior research and assist in the integration of the developed technique into the current body of literature. However, as tacit knowledge codification has not been explored under the techniques explored in section 2.4, this thesis looks to explore and exploit this capability through the application of verified and validated data-mining techniques.

As mentioned in section 3.2.2, determining a point from which to measure normal operational behaviour is critical for the effective application of multivariate distance metrics. By exploring the moderation of this point to represent the initial appearance of various faults, the exploration of fault diagnosis within this paradigm becomes available. As such, this objective will explore the capabilities of multivariate distance metrics in their ability to perform fault diagnosis.

Given that multivariate distance metrics are capable of quantifying asset or component condition in a univariate manner, this lends itself well to traditional time-series predictive techniques which have been explored extensively in many domains. Due to the non-linear nature of degradation phenomena, the exploitation of non-linear time series forecasting techniques provides a means to provide prognostics by predicting future asset condition. Similarly, these non-linear techniques can be exploited to derive a non-linear mapping between current operational behaviour (as assessed by the multivariate distance or similarity metric) and the estimated remaining useful life of the component.

By exploring the capabilities of multivariate distance or similarity metrics for tacit knowledge codification, early detection of degradation, fault diagnosis and also prognosis, this thesis aims to provide a general, holistic prognostics and health management (PHM) methodology for undertaking condition assessment of a given asset or component.

3.3   Research questions and methodologies

Having explored the motivation for the overall thesis question and the research objectives proposed in this thesis, these can be further broken down into research questions. In total, seven research questions are proposed to assist in answering the three research objectives. By identifying the critical elements for analysis within each of the objectives, two key points are observed. Firstly, this enables a sound methodology to be constructed for the analysis of each experiment to enable this analysis. Secondly, by exploring the critical elements of each objective, quantifiable hypotheses for each objective can be assessed in order to ensure that the research question is answered. The seven questions which will be answered by this thesis are:

*RQ1: Are time-domain features beneficial to assisting in the empirical quantification of bearing degradation based upon high frequency accelerometer data? If so, which features differ significantly and substantially as a result of degradation? (**RO1**)*

*RQ2: Is it possible to normalise a non-stationary signal based upon low frequency gearbox data from the influence of loading conditions, transient states and external factors? If so, can this be empirically derived? (**RO1**)*

*RQ3: Can the use of multivariate distance or similarity metrics be exploited to empirically derive a metric for condition assessment? If so, which metrics are the most sensitive to changes in asset health? (**RO2**)*

*RQ4: Can the use of AI search techniques be employed to assist in deriving a robust encapsulation of the data which can be employed in conjunction with distance metrics? If so, how do these techniques compare to current exact or heuristic techniques? (**RO2**)*

*RQ5: Can robust multivariate distance metrics be used as a means for rapidly labelling data for the development of expert systems and extraction of tacit knowledge from historical data repositories? If so, what level of accuracy and agreement can be attained? (**RO3**)*

*RQ6: Can the application of robust multivariate distance or similarity metrics be employed for the diagnosis of developing bearing faults? If so, how can this be performed, and with what accuracy can fault diagnosis take place? (**RO3**)*

*RQ7: Can the application of robust multivariate distance or similarity metrics be exploited by non-linear techniques for either the prognosis of asset condition by future condition prediction or remaining useful life estimation? If so, how accurately can these be performed? (**RO3**)*

Each of the seven research questions presented here are discussed in detail in sections 3.3.1 through to 3.3.7. The motivation for undertaking the research required to answer each question is presented, along with a subset of research questions which must be answered to effectively answer the overall research question. Once each relevant research question is answered, the overall research objective which the research questions correspond to can be answered. Similarly, once each research objective has been answered, it is possible to answer the overall research question.

As such, it is essential that each research question is sound and assists in answering the overall research objective. For this purpose, the motivation for each research question and the necessary sub-questions proposed to answer the research question are now presented.

3.3.1 High frequency time-domain features for condition assessment

*RQ1: Are time-domain features beneficial to assisting in the empirical quantification of bearing degradation based upon high frequency accelerometer data? If so, which features differ significantly and substantially as a result of degradation? (**RO1**)*

Motivation

In order to create a model which could be exploited for condition assessment, relevant features which accurately encapsulate the current operational behaviour of the component (such as bearing or gearbox) are required. Due to the application specific nature of many of the techniques discussed in section 2.4, no "off-the-shelf" feature set typically exists for developing condition assessment metrics. As such, it was essential to explore numerous time-domain features which could be utilised to develop a model which accurately characterised current operational behaviour.

The purpose of this research question was to establish a multivariate model for

high-frequency data which would enable the analysis of asset or component condition. If univariate models representing asset or component condition can be shown to in-adequately encapsulate the condition of a component, this serves as a basis to explore analysis within the multivariate domain as a potential means to overcome the limitations of univariate techniques. Thus, by exploring univariate features which individually may provide some additional context to the problem, the multivariate domain may enable this to be exploited through the use of multivariate distance metrics.

Sub-questions

In order to assist in answering the first research question, a number of additional questions were proposed:

1. What statistical features exist within the time-domain for the analysis of condition?

2. Of these statistical features, which are beneficial for the assessment of condition?

3. Are certain categories of features more beneficial than others for the assessment of condition? Such as:

   (a) Measures of central tendency.

   (b) Measures of variability.

   (c) Measures of shape.

   (d) Measures of position.

   (e) Measures of impurity.

4. Which of the identified features are capable of identifying degradation at the earliest onset?

Methodology

In order to answer the first question, a systematic review of the literature was performed to identify statistical features which could be exploited within the time-domain for the purposes of condition assessment. Following this, the second and third questions were answered by a full statistical analysis of the identified features. This was performed on high-frequency roller element bearing data which contained normal operation behaviour, suspension histories and also failure histories, allowing for a comparative evaluation of all of the identified features for the purposes of condition assessment. Data pertaining to normal operation behaviour was compared and contrasted to that prior

to failure to quantify any significant (or substantial) changes to the identified features. The final question was then answered by pursuing more in-depth statistical analysis, seeking to identify the earliest point within the data at which each feature differed significantly from normal operational behaviour, allowing for a relative ranking of each of the identified attributes. Due to the nature of typical condition monitoring data, a robust statistical process control (SPC) methodology utilising exponentially weighted moving average (EWMA) charts was employed for this analysis.

3.3.2 Low frequency time-domain techniques for condition assessment

RQ2: Is it possible to normalise a non-stationary signal based upon low frequency gearbox data from the influence of loading conditions, transient states and external factors? If so, can this be empirically derived? (RO1)

Motivation

Building upon the research undertaken as part of research question 1 (RQ1), a natural extension becomes apparent. Whilst the previous model identifies features which can be successfully applied for condition assessment in a high-frequency signal which is stationary in nature, this has the potential to not be applicable to low-frequency data, specifically when the low-frequency data represents a non-stationary signal (such as that of a wind turbine gearbox). As such, it is of interest to explore and identify techniques and the characterisations of low-frequency signals so that methodologies such as those identified in the previous research question can be employed on alternative data sources.

The purpose of this experiment was to encapsulate the behaviour of the wind turbine gearbox in a physical model. This would allow for the development of an extensible model which took into consideration many aspects of the gearbox; this is essential due to the quantity of data acquisition devices due to be employed in next generation wind turbines as discussed in section 2.3.

Sub-questions

In order to assist in answering the second research question, a number of additional questions were proposed:

1. What models exist for wind turbine gearbox condition assessment?

   (a) Are these based upon high-frequency SCADA data?

   (b) Do any models currently exist based upon low-frequency data?

2. Can these models take into consideration (or be extended to incorporate) factors
   such as:

   (a) Loading of the gearbox.

   (b) Turbine de-rating or curtailment.

   (c) Transient states.

   (d) External factors.

3. Which of the identified or created models can be utilised for analysis in the multi-
   variate domain, specifically with regards to being employed within a multivariate
   distance or similarity metric?

Methodology

In order to answer the first sub-question proposed by this research question, a system-
atic review of the literature was required. This identified currently available techniques
and models for the condition assessment of non-stationary signals which are acquired
from the wind turbine gearbox. This also identified the nature of the data employed in
these techniques or models – either high or low frequency – for the purposes of condi-
tion assessment. Due to the critical nature of the wind turbine gearbox as discussed in
section 1.5.1, physical models of degradation and failure will be given preference over
data-driven and statistically based models for the reasons discussed in section 2.4.5.

The second question complements the prior sub-question and can be answered
through a comparative analysis of the techniques or models identified throughout the
systematic review. Should no technique or model prove viable for normalisation against
turbine loading, transient states, de-rating, curtailment or external factors, then the
creation of a proprietary multivariate model would be necessary. The answer to the
final question arises as a direct result of the comparative analysis between existing
models, techniques and the chosen model.

3.3.3 Multivariate distance metrics for asset degradation quantifica-
tion

*RQ3: Can the use of multivariate distance or similarity metrics be exploited to em-
pirically derive a metric for condition assessment? If so, which metrics are the most
sensitive to changes in asset health? (**RO2**)*

Motivation

The results of research questions one (section 3.3.1) and two (section 3.3.2) provided motivation to explore the use of multivariate distance metrics for the purposes of condition assessment. As in many cases, univariate models are incapable of the accurate encapsulation of all aspects of asset or component behaviour, the use of multivariate models can potentially overcome this limitation.

However, in a multivariate model it may be difficult to understand the interactions between all of the attributes. As such, the assistance of a distance or similarity metric would enable the benefits attained through employing a multivariate model whilst presenting the notion of asset health in a univariate way. If multivariate metrics are found to be able to accurately and precisely perform this function, this provides an avenue to enable currently employed univariate models to be superseded by multivariate models – thus potentially giving earlier and more robust assessment of condition – without a substantial change to current PHM practices, techniques or infrastructure.

Sub-questions

In order to assist in answering the third research question, a number of additional questions were proposed:

1. What multivariate distance or similarity metrics are available?

2. Of these metrics, which have been explored within the context of condition assessment? Which have not?

   (a) Of those which have not been explored within the context of condition assessment? Why is this the case?

3. Are certain categories of metrics better suited than others for the assessment of condition in the multivariate domain?

   (a) Are certain metrics more suited to high frequency data?

   (b) Are certain metrics more suited to low frequency data?

4. Which of the identified metrics are easiest to compute?

5. Which of the identified metrics are most sensitive to degradation?

6. What are the limitations of using multivariate distance or similarity metrics for condition assessment?

Methodology

As per research questions one and two, the systematic review of the literature undertaken in chapter 2 was required to determine currently available multivariate metrics which could be employed for the assessment of either "distance" or "similarity". This would answer the first question and also assist in the answering of the second.

The second and third questions can be effectively answered through a comparative analysis within the context of condition assessment. Exploiting the models derived through the research conducted in this thesis to answer research questions one and two, the identified metrics can be compared and contrasted in their application to both high frequency (accelerometer) and low frequency (SCADA) data.

The fourth and fifth questions form part of the trade off analysis which is required when undertaking condition assessment. It may be the case that a metric is superior, however, cannot be computed effectively in real time. This would limit the application of the metric. Complexity analysis is sufficient to determine which metrics can or cannot be computed in real time for a given computational capacity, with a similar analysis undertaken to determine the most sensitive metric explored.

The final question is partly answered as a result of the trade off analysis performed to answer questions four and five, with the assistance of the systematic literature review in identifying limitations of these techniques beyond the scope of this thesis.

3.3.4   Encapsulation of normal operational behaviour

RQ4: *Can the use of AI search techniques be employed to assist in deriving a robust encapsulation of the data which can be employed in conjunction with distance metrics? If so, how do these techniques compare to current exact or heuristic techniques? (**RO2**)*

Motivation

Having established the pre-requisites to employing multivariate metrics for condition assessment in the previous research questions, it is essential that the metric is utilised in the correct manner. In the first research question, a sound set of features for condition assessment for high frequency bearing accelerometer data were established. In the second research question, an extensive physical model for low frequency wind turbine gearbox SCADA data was explored. Following this, in the third research question, the exploration of multivariate metrics (which are both sensitive to changes in behaviour and also capable of being calculated in real time) was performed. For a metric, the distance (or similarity) will be determined between two points. One of these points will represent the current operational behaviour. The second, will then represent a known

(or estimated) behaviour. Ensuring that this second point accurately and precisely encapsulates normal operational behaviour is key to ensuring that the metric performs as anticipated.

If no processing of collected data is performed prior to setting this reference point representing known behaviour for the metric, then noise and potential degradation will inherently be contained within the data. This contamination will reduce the sensitivity of the multivariate metric, as it will introduce noise into the calculation. The metric will be unable to assist in the identification of degradation which is of a lesser extent than that which was contained within the dataset utilised to determine the normal operational behaviour. As such, ensuring that no degraded data, noise or contamination is contained within the data is essential.

In this thesis, the exploration of the possibility of using artificially intelligent search algorithms to assist in the identification of degraded, erroneous and noisy data within signals is undertaken, so that this can be isolated to increase the sensitivity of the multivariate metrics employed for condition assessment.

Sub-questions

In order to assist in answering the fourth research question, a number of additional questions were proposed:

1. What techniques exist to identify normal behaviour?

    (a) How much data is required to exploit these techniques?

    (b) Are these capable of working in the multivariate domain?

    (c) Do these rely on heuristics to assist in the identification of normal behaviour?

2. Which of these techniques have been explored within the context of condition assessment?

3. Can AI search techniques assist in the identification of normal behaviour?

    (a) Which techniques are well suited to this?

    (b) Which techniques are not suited to this?

4. How computationally intensive are these techniques?

    (a) Can these techniques be performed in real time?

Methodology

Similar to the prior research questions, initially a further review of outlier analysis and artificially intelligent search algorithms was utilised. This enabled the identification of available techniques, limitations in the area and potential avenues of exploitation for this purpose. By cross referencing literature identified in this search with literature identified in section 2.4, techniques already exploited for the purposes of condition assessment could be identified. This would enable sub-questions one through three to be answered.

Sub-question four is answered through a comparative analysis within the context of condition assessment. By applying a selection of techniques identified in sub-questions one through three, the limitations, capabilities and practicalities of the techniques could be explored. This would assist in quantifying the trade off which exists between the algorithm selection.

3.3.5   Tacit knowledge codification

RQ5: *Can robust multivariate distance metrics be used as a means for rapidly labelling data for the development of expert systems and extraction of tacit knowledge from historical data repositories? If so, what level of accuracy and agreement can be attained? (RO3)*

Motivation

Given the capability of a metric to quantify the current operational behaviour of an asset or component, it is possible to exploit this for the purposes of retrospectively data-mining historical repositories of data in order to extract meaningful information or knowledge from this data. This would enable the analysis of vast quantities of data which have been previously unexplored due to the limitations of manually labelled data, or the economic constraints which would otherwise exist without an algorithmic approach to this process.

Manual labelling of historical repositories typically requires strong assumptions to be made. These assumptions enable the labelling of the data, but often introduce noise or uncertainty within the data. For instance, in order to identify "degradation" one may assume degradation occurs one day prior to failure. This would enable supervised data-mining techniques to be employed to identify degradation, however, the notion of degradation in this context would be highly subjective as it is based upon the supervised labels (which may – or conversely, may not – be adequate).

As such, the purpose of this experiment is to explore the use of metrics with known

distributions to enable the objective, autonomous and statistically sound labelling of the data for the purposes of data-mining.

Sub-questions

In order to assist in answering the fifth research question, a number of additional questions were proposed:

1. What algorithmic techniques exist to codify tacit knowledge?

2. Which of these techniques have been explored within the context of condition assessment?

3. What is the trade off which exists between these techniques?

4. How can we validate the codified knowledge?

   (a) What metrics exist for this purpose?

   (b) Can domain experts be utilised to assist in this process?

Methodology

Upon performing the systematic review to determine the various techniques which exist to answer the first sub-question, a comparative analysis was required to compare and contrast the codified knowledge to assist in answering the second question. Different techniques will present different knowledge in different formats – for instance, as a table, tree or list of rules – and as such, each technique will have its own capabilities and limitations.

   In order to quantify the agreement between the metrics employed for the labelled of the data and the data-mining techniques which exploited these labels, various metrics exist. For instance, Cohen's κ statistic (Cohen, 1960) can be employed for this purpose. Once the level of agreement between the models created throughout this thesis and the derived data-mined models are understood, the learnt models can be presented to domain experts for analysis, critical review and interpretation.

3.3.6 Fault diagnosis

RQ6: *Can the application of robust multivariate distance or similarity metrics be employed for the diagnosis of developing bearing faults? If so, how can this be performed, and with what accuracy can fault diagnosis take place?* (***RO3***)

Motivation

Given the capability of robustly defined multivariate metrics to perform asset or component condition assessment, the natural extension of this is to perform diagnostics for a given asset or component. Whilst the identification of anomalous behaviour may suffice in some domains, in others, information regarding the location, type and severity of the fault is required. This distinction highlights the difference between fault detection ("early detection") and fault identification ("diagnosis").

   If the failure mode is known prior to the failure, steps can be taken to mitigate the impact – financially, or otherwise – of the resulting maintenance action or potential failure. For instance, if degradation is identified on a bearing, it may be possible (depending upon the anticipated or observed failure mode) to purchase replacement components in advance. This would reduce downtime, and limit the spare inventory required to be held. For instance, if it is anticipated that a roller element (ball) within a bearing will degrade, a replacement component can be purchased in advance, and the replacement performed at the next scheduled maintenance event, before catastrophic failure of the asset or component occurs.

Sub-questions

In order to assist in answering the penultimate research question, a number of additional questions were proposed:

1. Are the identified time-domain features from RQ1 beneficial for diagnosis?

   (a) Do these differ significantly and/or substantially?

   (b) Are the previously explored features capable of diagnosis or just early detection?

   (c) Are certain categories of features better suited to diagnosis than others?

   (d) What data is required in order to perform this analysis?

2. Are multivariate metrics capable of identifying failure modes rather than normal behaviour?

   (a) With what diagnostic accuracy can this be performed?

   (b) What coverage of failure modes can be achieved in this context?

Methodology

To answer the first sub-question, we can build upon the research performed in RQ1 as discussed in section 3.3.1. The systematic review which was undertaken in chapter 2

identified various time-domain features which could be exploited for condition assessment and ultimately the early detection of degradation. By exploring these features in the context of diagnosis, the capabilities and limitations of these time-domain features can be identified. It may be the case that the subset of features identified as beneficial for the early detection of degradation are not beneficial – or may even hinder – the diagnostic process. As such, a full statistical analysis of the employed features is essential to identify those which are capable of discriminating between the different failure modes which exist due to degradation.

To perform this statistical analysis, data containing faults representing each failure mode must be exploited. Although it would be possible (in theory) to perform this analysis in real-time as faults occur, in order to verify the technique, it is beneficial to utilise data which is known to contain pre-existing faults. Online repositories of seeded fault data – such as that made available by Case Western Reserve University (2011)– are available for analysis, and are exploited in this experiment in order to perform diagnosis. To perform the statistical analysis, analysis of variance (ANOVA) was utilised, with both traditional and robust post-hoc analysis being performed (as required) to identify which features are capable of strongly discriminating between failure modes.

Once a diagnostic model has been verified, the second question can be explored by exploiting the research undertaken to answer the research questions RQ4 and RQ5 proposed in sections 3.3.4 and 3.3.5. This would enable the analysis of the capabilities of exploiting a robust multivariate distance metric for the purposes of diagnosis. It should follow that if features differ significantly amongst failure modes, then setting a given point of a distance metric to these previously known failure modes, it may be possible to exploit multiple metrics to perform diagnosis as well as the early detection of degradation.

3.3.7 Prognosis

***RQ7:** Can the application of robust multivariate distance or similarity metrics be exploited by non-linear techniques for either the prognosis of asset condition by future condition prediction or remaining useful life estimation? If so, how accurately can these be performed? (**RO3**)*

Motivation

Building upon the research undertaken to answer the research question proposed in section 3.3.6, it is natural to extend the analysis into the realm of prognosis. Although being able to anticipate or identify the failure mode which is expected to occur (or

already may have occurred), this is unable to provide a notion of time which is crucial for proactive (or preventive) maintenance strategies.

Prognosis would enable the estimation of the future condition of an asset or component, or enable an estimation of the remaining useful life of a component. This is valuable business intelligence which can be exploited to significantly reduce maintenance costs. As such, providing means for prognosis within the paradigm explored in this thesis is crucial.

The purpose of this experiment was to devise a methodology to exploit the health metric derived throughout this thesis in order to perform both asset condition prediction and also remaining useful life prediction based upon non-linear time series forecasting techniques. This would enable all aspects of a condition based maintenance (CBM) strategy or PHM system to be explored under the methodologies presented in this thesis. Thus, this thesis would present a set of methodologies which would enable health monitoring for the early detection, diagnosis and prognosis of any asset or component which warranted this analysis.

Sub-questions

In order to assist in answering the final research question, a number of additional questions were proposed:

1. What techniques are commonly utilised for prognosis currently within the literature?

    (a) What are the capabilities of these techniques?
    (b) What are the limitations of these techniques?

2. What training error can be achieved for remaining useful life estimation? How much data is needed?

Methodology

The first two sub-questions can effectively be answered in a similar manner to those of the previous research questions; by the aid of the systematic review undertaken in chapter 2. This would identify the range of techniques which have previously been exploited for condition prediction or remaining useful life estimation. Identification of non-linear techniques is essential due to the typically non-linear nature of degradation phenomena. Although linear techniques may have the potential to predict future asset or component condition or remaining useful life, it is anticipated that linear techniques will perform inadequately at these tasks.

Upon identification of applicable techniques for the prognosis of asset condition or estimation of remaining useful life, a comparative evaluation of the different techniques which have been identified can be performed. This would enable the third and fourth research sub-questions to be answered. By applying the techniques within the context of condition prediction or remaining useful life estimation, the error of these techniques and prognostic horizon of the technique can be attained.

Once these are known, the capabilities of the techniques and limitations can be discussed within the context of this thesis and condition monitoring in particular.

3.4 Datasets employed in this work

In order to derive an index or metric which can be exploited for the purposes of condition assessment, empirical evidence of the validity of the developed techniques must be presented. In this case, this is obtained though the application of the techniques developed within this thesis to historical failure data. As such, the availability of data (high frequency condition monitoring data, low frequency SCADA data or otherwise) is integral to the work in this thesis.

In order to effectively answer the seven research questions proposed throughout sections 3.3.1 to 3.3.7, three datasets were employed. These were:

- *High frequency roller element bearing accelerometer data:* Required to develop remaining useful life estimators, feature selection for the early detection of degradation and to validate the application of multivariate distance metrics for condition assessment.

- *Low frequency wind turbine gearbox SCADA data:* Required to create a low frequency gearbox model, to verify and validate the application of multivariate distance metrics and to derive expert systems based upon traditional data-mining techniques.

- *Seeded fault high frequency bearing accelerometer data:* Required to assist in verifying the developed bearing model, and to extend the analysis to diagnosis.

Each of these datasets are discussed in detail in sections 3.4.1 through to 3.4.3 covering both data acquisition and the experimental setup.

3.4.1 Reliawind data

Due to the University of Durham being a member of the "Reliawind" consortium (Wilkinson et al., 2010), the proprietary data contained within this repository was

made available for the purposes of this thesis. This data repository contains both high and low frequency data taken from numerous wind turbines. For the purposes of this thesis, in total, the data from 6 wind turbines across a single wind farm over a period of 28 months was available to be utilised. As this thesis is primarily concerned with exploiting low-frequency data, data acquired through the SCADA data system is discussed here. Of these 6 turbines, one suffered a catastrophic gearbox failure during the period of data acquisition, and was subsequently replaced.

With regards to low frequency data, in total, 190 channels of SCADA data across 11 subsystems of the wind turbine were available. The SCADA data system reports data at a 600 second (10 minute) interval for each channel. This corresponds to 6 data points per hour or 144 data points per day per turbine. Typical data collected from the SCADA data systems includes (but is not limited to):

- Wind speed and direction.

- Ambient temperature and humidity.

- Power production.

- Gearbox and oil temperature.

- Blade pitch and turbine yaw.

Similar to the SCADA data system, the corresponding SCADA alarm data was available. This system polls data on a second-by-second basis (1 Hertz) in order to notify the operator or owner of a wind turbine of a potential fault or issue. This separate system reduces the quantity of data required for analysis; if no alarms are present, there is no necessity for inspection of the current state of the wind turbine. On the other hand, if alarms are present, it provides a basis to inspect the SCADA, high frequency condition monitoring or on demand diagnostic systems in order to confirm or deny the presence of the fault (or error) which the alarm corresponds to.

The final data source available for analysis within the Reliawind dataset is that of the maintenance logs. These are typed statements from the maintenance operators corresponding directly to the work undertaken. Data pertaining to inspections, maintenance, failures and repairs are collected within these statements. The turbine in question, date, work undertaken (including the components changed and turbine subsystem in question) are available in this repository.

3.4.2 NASA bearing data

The high frequency bearing dataset utilised in this thesis was made available through NASA and collected by Rexnord technical services (Lee et al., 2007). This dataset pertains to the run-to-failure of three experiments containing four bearings each.

As can be seen in figure 3.2, the experimental setup consisted of a test rig holding four bearings on a single shaft with a 6000 lbs radial load. This was driven by an AC motor at a rate of 2000 RPM, coupled together with rub belts. Each of the four bearings were force lubricated, with computerised control of the flow and temperature of the lubricant. Debris was collected in the lubricant feedback pipe via a magnetic plug; this enabled evidence of degradation to be collected. The test was deemed concluded when the debris within the oil had amounted to a predetermined level which caused an electric switch to close.

The bearings employed within this data were Rexnord ZA-2115 double row bearings. Each row contains 16 rollers with a roller diameter of 0.331 inches and a pitch diameter of 2.815 inches. The tapered angle of contact was $15.17°$. For the purposes of data collection, each bearing was monitored via an accelerometer (PCB 353B33 high sensitivity quartz ICP ®). In addition to this, a thermocouple was attached to each bearing for use in the computerised the lubrication system. Data was sampled periodically (either at 5, 10 or 20 minute intervals depending upon the setup) though a
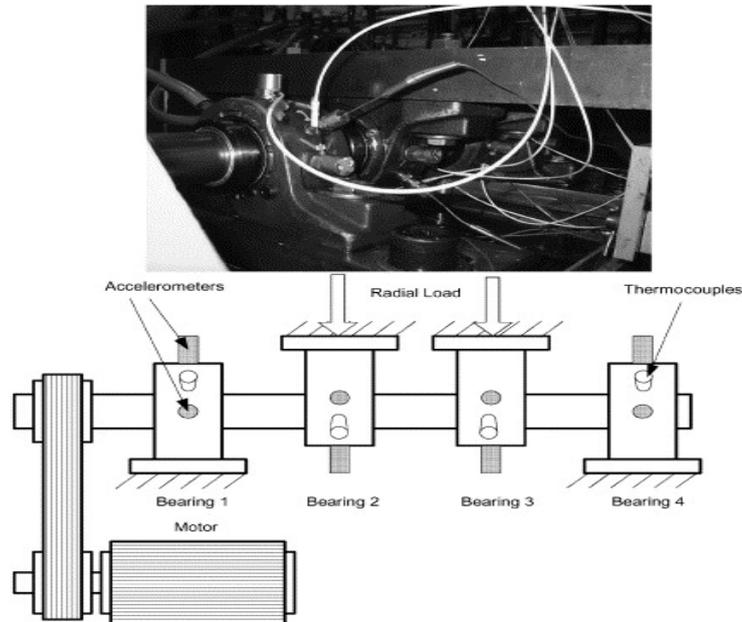
Figure 3.2: Experimental setup for run to failure bearing data acquisition (Lee et al., 2007).

National Instruments data acquisition device (DAQ Card-6062E). Data was collected at 20 kHz, with the data length equalling 20,480 data points.

In total, three run to failure experiments were performed between October 2003 and April 2004. These are described below:

- **Experiment 1:** In this experiment, eight channels of high frequency data were recorded every 10 minutes for one second over 34 days resulting in 2,156 data points representing the X & Y axes of four bearings. At the end of this test, an inner race fault had occurred in one of the bearings, with a roller element (ball) fault in another.

- **Experiment 2:** In this experiment, four channels of high frequency data were recorded every 10 minutes for one second over 6 days resulting in 984 data points. At the end of this test, an outer race fault had occurred in one of the bearings, with no other faults detected on the remaining three bearings.

- **Experiment 3:** In the final experiment, four channels of high frequency data were recorded every 10 minutes for one second over a period of 31 days resulting in 4,448 data points. At the end of this test, an outer race fault had occurred in one of the bearings, with no other faults detected on the remaining bearings.

3.4.3   Case Western Reserve University seeded fault dataset

In order to exploit health monitoring models for diagnosis, it is essential to correlate the propagation of degradation to failure modes. When developing these models, the verification and validation of their diagnostic capability can be attained when historical data containing coverage of known failure modes is available. As such, this thesis exploits publicly available high frequency bearing data with seeded failures for this purpose made available by Case Western Reserve University (2011).

The test rig employed for data acquisition in this case consisted of a 2 horse power motor, a torque transducer/encoder and a dynamometer, with the bearings supporting the motor shaft. This is shown in figure 3.3. Faults were seeded into the bearings using electro-discharge machining. This allowed for the control of both the fault size and also fault location. In total, five differing diameters of faults were seeded; .007 inches, .014 inches, .021 inches, .028 inches and .040 inches. Depths of the faults varied between .011 inches and .150 inches dependent upon the failure mode, diameter of the seeded fault and the location of the bearing. For the three smaller diameter faults, SKF bearings were employed. In this case, these were 6205-2RS JEM bearings, with the following characteristics:
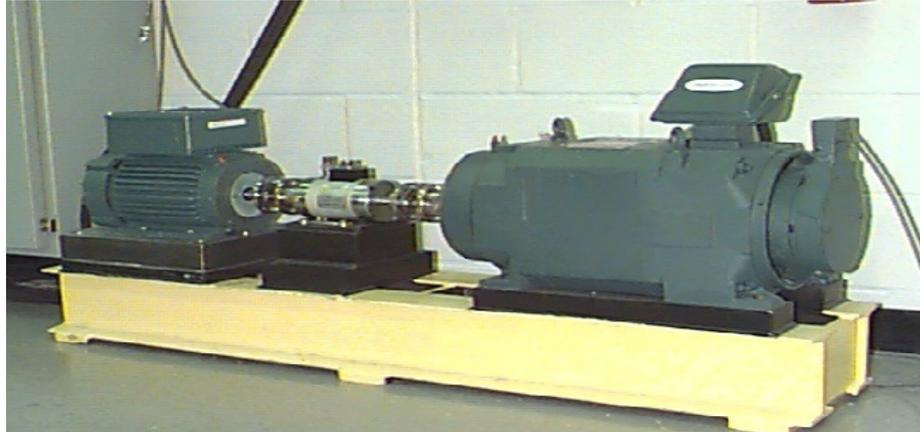
Figure 3.3: Experimental setup for seeded fault bearing data acquisition (Case Western Reserve University, 2011).

1. **Inside diameter:** .9843 inches

2. **Outside diameter:** 2.0472 inches

3. **Thickness:** .5906 inches

4. **Ball diameter:** .3126 inches

5. **Pitch diameter:** 1.537 inches

Accelerometers were attached magnetically to the base of the housing at the 12 o'clock position at the drive end of the motor housing to collect vibration data. Data was collected though a 16 channel DAT recorder at a rate of 48,000 samples per second and processed in Matlab. The ancillary data of both horse power and speed were collected via the torque transducer/encoder and recorded manually.

3.5 Chapter summary

In this chapter we have introduced the research to be undertaken in this thesis. The overall question to be answered within this thesis was presented, which is:

> *To what extent can the use of robust multivariate statistical techniques be employed and exploited as condition assessment metrics to assist in the early detection, diagnosis and prognosis of both bearings and gearboxes?*

Following this, the three overall research objectives (RO's) of this thesis which are required in order to answer the overall aim of this thesis were presented. This was subsequently broken down into seven questions (RQ's) which could be answered through experimentation. Each question was described in detail, providing the motivation for undertaking the research, the sub-questions which are required to answer the research question and a brief overview of the methodology employed in order to undertake the research.

Finally, the three datasets employed within this thesis were described. They consisted of:

1. ***Reliawind Data:***
   Low-frequency SCADA data (1 sample per 600 seconds) covering 190 channels of data collected from a collection of 6 wind turbines.

2. ***High frequency bearing accelerometer data (run to failure):***
   20 kHz data of both failure and suspension histories of bearings.

3. ***High frequency bearing accelerometer data (seeded fault):***
   48 kHz data containing electro-discharge machined failures of varying sizes.

As such, given the data available for analysis and both the research questions and objectives specified throughout this chapter, it is possible to explore and expand the current state-of-the-art with regards to the condition monitoring of assets and components for the early detection, diagnosis and prognosis of degradation. In order to perform this, models are required which encapsulate the current state of the asset or component. Chapter 4 explores the development of two models for the purpose of condition assessment, based upon both high frequency bearing accelerometer data and low frequency wind turbine gearbox SCADA data.

Chapter 4

Condition Assessment Models

The contents of this chapter appear in part in the following peer-reviewed publication:

- Godwin, J. L. & Matthews, P. C. (2014). *Robust Statistical Methods for Rapid Data Labelling.* In Bhatnatgar, V. Editor, *Data Mining and Analysis in the Engineering Field* (pp. 107–141). Hershey, Pennsylvania (USA): IGI Global.

This chapter presents two models for condition assessment; the "Moments model" for high frequency bearing data, and the "extended physics of failure model" for low frequency gearbox SCADA data. Statistical time-domain features are evaluated for their capability to identify the early signs of degradation within high frequency bearing data; the deviation of each feature over time is employed as a means to quantify feature quality, enabling features to be ranked relatively to identify leading features of degradation. Following this, state of the art gearbox physics of failure models are explored. These are refined, with extensions derived from this in order to normalise for turbine loading, transient states, external conditions and wind farm topology. The model is validated by identifying a failure on an independent turbine.
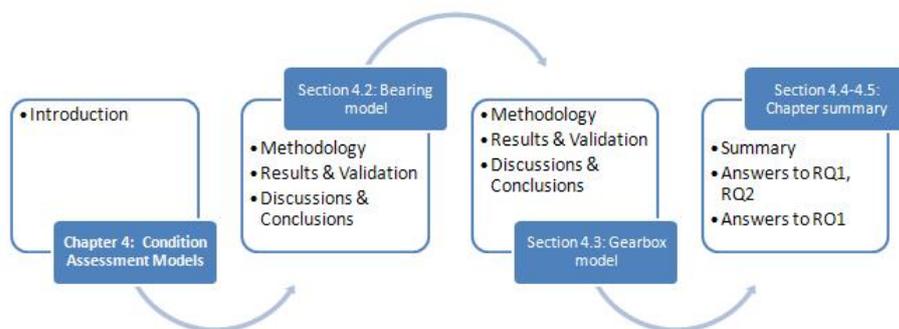
Figure 4.1: Overview of the structure of chapter 4.

4.1   Introduction

Having explored the limitations of the current state of the art in chapter 2 and defined the scope of this thesis in chapter 3, it is necessary to undertake the preliminary work in quantifying current operational behaviour of assets.

The basis for any condition monitoring or prognostic solution is a sound set of features which accurately and precisely encapsulate the current operational behaviour of the asset or component, and is free from noise or other undesirable contamination. As such, this chapter explores in detail the features identified throughout section 2.4.2 which can be exploited for the health monitoring in both high frequency and low frequency contexts. Specifically, this chapter looks at time domain statistical features for degradation assessment within the context of high frequency roller element bearing data, and physics of failure models of degradation within the context of low frequency gearbox SCADA data.

The identification of features (or residuals of features) which correlate with degradation is essential. As this thesis is concerned with the eventual prognosis of asset condition or estimation of remaining useful life, this would not be possible without these features. Poor selection of features would introduce noise into the derived health metric, which could cause both false negative and false positive responses. Within the context of this research, this would equate to the untimely failure of the bearing or gearbox as the metric failed to identify degradation (false negative), or increased maintenance expense through the identification of degradation which had not occurred (false positive).

This chapter explores the first research objective of this thesis as discussed in 3.2.1, namely:

RO1: *To explore if the use of time-domain features are appropriate as leading indicators of degradation with regards to both bearings and gearboxes for either low or high frequency data, utilising either statistical, data-driven or physical models.*

This is further broken down into two research questions as discussed in sections 3.3.1 (RQ1) and 3.3.2 (RQ2). These are:

RQ1: *Are time-domain features beneficial to assisting in the empirical quantification of bearing degradation based upon high frequency accelerometer data? If so, which features differ significantly and substantially as a result of degradation? (**RO1**)*

RQ2: Is it possible to normalise a non-stationary signal based upon low frequency gearbox data from the influence of loading conditions, transient states and external factors? If so, can this be empirically derived? (RO1)

These questions will be explored and answered in sections 4.2 and 4.3 (respectively).

4.2   Bearing health model

4.2.1   Introduction

In order to develop a condition assessment metric, it is necessary to determine a set of features which are strongly correlated to the current operational behaviour of the asset or component in question. This section explores the application of statistical time-domain features for the purposes of high frequency vibration data condition assessment for a bearing. In total 28 time domain features are explored across 5 categories. Robust statistical process control (SPC) is used to determine the capability of each feature for early detection of degradation and as a leading indicator of this degradation.

4.2.2   Dataset employed

In order to determine which of the selected features correlate to operational behaviour of the bearing, historical bearing data containing failures was employed for empirical validation and verification. In this case, the bearing dataset prepared by Rexnord and made available by NASA was employed (Lee et al., 2007) as described in section 3.4.2.

In order to comparatively evaluate the features, multiple bearings are required. In this case, historical data pertaining to the life cycles of two bearings were employed. These were:

- **Bearing 1 (Test 2):** A failure history consisting of 20,480 data points per file, across 984 files. Referred to as the "failed" bearing.

- **Bearing 2 (Test 2):** A suspension history consisting of 20,480 data points per file, across 984 files. Referred to as the "healthy" bearing.

There is a necessity to employ both a failure history (a bearing which ultimately failed) and also a suspension history (a bearing which ultimately did not fail). This is required in order to discriminate between features. The data in this dataset was sampled at 20 kHz (20,480 samples per second), with a one second sample taken every 10 minutes. In total 984 samples were taken over the period of six and a half days. The failed bearing was deemed to have an outer race fault at the end of this sample period.

4.2.3 Bearing model development

In order to exploit the acquired data for the purposes of feature selection, pre-processing of the data is required. This is required for a number of reasons. Firstly, missing values within the data must be managed. In this case, missing values were removed. This was performed as missing data does not contribute to the quality of a given feature, whilst increasing the computational complexity of the feature calculation. Implausible, duplicate and erroneous values remained within the data. This was for two reasons. Firstly, it is possible that these values are a result of degradation. As such, the removal of these values would potentially remove vital information from a signal, and reduce the quality of derived features. Secondly, given the difficulty in the identification of these values and the potential benefit which could be gained from the inclusion of these values, it was determined that these values should remain within the data (should they exist).

After pre-processing of the data, statistical time-domain features were extracted from the dataset. These features are detailed in appendix sections A.1 through A.5. In total, 28 features were identified through the literature search undertaken in chapter 2 and were explored for their correlation to the operational behaviour of the bearings. These were:

- **Measures of central tendency:** Mean, median, mode, trimean, root mean square (RMS) and the Winsorized mean.

- **Measures of variability:** Standard deviation, variance, interquartile range (IQR), range, minimum, maximum and the median absolute deviation (MAD).

- **Measures of shape:** Skewness, kurtosis, hyperskewness, hyperflatness, crest factor (CF) and the peak to average power ratio (PAPR).

- **Measures of position:** The 10th percentile, 25th percentile, 75th percentile and the 90th percentile.

- **Measures of impurity:** Shannon entropy, Chao-Shen entropy, Miller-Madow entropy, James-Stein style shrinkage estimator of entropy and a Bayesian estimate of entropy (Krichevsky-Trofimov).

Due to various phenomena which occur in reliability engineering, further processing of the data is required in order to prepare the features for analysis. Some of these phenomena include the running in of components, infant mortality failures and wear out failures. Due to manufacturing defects or slight variation due to inherent tolerances, each bearing will have a unique path to failure dependent upon these conditions,

the operational usage of the component and the external conditions the component is exposed to. Within a "type I" (population based) prognostic approach, the hazard function describing failures represents a bathtub and is referred to as the "bathtub curve" (Leemis, 1995) this is illustrated in figure 4.2. As can be seen in this figure, three distinct notions exist:

- ***Infant mortality (burn-in or early failure period):*** Decreasing failure rate.

- ***Useful life (or intrinsic failure period):*** Constant failure rate.

- ***Wear out (or end of life):*** Increasing failure rate.

The bathtub hazard rate function is defined in probabilistic terms as per Leemis (1995). Consider the probability of failure between times t and $t + \Delta t$:

$$P[t \leq T \leq t + \Delta t] = \int_t^{t+\Delta t} f(\tau)d\tau = S(t) - S(t + \Delta t) \tag{4.1}$$

Where $f(\tau)$ is the PDF and $S(t)$ is the survivor function (Leemis, 1995). Assuming that the item is working at time t yields:

$$P[t \leq T \leq t + \Delta t | T \geq t] = \frac{P[t \leq T \leq t + \Delta t]}{P[T \geq t]} = \frac{S(t) - S(t + \Delta t)}{S(t)} \tag{4.2}$$
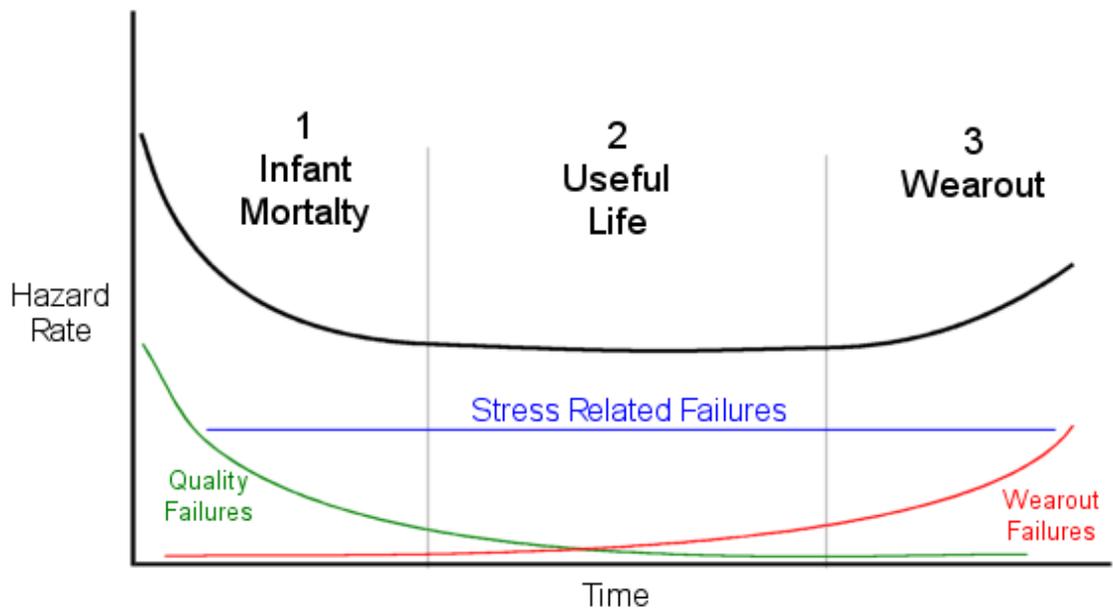
Figure 4.2: Traditional bathtub curve.

As Δt approaches 0, this becomes the instantaneous failure rate (the hazard function):

$$h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{S(t)\Delta t} = \frac{-S'(t)}{S(t)} = \frac{f(t)}{S(t)} \tag{4.3}$$

Thus, the hazard function is the ratio of of the probability density function to the survivor function. As such, a probabilistic interpretation of the hazard function is:

$$h(t)\Delta t = P[t \leq T \leq t + \Delta t | T \geq t] \tag{4.4}$$

Due to the potential variance introduced by the running in of bearings by manufacturing defects or inherent tolerances, the first 10% of the test data was disregarded. This was done to reduce the influence of this phenomena on the selected features. The noise introduced into the data by the running in of the bearings was mitigated by the removal of this data. As such, the first 98 data samples were removed from the dataset due to this.

As the bathtub curve refers to populations of assets, rather than individuals this may not seem intuitive. The hazard function represents the instantaneous rate of failure for the surviving population of assets at a given time, during the next time period. However, the decrease failure rate during the 'infant mortality' (early failure period) stage, provides some evidence that phenomena such as burn-in and manufacturing defects do exist and can cause catastrophic consequences. Although these phenomena are not as common as they once were, cleansing the data in this way assists in mitigating the impact of this (should these phenomena be present).

This assists in enabling normal operational behaviour of the bearing to be encapsulated. As such, we can utilise the remaining data in order to determine statistically significant differences between the normal operational behaviour of the bearing and the degraded operation of the bearing, however, various challenges exist. These can be broken two into two distinct categories; methodological and application-specific.

In terms of methodological challenges, the first of these issues is that various statistical assumptions required for traditional statistical hypothesis testing often do not hold in practice on data of this nature. For instance, although it is possible to perform a two-sample, two-tailed paired t-test to determine if degradation differs between a baseline condition (normal behaviour) and a given condition (degraded behaviour), this is non-nonsensical due to temporal dependency of the data. Furthermore, due to the inherent variance, large sample rate and noise within the data, statistically significant results will often be obtained although no substantial change has occurred. Similarly, traditional t-tests require normally distributed data, which may not occur

in practice for various reasons (which may include degradation). Due to this temporal dependency, typically auto-regressive techniques are employed for the analysis of time series data. However, due to the challenges specific to the analysis of degradation data, other problems also arise.

In terms of challenges specific to data pertaining to degradation, given that degradation can be thought of as a function of both time and operational usage this causes issues for the application of auto-regressive techniques. All bearing data covering a substantial period for analysis will contain degradation to some extent. As such it would be expected that the regression model gradient would differ from zero. Although it is anticipated that features highly correlated to degradation would have a steeper gradient than normal operational behaviour, given that degradation is a stochastic behaviour and often exponential in nature, large data bias would exist. A significant proportion of a given failure history would represent normal operational behaviour. Without left-censoring the data, normal operational behaviour would create a large bias which may reduce the sensitivity of the auto-regressive models. Although this would provide a means to quantify the features in terms of their capability, large quantities of data would potentially be disregarded for this analysis.

Reducing the quantity of data would reduce the power of statistical hypothesis testing. As degradation can be thought of as a deviation from normal operational behaviour, by encapsulating normal operational behaviour by exploiting this data source, deviations from this (such as potential degradation) can be identified easier. As such, other methodologies for this analysis were sought which could exploit the full range of data available within the dataset.

4.2.4   Statistical process control

In order to exploit the full range of data available, a robust statistical process control methodology was employed. Statistical process control (SPC) is a quality control process where samples are periodically taken and tracked about an estimated (or known) mean. Typically, a control chart showing $\pm 3\sigma$ is employed to track the current state of operation. Statistical process control provides a means to objectively quantify the variation of a process over time, with an emphasis on the early detection of "out-of-control" samples (Chen and Shaw, 2011).

For a normally distributed process, $\pm 3\sigma$ will cover 99.7% of the variance of all observations under normal operational behaviour. However, different sources of variance may develop and become apparent within the model. Within statistical process control, these are referred to as "common" and "special" source of variation. Common sources of variation typically include noise, natural patterns and active phenomena

within the system. Special sources of variation are often characterised by new phenomena within the system, variation from the historical reference and an inherent change to the underlying process.

However – as discussed previously – data from real-world applications typically does not conform to various assumptions required in statistical models. For instance, degradation data is notorious for its non-normality. As such, the basis for traditional (Shewhart) control charts is often invalid in practice. This can be mitigated in a variety of ways. These are briefly described below:

- *Box-Cox transformation:* This is a statistical transform (Box and Cox, 1964) which utilises power functions to manipulate variables in order to more closely match the normal distribution.

- *Change process monitoring (CPM):* This is a statistical technique which is capable of identifying changes in the mean, variance, regression or dependence of a time series, enabling the identification of degradation in bearing data (Brodsky and Darkhovsky, 1993).

- *Exponential weighted moving average (EWMA) charts:* These are parametric control charts which are regarded as robust against the assumption of normality (Chen and Shaw, 2011).

Although each of these options are capable of quantifying the quality of the time-domain features explored in this chapter, some are better suited than others. For instance, although the Box-Cox transform is widely utilised to normalise data, some variables cannot be transformed effectively. As such, unless additional prior information is known about the process, this technique may not be suitable in practice.

Similarly, CPM methodologies are effective non-parametric techniques at identifying change points. However, due to the various changes which can occur within a time series, unless the effect of degradation is known for a given feature a-priori, some non-parametric change point methodologies can lack statistical power.

Finally, exponential weighted moving average (EWMA) control charts are parametric control charts which re-weight the raw data enabling more recent data points to contribute more to the current process value than prior data points. EWMA charts are regarded as robust to the assumption of normality (Testik et al., 2003) and as such, are a practical tool for this analysis in particular, and can be recommended when data is not normally distributed (Stoumbos and Reynolds, 2004). Given the multivariate nature of this thesis in particular, it should be noted that multivariate extensions

are readily available to the univariate EWMA chart. The (univariate) exponentially weighted moving average is defined as:

$$z_i = \dot{\lambda} x_i + (1 - \dot{\lambda}) z_{i-1} \tag{4.5}$$

Where $0 \leq \dot{\lambda} \leq 1$ is defined as a constant. For the purposes of this thesis, and in line with Montgomery (2007), $\dot{\lambda}$ is fixed to 0.1. The upper and lower control limits of this chart are defined as:

$$CL = \mu_0 \pm L\sigma \sqrt{\frac{\dot{\lambda}}{(2 - \dot{\lambda})}[1 - (1 - \dot{\lambda})^{2i}]} \tag{4.6}$$

Where μ_0 represents the first observation sample average, L represents the limits (typically denoted in deviations) and $\dot{\lambda}$ is defined as above. For the purposes of this thesis, and in line with Montgomery (2007), L is set to 3 for all attributes.

As the conditions present during the acquisition of the data explored for condition assessment of bearings remained constant throughout, sources of variance (both common and special) would be captured. As such, should the process deviate and become "out-of-control", it may be due to special variance which has occurred as a result of degradation. As degradation can be thought of as a phenomena outside of control, the use of statistical process control provides a suitable means to identify features which correlate to the condition of the bearing.

The implementation of a statistical process control methodology requires two stages (also known as phases). Initially, the process must be established and the process variance of the attribute determined (phase 1). Following this, the derived limits can be employed to determine the deviation of current operational behaviour to the known standard (phase 2).

For this purpose, the establishment of process variance was performed on the attributes. In total, 40% of bearing life was used as a baseline from which to measure degradation against. This compromised 10% to 50% of asset life (as the first 10% of life has been removed prior to avoid the noise introduced through running-in phenomena).

Once this baseline has been established, it is possible to quantify the capability of a given feature. It should be noted that due to the degradation process, any baseline (i.e. "normal behaviour") will inherently contain degradation. As such, it is expected that data collected from a healthy bearing will eventually move beyond initial control limits due to this, however, at a later time period than that of the bearing which failed.

As such, it is necessary to define the anticipated behaviour of both strong and weak features so that a metric can be created in order to assist in attribute selection. The suitability of each feature was determined by answering the following set of questions:

1. *Is this feature consistently in control?* If the feature remains within control limits for both the healthy and failed bearings, it is unlikely to deviate substantially due to degradation.

   - *With regards to healthy bearings:* It is anticipated that normal operational behaviour will remaining within control limits during periods of normal operational behaviour.

   - *With regards to failed bearings:* It is anticipated that degraded behaviour will move beyond control limits and remain out side of control.

2. *Is this feature consistently outside of control?* If the feature remains beyond control limits for the healthy and failed bearings, it is likely to contain too much noise (for instance, false positives or false negatives) to be of value.

3. *When does the feature move beyond control?* Features which are more sensitive to degradation should deviate from control the earliest, for instance:

   - *With regards to healthy bearings:* It is anticipated that normal operational behaviour will remaining within control limits beyond that of the failed bearing.

   - *With regards to failed bearings:* It is anticipated that degraded behaviour will move beyond control limits and remain out side of control substantially earlier than that of the healthy bearing.

Having established the baseline and characterised the expected behaviour of the features, analysis of the features could then be performed. Statistical process control charts exploiting exponentially weighted moving averages (EWMA) were created compromising data from 10% to 100% of bearing life. Although it is hoped that various features can identify degradation prior to failure, the failure and suspension histories are employed until the data acquisition was concluded in order to assist in the characterisation of wear-out and failure. The 28 univariate features employed within the statistical process control methodology detailed above are described in detail in appendix A.

4.2.5 Results

For convenience, due to the large quantity of graphs presented in this section (84 in total), these have been placed in appendix B. Each of the five categories of features are presented separately for analysis.

Feature ranking metric

In order to rank the attributes on their potential for condition assessment, two factors were taken into consideration. First and foremost, the percentage of life at which the attribute left control. It is intuitive that features which accurately encapsulate degradation on the bearing should be able to perform early detection of degradation. As such, the earlier the feature deviates from control for the failed bearing, the stronger this potential feature is. Secondly, the percentage of data that was in control throughout the test is taken into consideration. As the suspended (healthy) bearing explored in this chapter will also inherently contain degradation information, stronger features are anticipated to detect this, and as such, will be reflected in the ranking of the attribute. The objective function to minimise in this case is defined as:

$$S_f = \frac{\frac{(OC_h + OC_f)}{2} + \frac{(IN_h + IN_f)}{2}}{2} \tag{4.7}$$

Where the score of feature f is denoted S_f, OC_h refers to the earliest point (as a percentage of bearing life) at which this attribute consistently left the predefined control limits (in this case, out of control for all samples covering at least 1% of life) for the healthy bearing (with OC_f referring to that of the failed bearing). Similarly, IN_h refers to the percentage of data points in control for the healthy bearing (with IN_f referring to that of the failed bearing).

Measures of central tendency

With regards to the measures of central tendency, the six attributes were analysed for their performance and encapsulation of behaviour of the bearings. The results of this analysis are presented below and summarised in table 4.1, along with the respective score and ranking of each feature.

The ranking attained from the scoring function for the measures of central tendency is:

1. Root mean square (ranked joint 1st of 28, $S_f = 41.43$)

2. Median (ranked 15 of 28, $S_f = 71.14$)

3. Winsorized mean (ranked 17 of 28, $S_f = 72.49$)

4. Tri-mean (ranked 20 of 28, $S_f = 73.23$)

5. Mean (ranked 25 of 28, $S_f = 90.18$)

6. Mode (ranked 28 of 28, $S_f = 98.12$)

Table 4.1: Performance of measures of central tendency.

	Suspension history		Failure history	
	Leaves control	Data in control	Leaves control	Data in control
Feature	(%)	(%)	(%)	(%)
RMS	48.10	20.12	53.10	44.41
Winsorized mean	70.10	85.87	66.10	67.89
Median	70.10	86.99	66.10	61.38
Tri-mean	70.10	87.80	66.10	68.90
Mean	70.10	92.78	N/A	97.87
Mode	N/A	98.88	N/A	93.59

Measures of variability

With regards to the measures of variability, the seven attributes were analysed for their performance and encapsulation of behaviour of the bearings. The results of this analysis are presented below and summarised in table 4.2, along with the respective score and ranking of each feature.

The ranking attained from the scoring function for the measures of variability is:

1. Standard deviation (ranked joint 1st of 28, $S_f = 41.43$)

2. Variance (ranked joint 1st of 28, $S_f = 41.43$)

3. Median absolute deviation (MAD) (ranked 6 of 28, $S_f = 43.51$)

4. IQR (ranked 8 of 28, $S_f = 44.91$)

5. Range (ranked 16 of 28, $S_f = 71.76$)

6. Minimum (ranked 18 of 28, $S_f = 72.78$)

7. Maximum (ranked 19 of 28, $S_f = 72.79$)

Measures of shape

With regards to the measures of shape, the six attributes were analysed for their performance and encapsulation of behaviour of the the bearings. The results of this analysis are presented below and summarised in table 4.3, along with the respective score and ranking of each feature.

The ranking attained from the scoring function for the measures of shape is:

1. Crest factor (ranked 7 of 28, $S_f = 43.95$)

2. Hyperflatness (ranked 9 of 28, $S_f = 45.03$)

3. Kurtosis (ranked 13 of 28, $S_f = 67.25$)

4. Hyperskewness (ranked 14 of 28, $S_f = 70.07$)

5. Skewness (ranked 23 of 28, $S_f = 77.87$)

6. Peak to average power ratio (ranked 27 of 28, $S_f = 91.89$)

Measures of position

With regards to the measures of position, the six attributes were analysed for their performance and encapsulation of behaviour of the bearings. The results of this analysis are presented below and summarised in table 4.4, along with the respective score and ranking of each feature.

The ranking attained from the scoring function for the measures of position is:

1. 90th percentile (ranked 4 of 28, $S_f = 43.34$)

2. 10th percentile (ranked 5 of 28, $S_f = 43.39$)

3. 25th percentile (ranked 10 of 28, $S_f = 45.19$)

4. 75th percentile (ranked 11 of 28, $S_f = 49.35$)

Table 4.2: Performance of measures of variability.

Feature	Suspension history		Failure history	
	Leaves control (%)	Data in control (%)	Leaves control (%)	Data in control (%)
Standard deviation	48.10	20.12	53.10	44.41
Variance	48.10	20.12	53.10	44.41
MAD	48.10	24.79	54.10	47.05
IQR	48.10	30.38	54.10	47.05
Range	91.10	87.50	56.10	52.34
Minimum	91.10	90.65	56.10	53.25
Maximum	92.10	87.50	57.10	54.47

Table 4.3: Performance of measures of shape.

Feature	Suspension history		Failure history	
	Leaves control (%)	Data in control (%)	Leaves control (%)	Data in control (%)
Crest factor	48.10	29.47	53.10	45.12
Hyperflatness	50.10	35.46	54.10	40.44
Kurtosis	79.10	76.93	59.10	53.86
Hyperskewness	86.10	81.91	65.10	47.15
Skewness	91.10	88.41	66.10	65.85
PAPR	93.10	89.74	95.10	89.63

Table 4.4: Performance of measures of position.

	Suspension history		Failure history	
	Leaves control	Data in control	Leaves control	Data in control
Feature	(%)	(%)	(%)	(%)
90th percentile	48.10	27.54	53.10	44.61
10th percentile	48.10	28.76	53.10	43.60
25th percentile	48.10	32.62	53.10	46.95
75th percentile	57.10	39.43	54.10	46.75

Measures of impurity

With regards to the measures of impurity, the five attributes were analysed for their performance and encapsulation of behaviour of the bearings. The results of this analysis are presented below and summarised in table 4.5, along with the respective score and ranking of each feature.

The ranking attained from the scoring function for the measures of impurity is:

1. Miller-Madow entropy (ranked 12 of 28, $S_f = 65.45$)

2. James-Stein shrinkage entropy (ranked joint 21 of 28, $S_f = 76.15$)

3. Shannon entropy (ranked joint 21 of 28, $S_f = 76.15$)

4. KT entropy (ranked 24 of 28, $S_f = 80.45$)

5. Chao-Shen entropy (ranked 26 of 28, $S_f = 91.49$)

4.2.6 Discussion

Each individual feature (and category of features) provide some insight into the condition of both bearings. Naturally, some features will provide more insight than others, and these are explored here. For convenience, the time series of each feature for both the failure history (the failed bearing) and the suspension history (the healthier of the

Table 4.5: Performance of measures of impurity.

	Suspension history		Failure history	
	Leaves control	Data in control	Leaves control	Data in control
Feature	(%)	(%)	(%)	(%)
Miller-Madow entropy	69.10	83.84	55.10	53.76
James-Stein entropy	94.10	88.82	61.10	60.57
Shannon entropy	94.10	88.82	61.10	60.57
KT entropy	94.10	88.41	71.10	68.19
Chao-Shen entropy	92.10	89.73	95.10	89.02

two bearings which had not failed at the end of the test) are presented in appendix B. Due to the large quantity of figures presented in B, two of these have been reproduced in this section to assist the reader in understanding the methodology undertaken and the scoring metrics employed.

Measures of central tendency

With regards to the mean, this was found to be the 4th least suitable feature explored in this chapter. This feature was unable to correctly identify degradation before failure. No trend was identified on either the healthy or failed bearing. Although a small proportion of data points on the failed bearing were identified as out of control at the end of bearing life, the feature consistently found points within control during the expected period of degradation. Similarly, more points were in control for the failed bearing (97.87%) than the healthy bearing (92.78%). This was unexpected and would cause potential false positive or false negative results, and as such, it is felt that the mean is not a viable feature for bearing degradation assessment. No clear trend can be observed in figure B.3. Potential degradation was identified in the healthy bearing at 70% of life, but not in the failed bearing. This is inconsistent with what is anticipated and as such, reduces the value of this attribute.

Similarly to the mean, the mode is not capable of consistently identifying degradation on the failed bearing or potential degradation on the healthy bearing. Although more points were in control for the healthy bearing than the failed bearing (98.88% and 93.60% in control, respectively), no observable trend or consistency is present, as is shown in in the appendix in figure B.9 and reproduced here for clarity in figure 4.3. No consistent detection of degradation (based upon control limits) was identified for either bearing; leading to the lowest rank of any feature explored (28th).

As can be seen in figure 4.3, the mode as a feature does not perform well for early fault detection. The normal operation of this (the failed bearing) has substantial variance, and as such, little to no trending behaviour can be identified as a result of this. Immediately prior to failure, some of the data points are regarded as being out of control. It is likely that this is due to the increased variance within the signal due to the degradation. It should be noted that points are regarded as out of control prior to the end of the test, however, the bearing almost immediately returns to within control limits. This is (typically) not desirable and may not be representative of the true degradation which is effecting the bearing. The failed bearing is difficult to distinguish from the healthy bearing, even at the end of the test (as can be seen in figures B.7). As such, this feature is incapable of the early detection of potential degradation in this case.
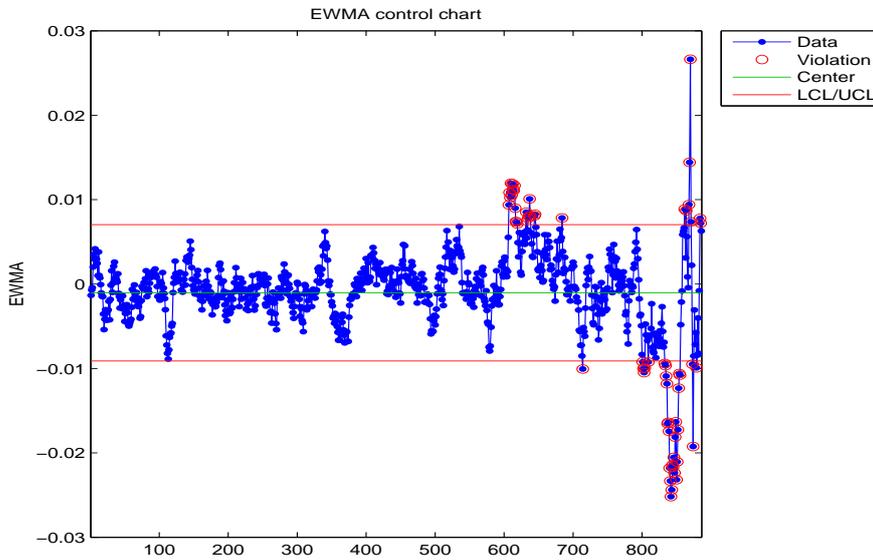
Figure 4.3: The mode of the failed bearing.

The Winsorized mean was able to detect degradation on the failed bearing (at 53% of life), and as such, out performed the mean. However, this feature identified the potential degradation which occurred on the healthy bearing at the same moment as the mean (at 70% of life). Substantially more data is deemed to be out of control in this case for the failed bearing (67.89% in control) than the healthy bearing (85.87% in control). This is expected, and provides some potential for use as a feature for bearing degradation assessment. A minor trend of degradation can be observed in figure B.18, however, some noise is present for both bearings.

Both the median and tri-mean performed similarly well; both performed better than the traditional mean and the mode, but not as strongly as both the RMS or Winsorized mean. The median was unable to identify degradation any earlier than the tri-mean (with both identifying degradation at 66% of life for the failed bearing and potential degradation at 70% of life for the healthy bearing). The median is seemingly more sensitive to degradation in both cases, hence the ranking of this feature. For the failed bearing, 61.38% of data in is deemed in control, with the tri-mean deeming 68.90% of data in control. This is echoed in the results for the healthy bearing, with 86.99% and 87.80% of data being regarded as in control for the median and tri-mean (respectively). A minor trend similar to that of the Winsorized mean (but more noisy) can be observed in both the median and tri-mean results, as observed in figures B.6 and B.12.

The best performing measure of central tendency in this context is the root mean

square (RMS). As can be observed in figure B.15 and reproduced below in figure 4.4, trends can be identified in the failed bearing leading to degradation and also the healthy bearing leading to potential degradation. This feature was able to identify the degradation on the failed bearing the earliest of all measures of central tendency (at 53% of life), and was highly sensitive (with only 44.41% of data being deemed within control). The feature was also able to strongly identify potential degradation on the healthy bearing, at 48% of life (with 20.12% of data being in control). Although this result is unexpected for the healthy bearing, this may be due to the sensitivity of the EWMA charts and the defined control limits at detecting deviation from a stationary position.

As can be seen in figure 4.4, in this case, clear distinctions can be made between the healthy and failed bearings. Strong trending characteristics (in comparison to other features, such as the mode) can be identified and exploited for the purposes of early detection and condition assessment. Substantial changes from the initial bearing conditions are apparent towards the end of the test. The normal operation for both bearings is characterised as one would anticipate; relatively stable condition throughout the components life until the end of life period where significant changes occur, potentially due to degradation. It should be noted that a slight increase towards the end of life for the healthy bearing is also apparent. This may signify that some slight potential degradation may have occurred on this bearing, however, not enough to cause failure.
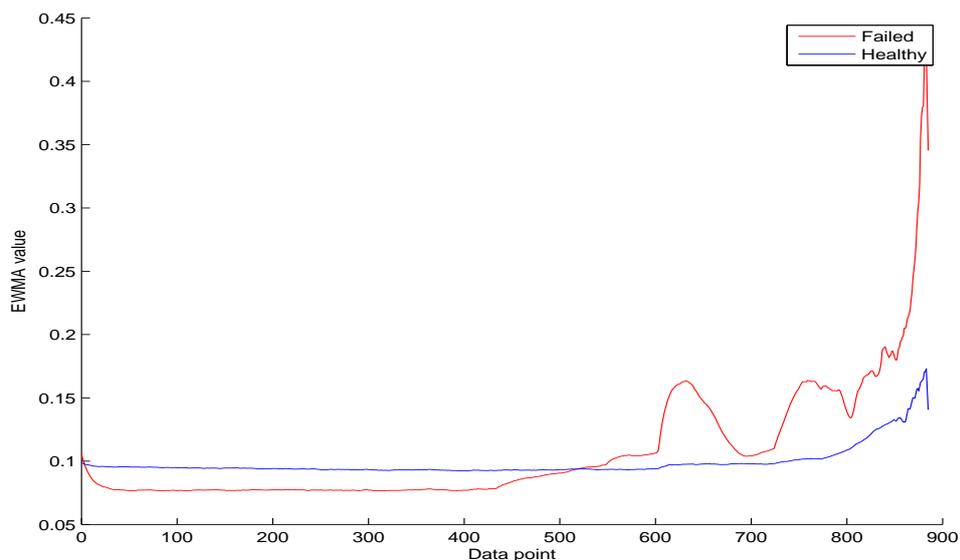
Figure 4.4: The RMS of the failed and healthy bearing.

This hints at the sensitivity of this feature and that this feature has the potential for the early detection of faults in this case.

Overall, the measures of central tendency appear to perform at a standard lower than one would anticipate for the purposes of condition assessment (aside from the root mean square). This could be due to a number of factors, such as the features being insensitive to degradation phenomena, or noise within the dataset. Although the robust measures of central tendency out-perform the traditional measures of central tendency, it may be that measures of central tendency simply cannot encapsulate degradation phenomena effectively. As such, the benefit gained from robust estimation of these parameters is limited. As the root mean square was able to identify degradation in the failed bearing (and potential degradation in the healthy bearing) earlier than other features, and trends of degradation were visible (figure B.15), we can be satisfied that it is the limitations of the features explored here rather than external influence which ultimately led to the consistently poor results attained.

Measures of variability

Both the maximum and minimum features are found to be adequate features for the purposes of bearing degradation and condition assessment. Trends which may be indicative of degradation can be identified in both the healthy and failed bearings as illustrated in figures B.33 and B.36. The minimum slightly out performs the maximum in terms of the early identification of degradation on the failed bearing (56% of life, as opposed to 57% of life). Both the maximum and the minimum identify potential degradation on the healthy bearing later than other features (91% of life and 92% of life, respectively), and this is reflected in the relative ranking of the features. It it interesting to note that whilst the minimum slightly outperforms the maximum, more data is regarded as in control by the minimum for the healthy bearing (90.65%) than by the maximum (87.50%). Both the minimum and maximum features returned comparable control proportions for the failed bearing (53.25% and 54.47%, respectively).

The range was found to be similar to the minimum and maximum features. Similar trends and observations exist within the time series of both the healthy and failed bearing (as illustrated in figure B.30). Degradation and potential degradation was identified at the same period as the minimum on both the failed and healthy bearings (56% and 91% of life, respectively). Fewer samples were in control for the failed bearing (52.34%) than the minimum and maximum, suggesting the heightened sensitivity of this feature given the early detection of degradation. It should be noted that this feature had more out of control points (87.50% of points were in control) than the minimum (90.65% of points in control); meaning it is potentially less sensitive than the minimum

on the healthy bearing (but equal to that of the maximum). This is unexpected, however, the sensitivity to degradation attained on the failed bearing is higher than that of either the minimum or maximum, which affords the feature the higher ranking.

The interquartile range was found to be an improvement over the traditional range. This is potentially due to limiting the influence of outliers, noise and potential influence from external conditions by removing the first and also the final quartile of the data. As can be seen in figure B.27, similar trends are identified as for other strong features. The interquartile range was able to identify the degradation on the failed bearing almost as early as the strongest features (at 54% of life with 47.05% of data in control). Similarly, this feature was able to identify potential degradation on the healthy bearing at 48% of life; comparable to the strongest features explored in this work (with 38.39% of data being in control).

The median absolute deviation (MAD) was found to be a strong indicator of potential degradation. Degradation is detected early on both the failed bearing (at 54% of life) and potentially on the healthy bearing (at 48% of life). Strong trends are observed within the time series, as shown in figure B.39. Data is found to be out of control earlier and more consistently than the previously discussed measures of variability, with 47.05% of data being in control for the failed bearing and 24.79% of data being in control for the healthy bearing. Given the simplicity and robustness of this feature, favourable performance was found.

The final two measures of variability explored are the variance and standard deviation. These are found to be (jointly, in conjunction with the RMS) the best performing features explored in this work. Both the variance and standard deviation display strong trending behaviour (figures B.24 and B.21), and identify degradation on the failed bearing at 53% of life (the earliest of any feature explored in this work). Similarly, potential degradation is identified on the healthy bearing at 48% of life (again, the earliest of any feature explored in this work). Both of these features classify the data as in control in a comparable manner for the failed bearing (44.41% of data for both features). With regards to the healthy bearing, identical results are achieved. Both features report 20.12% of data in control. This is most likely due to the variance being the standard deviation squared, and thus their high correlation.

Overall, the measures of variability perform consistently better than the measures of central tendency, and typically produce strong metrics which can be explored for the early identification of potential degradation. This could be due to a number of factors, and is likely driven by degradation changing the underlying distribution of the data. If the mechanism of degradation is thought of as an additional source of variability (i.e. a "special" source of variance under SPC terminology), this provides some insight into

why these features are more capable than the measures of central tendency explored
previously. It should be noted that similar to the measures of central tendency, the more
robust measures (such as the interquartile range) perform better than their counterpart
(for instance, in comparison to the traditional range, or the maximum and minimum).

Measures of shape

The peak to average power ratio (PAPR) performed poorly. Degradation on the failed
bearing was detected late (at 95% of life), with 89.63% of data being in control through-
out the life of the bearing. Similarly for the healthy bearing, although potential degra-
dation is detected, it is also late (at 93% of life). As such, and given the noisy nature
of this feature as seen in figure B.57, this leads to the low ranking of this feature.

Surprisingly, the crest factor was found to perform well. Given the consistency of the
standard deviation and variance, this was unexpected. Degradation was detected by the
crest factor on the failed bearing (at 53% of life), with potential degradation detected
on the healthy bearing at 48% of life. This feature was also sensitive, with 45.12%
and 29.47% of data being in control for the failed and healthy bearings (respectively).
Similar to the peak to average power ratio, similar noisy trends are observed in figure
B.54, limiting the ranking of this feature.

Both skewness and kurtosis were found to be typical features for bearing condi-
tion assessment. Although both features were able to identify the degradation on the
failed bearing and potential degradation on the healthy bearing, kurtosis performed
somewhat more consistently than skewness. As can be seen in figure B.42, although
degradation is detected early on the failed bearing (66% of bearing life), the deviation
from control limits returns to within control for a period of time prior and remains
noisy until failure. Whilst this is a signature which could be exploited in itself, in
comparison to figure B.45 (representing kurtosis), a more consistent trend can be seen.
Kurtosis was able to identify degradation earlier than skewness (at 59% of life on the
failed bearing). This trend continues with the healthy bearing, with kurtosis identify-
ing potential degradation earlier than skewness (at 79%, in contrast to 91%, of life).
Kurtosis was similarly more sensitive than skewness, as can be seen in table 4.3.

The final two measures of shape explored for bearing condition assessment are the
5th and 6th moments (higher order moments) which are referred to as hyperskewness and
hyperflatness. The time series of these features are presented in figures B.51 and B.48.
Both of these higher order moments outperformed the many traditionally explored
features (including skewness). Hyperflatness outperformed hyperskewness in terms of
both early detection – degradation was detected at 54% of life for the failed bearing
and potential degradation at 50% of life for the healthy bearing, as opposed to 65%

and 86% (respectively) – and also in terms of sensitivity (as can be seen in table 4.3).

Overall, the measures of shape lack consistency in comparison to the other measures explored in this work, however, they have the potential to produce strong metrics which can be explored for the early identification of potential degradation. Similar to previous work, skewness and kurtosis were found to be adequate features for the purposes of identification of potential degradation and found to have average sensitivity. It is interesting to note that the higher order moments may potentially perform better than the lower order moments (in this case).

Measures of position

The measures of position explored are perhaps the most consistent of all features explored in this work. Both the 10th and 90th percentiles identify degradation on the failed bearing at 53% of life, and the potential degradation on the healthy bearing at 48% of life (the same detection point as the standard deviation, variance and RMS). The 90th percentile is marginally more sensitive than the 10th percentile with less data in control for the healthy bearing (27.54% and 28.76% of data in control, respectively). However, for the failed bearing the 10th percentile is slightly more sensitive than the 90th percentile (43.60% and 44.61% of data in control, respectively). Both features exhibit similar trends to other favourable features. These are presented in figures B.60 and B.69.

The 75th and 25th percentile perform similarly to the other measures of position explored, however, perhaps due to the reduced sensitivity to outlying data (due to their position), these features are not as sensitive to degradation as the 10th and 90th percentiles.

The 25th percentile performs somewhat better than the 75th percentile. With regards to the failed bearing, detection of degradation is earlier (53% of life as opposed to 54% of life), and of comparable sensitivity (46.95% of data in control, as opposed to 46.75%). For the healthy bearing, the 25th percentile detects potential degradation much earlier than the 75th percentile. In this case, at 48% of life, rather than at 57% of life. This is potentially due to the asymmetry of the underlying distribution of data due to contamination from a special source of variance (for instance, due to degradation). Sensitivity of the 25th percentile also outperforms the 75th percentile on the healthy bearing, with 32.62% of data being in control (as opposed to 39.43%). The trends exhibited by these features are similar to the other measures of position explored, and are illustrated in figures B.66 and B.63.

Overall, the measures of position perform more consistently than those of other categories (such as variability), and typically produce strong features which can be

employed as metrics for the identification of potential degradation. Given the potential range of positions which could be explored for this purpose, it would be impractical to assess all of these. As the benefit attained from the use of measures of position is that of incorporating outlying or extreme data, subjective judgement should be employed when choosing a measure of position. If the position is too heavily influenced by the main body of the distribution, the information gained may be limited. For instance, the median (the 50th percentile) performs worse (ranked 15th) than the measures of position explored here. Similarly, if the position is too heavily influenced by the tails of the distribution, noise and extreme data which are not correlated to potential degradation may reduce the quality of the attained metric.

Measures of impurity

The measures of impurity explored are interesting for their diverse behaviour and characterisation of degradation. Firstly, the Chao-Shen entropy function was the lowest ranked of all impurity functions explored for their capability to identify degradation. This feature was only able to identify degradation on the failed bearing or potential degradation on the healthy bearing at the later stages of life (95% and 92% of life, respectively). Similarly, this function lacked sensitivity as can be observed in figure B.75 and in table 4.5.

Krichevsky-Trofimov (KT) entropy outperformed Chao-Shen entropy, however, this still performed poorly. Although earlier detection of degradation is attained on the failed bearing (at 71% of life), this is still later than many other features explored. Similarly, KT entropy achieved later detection of potential degradation on the healthy bearing (at 94% of life) than Chao-Shen entropy. Although this technique is more sensitive than Chao-Shen entropy, as can be seen in table 4.5, it is not as sensitive as the other measures of impurity explored in this chapter.

Shannon entropy performed somewhat better in that this feature was able to identify degradation at 61% of life on the failed bearing, with potential degradation on the healthy bearing being identified at 94% of life. This is similar to the result obtained by the James-Stein shrinkage estimator of entropy. This was similarly able to identify the degradation on the failed bearing (also at 61% of life) and able to identify the potential degradation on the healthy bearing (at 94% of life). The James-Stein shrinkage estimator of entropy was as sensitive as the traditional Shannon entropy for both the healthy and failed bearings. It is interesting to note that identical results were reported for both bearing from these features. As such, these features were ranked similarly and observed similar noisy trends, as can be seen in figure B.72 and figure B.84.

Miller-Madow entropy was found to be the most suitable measure of impurity ex-

plored in this case. This feature has a more easily observable trend than the other measures of impurity explored, as can be seen in figure B.81. Miller-Madow entropy was able to identify the degradation on the failed bearing (at 55% of life), and also the potential degradation on the healthy bearing (at 69% of life). Sensitivity was found to be higher for this feature than the other measures of impurity explored. For the failed bearing 53.76% of data was found to be in control, whereas for the healthy bearing, 83.84% of data was found to be in control.

Overall, the measures of impurity perform the least consistently out of all categories of features explored in this work. Although one viable metric is identified (Miller-Madow entropy), the remaining four features are of limited use in practice on this dataset. This could be due to the sensitivity of the techniques (for instance, picking up additional sources of information and thus adding noise to the feature), or simply that these techniques are not well suited to this dataset. Although a viable feature has potentially been identified, further analysis is required to determine is this is due to the quality of feature, or if this is an artefact caused by the underlying data.

4.2.7 Feature validation

Due to the inconsistency attained across various categories of features explored (such as measures of impurity), validation of the features was sought on an independent set of additional bearings. This would serve multiple purposes, namely:

- To ensure the generalisability of the features explored in this context to additional bearings.

- To ensure the consistency of the features explored.

- To ensure stochastic artefacts within the data are not responsible for the previous results.

In order to validate the features explored in section 4.2, an additional dataset was explored. Similarly to the previous experiment, this was taken from the Rexnord bearing dataset made available by NASA (Qiu et al., 2006). The experimental setup employed was identical to that described in section 4.2.3, however, bearings from an independent test were used. These were:

- *Bearing 1, channel 1 (Test 1):* A suspension history (referred to as the "healthy" bearing).

- *Bearing 4, channel 1 (Test 1):* A failure history (referred to as the "failed" bearing).

As in this test data was made available for accelerometers associated to both the X axis and Y axis of the bearing, the X axis was used in this case for both bearings for consistency.

Following the design of the original experiment, EWMA charts were again employed, with the first 10% of data removed from the analysis (due to wear in phenomena), 10% to 40% of bearing life was once again utilised to determine normal operational behaviour. As the same experimental setup is utilised with identical bearings and loading, it is anticipated a similar ranking of features will be attained. However, in the case that inconsistent rankings are attained, this validation assists in identifying features which are both robust to the underlying data and also insensitive to stochastic artefacts within the data.

4.2.8   Results and discussion

As can be seen in table 4.6, various issues of consistency arise. First and foremost, it should be noted that measures of central tendency again perform poorer than expected, compromising five of the six lowest ranked features. Of these, only the RMS was able to identify potential degradation on either bearing (corresponding to 31% of life on the second failed bearing and 37% on the second healthy bearing). As such, measures of central tendency (aside from the RMS) in this case do not perform adequately for the purposes of condition assessment.

With regards to the measures of variability, these remained consistently strong. In this case, the standard deviation and variance remained the two strongest features as in the previous experiment. As such, they are the most consistently strong features explored, outperforming the best features from the other categories of features explored. The median absolute deviation (MAD) and interquartile range (IQR) also performed consistently well in this validation, being ranked in the same region as in the prior experiment. The range, minimum and maximum were also consistent in their rankings.

One feature which underperformed was that of skewness (ranked at $20^{th}/28$ in the second test). This was unexpected given the prevalence of this feature within the literature as a means for condition assessment. It had been anticipated that the poor performance of this feature in the previous experiment was an artefact of the specific bearing, however, the verification shows that this may not be the case. As such, this is potentially due to the data acquisition setup, or the quality of the feature itself. Both of the higher order moments (hyperflatness and hyperskewness) performed consistently as per the prior experiment. It is interesting to note that the peak to average power ratio (PAPR) performed consistently with the crest factor (CF) in this experiment. This was not the case previously, leading to the conclusion that the prior result pertaining

Table 4.6: Performance of time-domain features on validation dataset.

Feature (rank)	Suspension history		Failure history	
	Leaves control (%)	Data in control (%)	Leaves control (%)	Data in control (%)
Measures of central tendency				
RMS (6/28)	37.05	23.42	31.01	10.67
Mean (24/28)	73.05	62.76	N/A	75.88
Median (25/28)	73.05	70.13	N/A	70.83
Winsorized mean (26/28)	N/A	67.86	N/A	73.98
Trimean (27/28)	N/A	72.87	N/A	74.12
Mode (28/28)	N/A	91.98	N/A	86.97
Measures of variability				
SD (1/28)	31.05	24.63	31.05	3.85
Variance (2/28)	31.05	25.42	31.05	3.57
MAD (4/28)	31.05	21.85	31.05	11.92
IQR (7/28)	31.05	25.79	33.05	15.77
Range (17/28)	N/A	82.79	40.05	12.57
Minimum (19/28)	N/A	90.54	50.05	24.54
Maximum (21/28)	N/A	91.56	61.05	25.93
Measures of shape				
Hyperskewness (5/28)	40.05	27.88	31.05	2.55
Hyperflatness (12/28)	74.05	34.14	31.05	3.01
Kurtosis (16/28)	70.05	65.54	32.05	10.53
Skewness (20/28)	N/A	99.07	31.05	39.61
Crest Factor (22/28)	N/A	92.95	50.05	40.54
PAPR (23/28)	N/A	92.95	68.05	40.31
Measures of position				
10th percentile (8/28)	31.05	29.36	31.05	17.35
90th percentile (9/28)	31.05	20.45	41.05	20.50
25th percentile (13/28)	37.05	26.35	61.05	26.67
75th percentile (14/28)	40.05	35.44	60.05	22.73
Measures of impurity				
KT (3/28)	31.05	27.46	31.05	4.04
Shannon (10/28)	39.05	32.61	41.05	14.19
James Stein (10/28)	39.05	32.61	41.05	14.19
Chao-Shen (15/28)	76.05	36.09	41.05	14.42
Miller-Madow (18/28)	N/A	73.24	47.05	28.20

to the potential strength of the crest factor as a feature, was an artefact of the data employed, and not of the quality of the feature itself. Overall, the measures of shape performed better in this second experiment than in the prior experiment.

The measures of position explored within this chapter performed more consistently than other measures explored. Specifically, the 10th and 90th percentiles performed comparably (ranked 8th and 9th, respectively), along with the 25th and 75th percentiles (ranked 13th and 14th, respectively). Given that the positions towards the tails of the distributions performed better than the quartiles and the median, coupled with the

strong performance of the measures of variability, it can be inferred that in this case degradation causes increased variance within the distribution. This would assist in explaining the quality of the explored measures of variability.

Furthermore, with regards to the measures of impurity, aside from Krichevsky-Trofimov entropy (ranked 3rd of 28), the four remaining measures of impurity were ranked from 10th through to 18th. This is contradictory to the previous experiment (where typical rankings of 21st to 26th were attained). This adds further weight to the argument that degradation is inherently additional information incorporated into the underlying bearing signal, although further empirical evidence should be sought due to the inconsistent nature of these results.

4.2.9 Feature ranking

Due to the inconsistencies obtained in the ranking of predictors through the two experiments, a final ranking of features was sought. In order to derive a ranking, the score (S_f) of each feature was normalised using min-max normalisation and averaged across the two tests. This was preferred over utilising averaged ranked scores due to the nature of degradation. As degradation became apparent at differing stages in each test (as expected), this would inherently bias scores towards those features which performed stronger in the test with the earliest identifiable degradation. Although utilising the average ranked score could have been employed, it was felt that ranking based upon normalised score was more intuitive. The final ranking of features attained through this work can be found in table 4.7.

4.2.10 Limitations and threats to validity

Due to the nature of degradation, limitations exist in the experimental design employed in this work. These limitations are discussed in detail here.

First and foremost is the use of statistical process control in order to quantify and rank the features explored. In the context of this work, this is a natural paradigm for the exploration of features. However, in the case that external factors are of importance (such as a wind turbine gearbox), the applicability of this methodology would be limited. In this case, a traditional ARIMA methodology should be explored, with the addition of a seasonal component (SARIMA) in the case of asset or components affected by seasonal variations. Similarly, it is anticipated that for non-stationary signals, the applicability of this methodology is limited. In this case, re-sampling into the revolutionary domain as per Eric Bechhoefer (2009) would be recommended prior to this analysis.

Table 4.7: Overall ranking of bearing features.

Feature	Test 1 Score	Test 2 Score	Final Rank
Standard deviation	41.43	22.64	1
Variance	41.43	22.77	2
RMS	41.43	25.55	3
Median absolute deviation	43.51	23.96	4
10th percentile	43.39	27.20	5
90th percentile	43.34	28.26	6
IQR	44.91	26.41	7
Hyperflatness	45.03	35.56	8
25th percentile	45.19	37.78	9
75th percentile	49.35	39.56	10
Hyperskewness	70.07	25.38	11
KT entropy	80.45	23.40	12
Crest factor	43.95	70.88	13
James-Stein shrinkage entropy	76.15	31.72	14
Shannon entropy	76.15	31.72	14
Kurtosis	67.25	44.54	16
Miller-Madow entropy	65.45	62.12	17
Range	71.76	58.85	18
Chao-Shen entropy	91.49	41.90	19
Minimum	72.78	66.28	20
Maximum	72.79	69.63	21
Skewness	77.87	67.43	22
Median	71.14	78.50	23
Winsorized mean	72.49	85.46	24
Tri-mean	73.23	86.75	25
PAPR	91.89	75.33	26
Mean	90.18	77.92	27
Mode	98.12	94.74	28

Secondly, the use of statistical process control requires an objectively defined period to regard as normal operational behaviour. In this case, data from 10% to 50% of life was used for this training period. In both experiments, if degradation is apparent within this training period, the sensitivity of techniques from a statistical process control perspective is reduced. By moving away from a traditional statistical process control methodology and using this paradigm as the basis of a metric to quantify the features (rather than utilising SPC as a basis to justify maintenance actions on the bearings themselves), the implications of this are mitigated. If in the event data leaves control immediately after the training period, this would be a basis to review the data employed for the statistical process control chart. To ensure that this had not occurred within this work, the training period was revised to between 10% to 20% of bearing life. For both bearings within both experiments, the results attained from this for the features did not effect the ranking of the features in either experiment.

Thirdly, the use of a statistical process control methodology potentially limits the identified capability of some features. Although some features leave control early indicating the identification of degradation, some return to within control limits shortly afterwards. This is demonstrated by skewness in figure B.42. Although this is detrimental to the overall quality of the feature (and is reflected as such in the "data in control (%)" value), this signature itself may be beneficial in the assessment of bearing condition. It is possibly due to the high sensitivity of the feature in conjunction with the healing phenomena observed on bearings (Qiu et al., 2006) that this occurs. As such, it may be the case that features are inadvertently being penalised due to this higher sensitivity.

Fourthly, it should be noted that each bearing has a unique signature as can be seen in the figures presented in appendix B. This is due to two main reasons. Firstly, slight variation in the manufacture and the presence of defects on a given bearing will change the degradation path and behaviour of the component. Secondly, due to the 6000 lbs radial load applied, different bearings within the test rig will have been subjected to slightly different variance in this load, thus changing the behaviour of the bearing and the data which was collected.

Finally, with regards to the underlying data, many artefacts are present within the data. Specifically, for the second experiment performed, various points within the data can be identified as being anomalous. As the data utilised in this experiment was collected by a third party, no comment can be given as to the nature of the cause of these artefacts. There is the potential for both external factors (such as physical movement of the experimental setup or power surges effecting the drive motor) and internal factors (such as manufacturing defects or healing phenomena) to be responsible for these. Given the extreme deviation of these points and their rarity, these can be assumed to be outliers and handled as such if necessary.

4.2.11   Derived bearing model

Based upon the analysis performed in this section, a a multivariate model for the condition assessment of bearings is proposed. This is a 4-dimensional model and encompasses the following features:

- Standard deviation.

- Root mean square (RMS).

- Skewness.

- Kurtosis.

The standard deviation was selected due to strong performance in both of the experiments (ranking first in the first experiment and second in the second experiment). Furthermore, this is always a positive value and is widely utilised in many areas of reliability analysis.

The signal RMS was also selected due to the strong performance in both experiments. Although this feature was ranked sixth in the second experiment (and ranked joint first in the first experiment), this feature is widely recognised and understood. It is often used in signal analysis and is widely employed in practice already. Similarly to the standard deviation, this value is always positive.

The skewness was selected based upon prevalence within the literature, rather than the performance in the experiments undertaken in this work. As skewness did not perform as well as had been anticipated (ranked 23rd and 20th in the experiments, respectively) it is likely that this is due to utilising two similar forms of data for this analysis (both sets utilised the same experimental test setup for data acquisition).

The kurtosis was chosen for reasons on a similar basis to that of skewness. Kurtosis is widely prevalent within the literature and typically utilised in conjunction with skewness. Although kurtosis performed better than skewness (ranked 13th and 16th in the experiments, respectively), it is likely that this is due to the dataset employed, rather than the quality of the feature itself.

Although the skewness and kurtosis did not rank highly in this exploratory work, this may be due to the metrics employed for feature ranking. Both of these features leave control limits at similar moments to stronger features. However, due to their return to control – perhaps due to healing phenomena as discussed by Qiu et al. (2006) – are penalised by the metrics employed here. As such, it may be due to this that the features appear to perform worse than anticipated.

It is interesting to note that this model is made up of the first four moments. Although RMS is not the first moment, as it estimates the central point of a distribution (as the first moment traditionally does), this feature is chosen for its superior performance to the other measures of central tendency. The standard deviation (the second moment), skewness (the third moment) and the kurtosis (the fourth moment) provide a means to encapsulate the distribution of the behaviour of the bearing. Given the potential use of the higher moments (hyperskewness and hyperflatness), the potential exists to explore these in conjunction with the first four traditional moments, however, further work is necessary to explore the impact of these features.

As such, the model compromises four features from across three distinct categories of statistical time domain features. Namely, a measure of central tendency (RMS), a measure of variability (standard deviation), and two measures of shape (skewness

and kurtosis). As such, it is anticipated that this model will enable the encapsulation of bearing operational behaviour for the purposes of early detection, diagnosis and prognosis. For the purposes of this thesis, this model will be referred to as the "Moments model" for the purposes of distinguishing this model from previous literature.

4.2.12 Conclusion

This section has detailed a methodology to quantify the capability of suitable time-domain statistical features for the purposes of high frequency bearing condition assessment. In total 28 features were explored, across five distinct categories of features. Namely, these were; measures of central tendency, measures of variability, measures of shape, measures of position and measures of impurity. A robust statistical process control paradigm was employed utilising exponentially weighted moving average (EWMA) charts to determine the suitability of features based on their capabilities of both early detection and also their sensitivity. Preliminary exploration of the features was performed, with validation of the methodology and features performed on an additional independent dataset. Measures of variability were found to perform the most consistently across both tests, with measures of central tendency typically performing worse than expected. Features more commonly explored within the literature such as the median and the peak-to-average-power ratio (PAPR) were found to be out performed by many features, including traditional features such as measures of shape (specifically, skewness and kurtosis) and also measures of impurity (specifically, Shannon entropy).

4.3   Gearbox health model

4.3.1   Introduction

In order to contrast against the data-driven development of the high frequency bearing condition model explored above, the second condition model explored in this thesis explores the potential use of low frequency SCADA data for the purposes of condition assessment and health monitoring of the wind turbine gearbox.

As discussed in section 1.5, the wind turbine gearbox is the most critical component of the wind turbine, with an estimated 40 - 80 days of downtime should an offshore failure occur (Dinwoodie et al., 2012; Tavner, 2012). Due to the high consequence of failure of the gearbox, failure data is scarce. This is due to the over-maintenance of critical components due to the high consequence of failure; it is often simpler to replace components which do not necessitate maintenance simply due to the criticality of the component. This ensures high availability of the component, however, is far from optimal in economic terms. As such, the verification and validation of a potential condition assessment model is difficult empirically, as few failure signatures exist within the data itself.

Due to this – and utilising notions expressed in section 4.2 – by accurately and precisely encapsulating normal operational behaviour of a wind turbine gearbox, it is then possible to monitor the residuals of this model to determine the current health of the gearbox. As over 95% of wind turbine gearbox failures are due to gear or bearing failures (as explored in section 1.5.1), it is anticipated that a model capable of modelling the behaviour of these components would assist in a reduction of failures, or a reduction in the over-maintenance of components; reducing overall annual maintenance budgets.

Building upon the notions of distance (or similarity) and normal operational behaviour which have been at the core of this thesis so far, the development of a wind turbine gearbox model within this paradigm was sought.

Within the context of wind turbines, the expected behaviour of a wind turbine is typically characterised by the power curve. The power curve for a Vestas 1650 kW wind turbine is presented in figure 4.5. As can be seen, the turbine is characterised in terms of the power production for a given wind speed. Due to uncertainty in the forecast of wind, it may be necessary to curtail or de-rate the turbine to reduce the quantity of generation delivered. De-rating the turbine limits the maximum production by a specified amount, whereas curtailment reduces the overall power output by a given percentage of production capability. This enables flexibility in the quantity of generation and the grid frequency response (amongst other factors).

The power curve provides a natural means to integrate wind turbine behaviour with
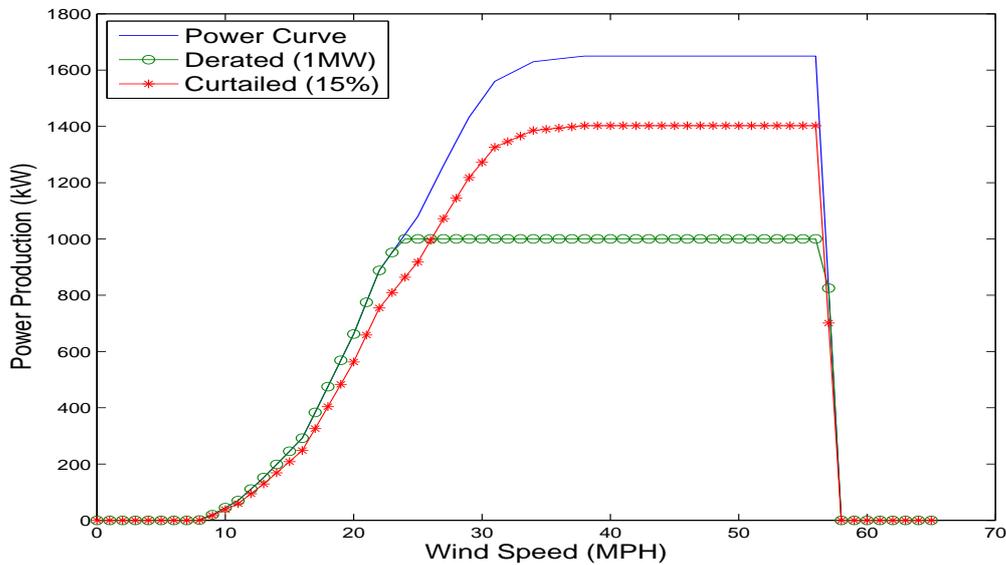
Figure 4.5: Example Vestas 1.65 MW wind turbine power curve with additional power curves showing curtailment (by 15%) and de-rating (to 1 MW).

a condition monitoring approach based upon the encapsulation of normal behaviour. It is intuitive to think of degradation as a deviation from the power curve for a variety of reasons. For instance, if the wind turbine is designed to produce 1650 kW at maximum output, however, is only currently delivering 1500 kW, this 150 kW discrepancy can potentially be due to increased inefficiencies within the gearbox, reducing the quantity of power generated.

However, curtailment and de-rating of the turbine cause degradation like artefacts within these models. For instance, in the work undertaken in Kusiak and Verma (2013), clustering of the power curve is performed to characterise behaviour. These clusterings are only valid under normal operational behaviour, and not in derated or curtailed scenarios. In many cases, due to proprietary OEM data concerns, curtailment or de-rating information is not available within the SCADA data collected from a wind turbine. This data was not available in the Reliawind SCADA data repository, and as such, alternative approaches were required in order to quantify degradation within the gearbox.

Dataset employed

In order to verify or validate a model of gearbox behaviour, it is beneficial to utilise data collected during both normal operational behaviour of the gearbox, and also data containing failure signatures prior to, during, and immediately after a failure. For this

purpose, data from the Reliawind consortium was utilised (Wilkinson et al., 2010).

Data from the Reliawind consortium contains data from across six various manufacturers, with low frequency SCADA data and high frequency condition monitoring and on demand diagnostic data from over 100 wind turbines. Due to the commercial sensitivity of this data, identifiable information has been anonymised. For the purposes of this thesis, data from one manufacturer is employed, from one wind farm. This wind farm comprised over 20 wind turbines, each with a power rating of 1.5 MW peak. Low frequency SCADA data (600 second sample rate) across 190 channels is explored throughout this thesis. In total, data was collected over a period of 28 months, from 6 wind turbines. Due to the scarcity of catastrophic wind turbine gearbox failure data, only one instance was available for analysis throughout this thesis. Due to the difficulty in deriving empirical, data-driven models from a single failure, alternative approaches were sought.

4.3.2 Gearbox model identification

The systematic review of the literature undertaken in chapter 2 identified multiple models which could be exploited for the purposes of condition assessment of the wind turbine gearbox. Many of these models exploited the notion of normal operational behaviour in order to identify states of potential degradation or damage. The most notable of the models within the literature are described below.

Gacia, Bobi and Pico, 2006

In their paper Garcia et al. (2006), the *"SIMAP"* system is presented which was developed to provide a comprehensive means for the detection and the identification of wind turbine gearbox cooling system faults. In their work, a model of the wind turbine gearbox is described based upon physical laws. This is then exploited utilising bearing temperature, nacelle temperature, power generated and the current behaviour of the cooling system for use in an artificial neural network.

Training of the neural network on data associated with normal behaviour enabled the identification of anomalous data. A prediction of the current state based upon previous states can be performed, with the residual of this value associating with potential degradation. During normal operational behaviour, the predicted values will have a lower error than those during periods of degraded or failed behaviour (due to the training process). By identifying data outside of the expected 95% confidence intervals, anomalies are identified. Following this, fuzzy rules are explored and utilised in order to provide transparent rules for the system to perform diagnosis.

Zaher et al. 2009

In the work of Zaher et al. (2009), the SIMAP model is extended and refined. Fewer model inputs are utilised, but a similar methodology is employed. SCADA data collected at 10 minute intervals is utilised; similar in many regards to the data available for this thesis. Three months of training data were utilised to train the neural network to identify normal behaviour. After validating the approach in a similar manner to that of SIMAP, it is extended through the use of multi-agent systems (MAS) to assist in the management of the wind turbine.

This work differs from Garcia et al. (2006) in that the focus is on the gearbox rather than the cooling system. As such, both of these models could be combined in an ensemble approach to provide a holistic means for the detection of gearbox subsystem faults. As both models exploit neural networks, their expert knowledge base (the rules which represent the underlying behaviour of the component) are not transparent. As such, alternative models which do not rely on these techniques were sought.

Kusiak and Verma, 2013

Differing from the previous work, in their analysis, the work undertaken in Kusiak and Verma (2013) exploits three performance curves for wind turbine gearbox condition assessment. The curves employed are that of the power curve (wind speed vs. power produced), the rotor curve (wind speed vs. rotor speed) and the blade pitch curve (wind speed vs. blade pitch angle). Similarly to Zaher et al. (2009), 10 minute interval SCADA data is employed for the analysis.

Distance metrics are explored, with the performance curves being broken down into smaller clusters in order to exploit the performance charts. Although strong results are achieved, extensive manipulation of the data is required (Box-Cox transformation, multivariate control charts, and derived multivariate skewness and kurtosis).

These manipulations are required as it is not possible to define a central point on these control charts from which to measure deviation from, and as such, the notion of distance is nonsensical. Furthermore, the authors acknowledge this as a limitation of the research and accept that curtailment of the wind turbine is identified as abnormal conditions. No neural networks are utilised in this research, and as such, the technique requires less data a priori and is more transparent overall.

Feng et al., 2011

Finally, in the work of Feng et al. (2011) a different approach is taken. A model built upon the laws of thermodynamics is built, modelled by relating inefficiencies of gears

to the work being performed. In their paper, the model validity is demonstrated in multiple ways, with three new performance charts being proposed. These consist of a histogram with data binned by month, a plot of rotor speed against temperature rise and a plot of power output against temperature rise. For their analysis, 10 minute interval SCADA data was employed.

This model requires no a priori knowledge, no "black box" techniques – such as the neural networks employed in Garcia et al. (2006) and Zaher et al. (2009) – and no extensive manipulation of the data such as in Kusiak and Verma (2013). Given the reduced computational complexity of this model in comparison to other available models, the model transparency and in keeping of the theme of normal operational behaviour and similarity, this model was selected for use in this thesis.

4.3.3 Gearbox model development

The gearbox model explored in this thesis build upon the work undertaken in Feng et al. (2011). This gearbox model based upon physical models is described as in their work as follows. Firstly, the model assumes that the heat generated by a gear is proportional to the work done by the gear:

$$q \propto w \propto \Delta t \tag{4.8}$$

Where the heat generated is denoted by q, the work done is denoted by w and the rise in temperature is denoted Δt. It follows that for the efficiency of a given gear, denoted η_{gear}, the energy is dissipated as heat to the gear, thus giving:

$$Q_{gear} = (1 - \eta_{gear})\frac{1}{2}l_{gear}\omega_{gear}^2 = k_{gear}\Delta t_{gear} \tag{4.9}$$

Which can be expressed in notion of inefficiency as:

$$1 - \eta_{gear} = \frac{2k_{gear}\Delta t_{gear}}{l_{gear}\omega_{gear}^2} \tag{4.10}$$

As the work done by the wind turbine gearbox can be expressed in the form:

$$P_{out} = W - Q_{gear} \tag{4.11}$$

Where P_{out} denotes the power output, W denotes the work done and Q_{gear} denotes the energy dissipated (as from equation 4.9). It follows that:

$$P_{out} = \eta_{gear}\frac{1}{2}l_{gear}\omega_{gear}^2 \tag{4.12}$$

This is manipulated in Feng et al. (2011) to give:

$$\frac{1 - \eta_{gear}}{\eta_{gear}} = k_{gear}\frac{\Delta t_{gear}}{P_{out}} \tag{4.13}$$

Which, when rearranged in terms of Δt_{gear} gives:

$$\Delta t_{gear} = P_{out}\frac{1}{k_{gear}}(\frac{1}{\eta_{gear}} - 1) \tag{4.14}$$

Thus, this model describes the relationship between the efficiency of a gear (η_{gear}) and the rise in temperature (Δt_{gear}). As such, it holds that as a gear wears the efficiency of the gear reduces, and the temperature (Δt_{gear}) must increase for the same generation of power (P_{out}).

This enables the identification of normal operational behaviour by empirically defining Δt_{gear} for a given power output. Thus, determining Δt_{gear} for all power outputs enables the indirect quantification of the inefficiencies of the gears within the gearbox.

Model limitations

Due to the simplicity of the model developed in Feng et al. (2011), noise is introduced through additional factors which must be taken into consideration when employing this model in practice. This thesis utilises and extends this state-of-the-art model to normalise low frequency SCADA data for:

- **External conditions:** Seasonal variations effecting temperature will add noise to the analysis, and must be normalised for.

- **Turbine loading:** As the gearbox is expected to produce more heat at higher power outputs, this must be taken into consideration.

- **Transient states:** Due to the variability of wind, sudden gusts may increase power production rapidly over a short period. This would increase the heat disproportionally to what would otherwise be expected, and lead to the potential false positive identification of degradation if not normalised.

- **Wind farm topology:** Wake effects from nearby turbines could potentially interfere with power production, and should be taken into consideration.

These normalisations provide a means to extended the physics of failure model presented in Feng et al. (2011) to overcome these limitations. For simplicity, the physics of failure model incorporating these normalisations will be referred to as the "extended

physics of failure model" in order to distinguish this model from previous research. The normalisations undertaken in this model are discussed in detail below.

4.3.4   Normalisation against external conditions

First, normalisation against external conditions is performed. Due to the seasonal variation of temperature, if this is not taken into consideration, false positive degradation may appear in the summer months due to heat. Similarly, this could disappear in the winter months and cause potential false negative readings. Three potential data sources can be used for this normalisation, namely:

- Wind farm ambient temperature.

- Wind turbine ambient temperature.

- Wind turbine nacelle temperature.

The first of these is immediately discounted; although the ambient temperature of the wind farm would provide a potential means to of normalisation, the readings available at the wind turbine would provide a more reliable means to perform the normalisation. The model developed by Feng et al. (2011) uses the nacelle temperature.

Analysis utilising Spearman's Rho correlation between the wind turbine ambient temperature and wind turbine nacelle temperature taken from the SCADA system of 3 wind turbines over a period of 28 months revealed a statistically significant relationship ($r_s(370390) = .84, p < .001$), this is regarded as a strong relationship (Dancey and Reidy, 2007) and demonstrates that the ambient temperature at the wind turbine is a viable alternative in the case of nacelle temperature data not being available.

4.3.5   Normalisation against turbine loading

The first contribution of this thesis to the gearbox model is the specification of a robust performance chart which is valid regardless of turbine de-rating or curtailment. Although not realised in their original work, the model developed by Feng et al. (2011) overcomes the limitations of the performance charts expressed in Kusiak and Verma (2013) in that regardless of the current derating or curtailment of the wind turbine, the chart is still valid. As such, by employing this model as a performance chart we have a novel means to perform analysis based upon the empirically determined normal operational behaviour of the gearbox.

By plotting the rise in temperature (from the external conditions baseline described above), against wind turbine power output, if low frequency SCADA gearbox data is

employed to empirically and robustly define the central point of these temperature rises, this provides a baseline of current health which can be monitored over time.

Power production was binned into 50 kW bins as per Feng et al. (2011). This enables similar behaviours to be grouped together, whilst mitigating the issues caused by data scarcity. Due to the low frequency of SCADA data, it is possible that minimum data is collected for a wind turbine for various loading parameters. As such, the distribution of values for some power production levels may not be sufficient to adequately determine a robust central point. By binning data into 50 kW bins, this is avoided whilst characterising the behaviour of the gearbox. Inherently, some trade-off exists: smaller bins may enable a more accurate reflection of gearbox condition, but there may not be sufficient data to adequately determine the normal operational behaviour for each bin. Similarly, larger bins may reduce model accuracy, but are more likely to contain sufficient data to adequately determine the normal operational behaviour of the gearbox.

As can be seen in figure 4.6, as the gearbox generates more power, the temperature increases. This is to be expected and assists in the verification of the normalisation performed. At approximately the 500 kW point, a drop in gearbox temperature is observed. It is likely that this is due to the turbine moving from a static pitch configuration to a variable pitch configuration at around this power production level. It
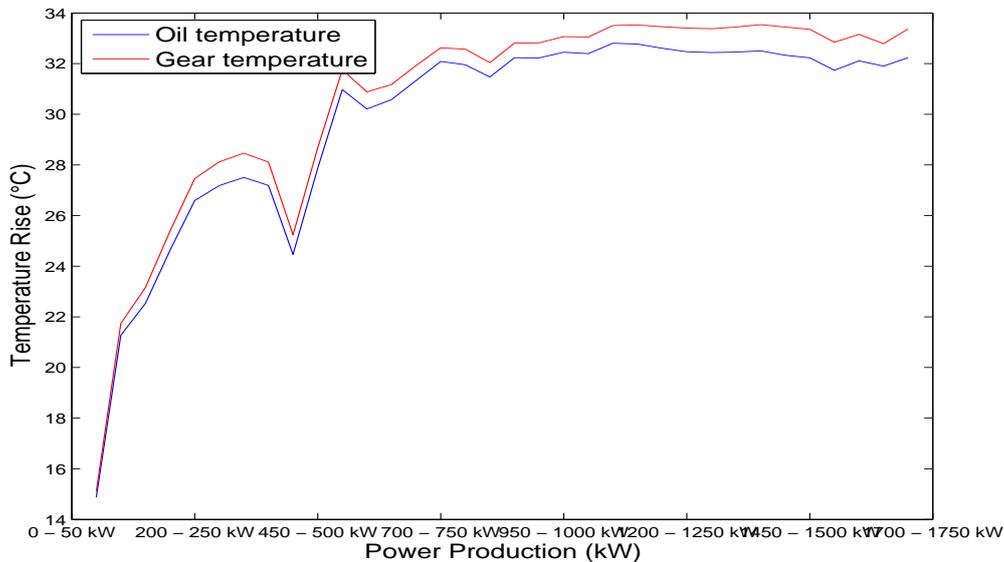
Figure 4.6: Normalisation of SCADA data against turbine loading based upon 10 minute SCADA data.

is also interesting to note that the gear temperature is consistently hotter than the oil temperature.

4.3.6 Normalisation against transient states

The first novel contribution that this thesis makes to the physics of failure model of Feng et al. (2011) is that of normalisation against the transient states which a wind turbine is typically subjected to. Due to the stochastic nature of wind – and the nature of this model in particular – this normalisation is important in the assistance of the removal of noise from the underlying SCADA signal.

For instance, due to the constant sampling rate of the SCADA system, various behaviours (such as gusts) may not be captured. As such, it is possible that phenomena such as strong gusts could cause noise in the model by moving the turbine from a non-producing state to producing at maximal capacity within a single recorded SCADA period. Due to the lagging nature of the SCADA temperature data, the normalisation against turbine loading detailed above would be incorrect. In this case, by approximately 15 degrees (approximately 4 deviations).

As such, normalisation against these transient phenomena is performed. This is done by empirically defining the typical (central) offset required for each possible transition state. For reasons similar to those detailed previously, this is done in discrete 50 kW bins. Each bin thus contains data pertaining to the temperature offset which typically occurs for a power generation transition of a specified size (from -1750 kW to 1750 kW; a total of 70 transitions). Only SCADA records representing transient states are employed (consecutive SCADA records representing transitions between discrete bins) to determine the typical rise (or fall) in temperature representing a transient state. For each transient record, the difference in temperature between the previous state and current state is placed into its relative discrete bin. For each discrete bin, the arithmetic mean is taken as the central point for the typical rise (or fall) of gearbox temperature due to the observed transient phenomena. This is shown in figure 4.7.

As can be seen in figure 4.7, rapid transitions between states of low power generation and high power generation have the potential to cause rapid changes in the temperature of the wind turbine gearbox. This is to be expected given the nature of this asset, and assists in the verification of this normalisation to ensure that it is fit for purpose. Normalisation is then performed by offsetting the model value (which has been normalised for loading) against this newly derived offset (for transient states). As such, the model is normalised against external conditions, the loading of the wind turbine and the transient nature of wind.

Performing this normalisation helps removes artefacts in the time series caused by
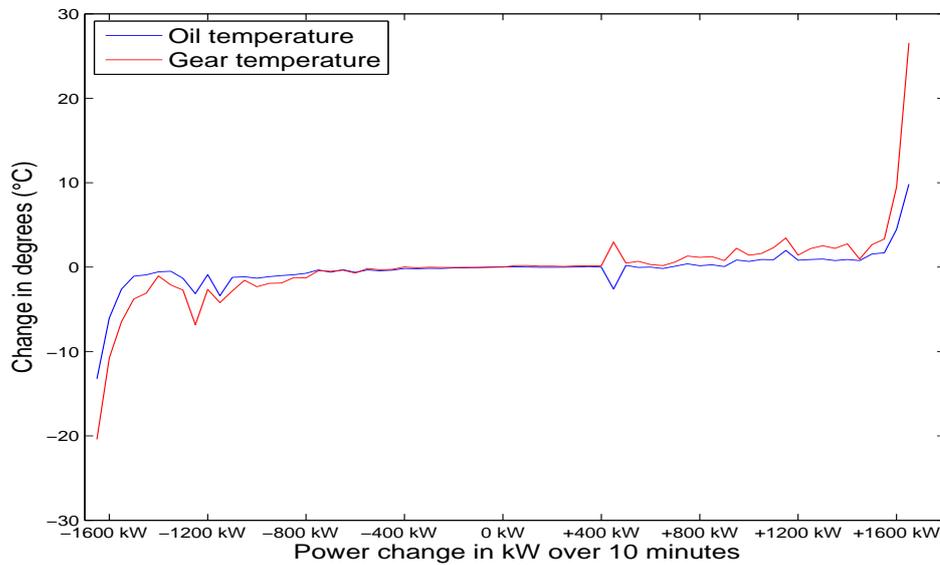
Figure 4.7: Normalisation of SCADA data against transient states based upon 10 minute SCADA data.

these transient phenomena.

4.3.7   Normalisation against wind farm topology

The final normalisation which can be undertaken on the model is against the role of wake effects which can occur within a wind farm. Wake is the name given to the highly turbulent air behind a wind turbine. As the wind turbine extracts energy from the air, a proportion of energy will not extracted due to the Betz limit (Albring, 1967). This remaining air contains a lower quantity of energy, and is typically much more turbulent (Tavner, 2012). Often in a wind farm, turbines are placed in a grid formation. Due to these wake effects, wind turbines located down-wind of other turbines in this formation will typically be presented with lower quality wind, which is more turbulent and potentially fatigues or degrades the components of the wind turbine at an increased rate than less turbulent air.

Inherently, the extended wind turbine gearbox model presented in this thesis takes wind farm wakes into consideration. A wind turbine within the wake zone of other wind turbines will inherently produce less power. As the performance chart which is exploited in this thesis utilises the temperature rise plotted against the power generated, this is taken into account. However, the extent of the wake effect (and thus its influence on down-wind turbines) will vary dependent upon the wind direction and as such, must

be considered.

Given the economic implications of poor planning, wind farms (both onshore and offshore) typically exploit locations which afford a high wind speed from a prevalent wind direction. Due to this, a significant proportion of wind comes from a known direction, and this information is exploited prior to the installation of the wind farm for the purposes of ROI assessment. As such, many of these wake interactions will be understood prior to the installation the farm, with the wind farm topology optimised to minimize the periods of wake encountered in order to increase production. Thus, periods of wake interaction will typically be minimal.

As each wind turbine would need to be normalised independently for the influence of wake effects, and given that only a small quantity of wind directions for a farm would produce wake which are not inherently captured in the normalisation against transient states (as discussed above), data scarcity becomes an issue. Similar to the previous normalisations, aggregation across directions would be required in order to populate the distribution to derive a central point from which we can normalise the effect of the wake. This is further exacerbated by the requirement to do this for each of the discrete turbine loading bins we have previously normalised our data into. As such, without significant quantities of historical data from this wind turbine in particular, normalising the model against the effect of wake through empirical wind direction is not practical. Utilising the normalisation against transient states as a normalisation to take into account the wake effects is therefore recommended.

Furthermore, two further issues exist with normalising the model based upon wind direction. Firstly, as we do not know if a given wind turbine is curtailed or de-rated, it would be difficult to determine the influence of wake effects empirically as we would not be able to identify data points caused by a down-wind turbine being curtailed or de-rated (artificially reducing the effect of the identified wake on a given turbine). This would introduce additional noise into the model. Secondly, as in this case we do not know the topology of the wind farm, it is difficult to empirically verify the normalisation to ensure that it is performing as expected. By removing the requirement of this a priori data (by utilising the normalisation against transient states), we can enhance the robustness of the model.

In the case that the topology of a typical wind farm is known and curtailment data is available, it is possible to empirically model these interactions to take wake effects into consideration. Work done by Yan and Zhang (2014) utilises historical low frequency SCADA data to empirically determine the influence of wake effects on a given wind turbine. In their work, a metric (the normalised wind speed difference, or WSD) is employed to construct performance charts where this wind speed difference between

two wind turbines is plotted against the wind direction (in degrees). This metric is defined as Yan and Zhang (2014):

$$WSD_3 = \frac{ws_1 - ws_2}{(ws_1 - ws_2)/2} = 2 - \frac{4}{ws_1 - ws_2 + 1} \tag{4.15}$$

Where in this case ws_i represents the wind speed of wind turbine i. As such, given two adjacent wind turbines, it is possible to create a performance chart showing the interactions between the turbines due to wake effects. It would therefore be possible to create revised power curves for a wind turbine based upon empirical wind data, accounting for these interactions. This is an alternative in cases where temperature data may not be available.

4.3.8   Results

Given the rarity of gearbox failures in our dataset (due to their design life and potential over-maintenance), only one failure is present for analysis in this thesis. As such, the results here pertain to data taken from 3 wind turbines over the course of 28 months. Due to the commercially sensitive nature of the data, physical locations and wind turbine model information cannot be provided. The turbines are as follows:

1. **Turbine 1:** This represents a wind turbine which suffered no gearbox failure over the specified period.

2. **Turbine 2:** This represents a wind turbine which suffered a catastrophic gearbox failure during the specified period.

3. **Turbine 3:** This represents a wind turbine which suffered no gearbox failure over the specified period.

As the model presented here is viable for both gearbox oil temperature and gearbox gear temperatures, both of these are presented for each turbine. Three independent wind turbines were used to determine the empirical normalisations against both turbine loading and also transient states. This ensured that the model could be employed on new wind turbines in practice as the normalisations were independent of the three turbines explored in this work. Utilising the three turbines of interest in this work to set these parameters may potentially over-fit the model and as such, may reduce the generalisability of the technique to other wind turbines. These three turbines are discussed below.

Turbine 1

Figure 4.8 presents the extended physics of failure model values for both the planetary gear temperature and gearbox oil temperature, aggregated daily (using the traditional mean). As the figure shows, both of these features are strongly correlated. Pearsons' product moment correlation coefficient revealed a correlation of $r(879) = .99, p < .001$. This suggests that a uni-variate model utilising this physics of failure technique may be suitable for the analysis of wind turbine gearbox condition. However, it should be noted that as they are not perfectly correlated, benefit may be gained from analysis in the bivariate domain in this case.

Although many negative artefacts can be observed, these are similarly observed in figure 4.9 and figure 4.10. This due to two factors: firstly, instances of missing and erroneous data from the wind turbine SCADA system, and secondly, the aggregation of these values using the mean to provide a quantitative value for the gearbox condition each day. As all three wind turbines were located on the same farm, these aggregations are all comparable in this case.

Turbine 2

In this case, a catastrophic failure of the wind turbine gearbox occurred between days 300 and 400. This is clearly visible in figure 4.9, assisting in the verification of the
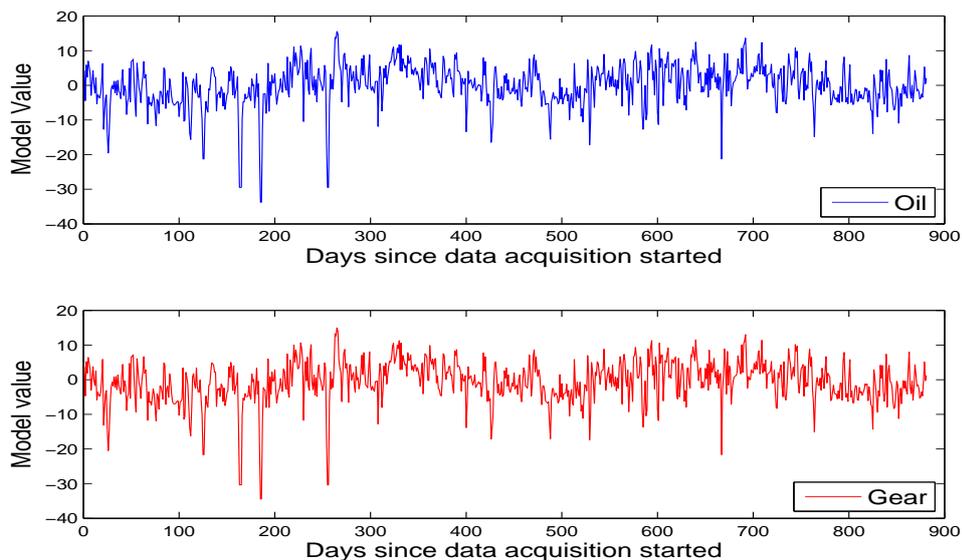
Figure 4.8: Wind turbine 1 health metric over 2 months.
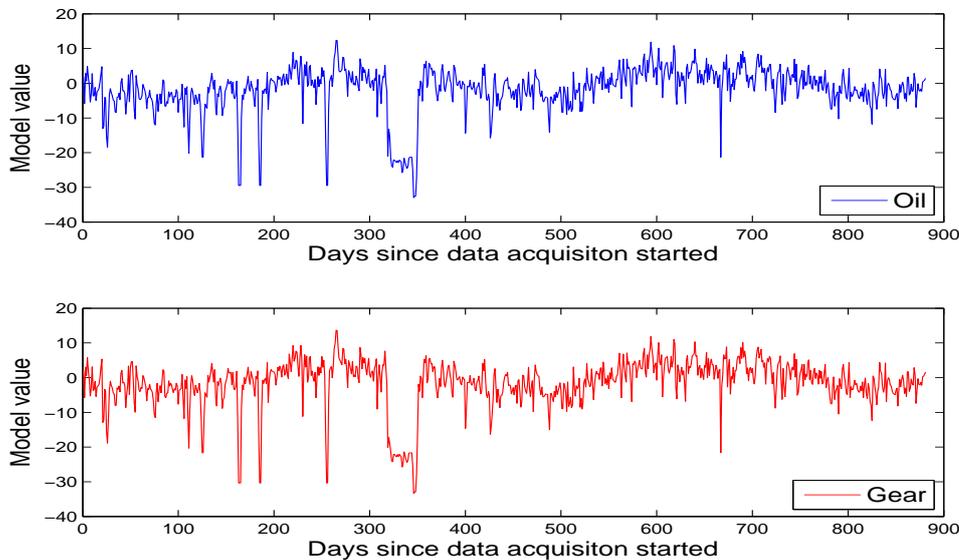
Figure 4.9: Wind turbine 2 health metric over 2 months.

extended physics of failure model. As this is not observed in either figure 4.8 or figure 4.10, and given the failure date in the maintenance logs, we can be almost certain that the failure occurred at this point.

Similar to turbine 1, the oil temperature and planetary gear temperature in this case also appear highly correlated. Pearsons' product moment correlation coefficient also revealed a strong correlation in this case ($r(879) = .99, p < .01$). As in figure 4.8, large negative artefacts are present in the model time series. These occur at precisely the same location as in both figure 4.8 and figure 4.10. As such, this provides some further evidence that these artefacts are due to poor data quality, rather than the developed model itself.

Turbine 3

In the final wind turbine, no failure was present. This is shown in figure 4.10. Both model variables were again strongly correlated ($r(879) = .99, p < .01$). Negative artefacts occur at the same location as both previous wind turbines. This would suggest that the problem with the SCADA system was across the entire wind farm (a problem with the central data acquisition server, rather than the wind turbine SCADA system itself). This was verified by manual observation of the acquired data, noting substantial periods of missing data during these artefacts. In some instances, data acquisition was offline for a consecutive period of almost 4 days. During these periods, it is not possible
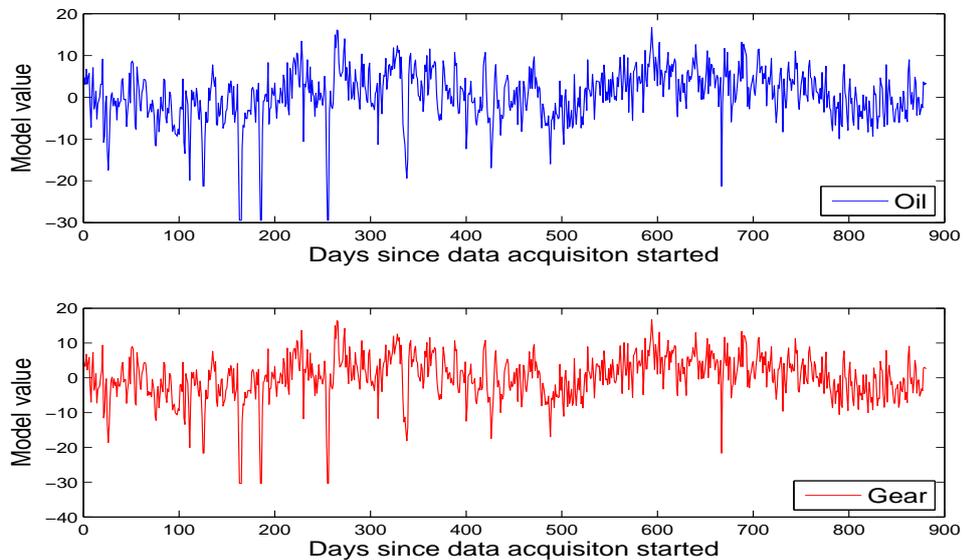
Figure 4.10: Wind turbine 3 health metric over 2 months.

to determine the gearbox health using this model. However, as these periods can be identified, alternative measures can be taken.

It is interesting to note that in all three time series, some seasonal variation can be observed. This would suggest that the normalisation presented by Feng et al. (2011) is not normalising for external conditions as effectively as is possible.

4.3.9   Discussion

Firstly, the quality of data should be discussed. The artefacts within the extended physics of failure model time series can be attributed to this. As such, it is interesting to explore this problem. Due to this, missing SCADA records (aggregated daily) were quantified and plotted in order to define the extent of this issue. This can be seen in figure 4.11. In terms of availability, the SCADA system for this wind turbine was available 97.46% of the time – or was unavailable 2.54% of the time – without taking erroneous, implausible or duplicate records into consideration. In total, over a period of 28 months, 3222 SCADA records were missing (approximately 537 hours of data). This represents an average of 115 records per month (approximately 19 hours of data per month).

Secondly, it should be noted that in figures 4.8 through to 4.10, seasonal variation can be observed. Although this is limited due to the normalisation undertaken, it is still present. This is potentially due to the thermal inertia of the gearbox. As this is
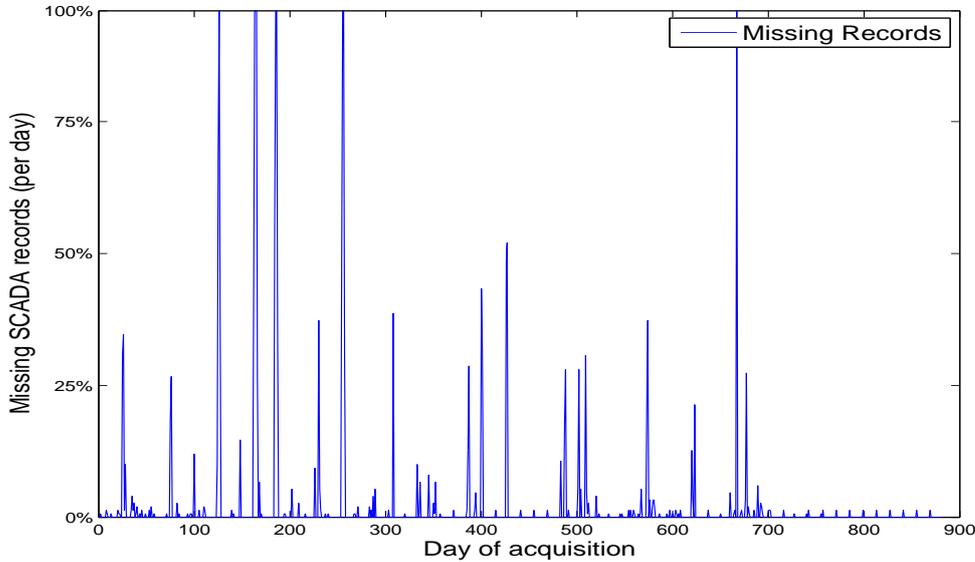
Figure 4.11: Number of missing SCADA records over time.

not modelled due to the simplistic nature of this model, it is not surprising that this is present. Further normalisation of this would require empirical data not available during the course of this thesis, and as such, is beyond the scope of this work.

It is also interesting to note the correlation between the various turbines. For instance, turbine 1 and turbine 3 are highly correlated in terms of both oil temperature ($r(879) = .85, p < .01$) and planetary gear temperature ($r(879) = .86, p < .01$). This suggests that it may be possible to establish a baseline value for a wind farm, from which individual wind turbines deviate from. This also hints that wake effects are being taken into consideration in the model (as anticipated). It is unlikely (albeit, possible) that two independent wind turbines would be as highly correlated if wake effects were not taken into account. Without topographical knowledge of the wind farm (a priori or otherwise), this analysis is not possible.

As the three independent wind turbines (which were used to empirically determine loading and transient behaviours) did not suffer from gearbox failures over the period of data acquisition, the model has effectively been trained on normal operational behaviour of the wind turbine gearbox. As such, it is promising that the time series presented in figure 4.9 has a strong signature at the time that a catastrophic gearbox failure occurred. As the system had not been exposed to a prior failure, the identification of this failure is promising. This would enable the deployment of this extended physics of failure model into a current wind farm to quantify current gearbox behaviour,

without the necessity of failure histories (as often required in supervised learning approaches, such as neural networks). Given the cost of wind turbine gearbox failure as discussed in section 1.5, this is an immediate benefit of the research undertaken in this thesis.

4.3.10 Conclusion

This section has presented an extended physics of failure model for the quantification of wind turbine gearbox condition based upon empirical data collected from the on board low frequency SCADA system. The model proposed by Feng et al. (2011) was explored, extended and validated on three independent wind turbines. Two novel contributions to this model are made. They are:

1. ***Normalisation:*** The model is extended to take into account both turbine loading and transient states. Similarly, this also normalises against the effect of wind turbine wake due to the paradigm of the model employed.

2. ***Robust performance chart:*** The model is used to design and specify a novel and robust wind turbine performance chart which is valid regardless of wind turbine curtailment and de-rating.

Six time series of the model are presented on three wind turbines, encompassing two model attributes (gearbox oil temperature and planetary gear temperature). No failure data is used to empirically define wind turbine loading conditions and transient states. The model is able to identify the catastrophic gearbox failure which occurred within the data, whilst requiring no additional sensors or capital investment. Normalisation of the data is computationally simple and can be performed in real-time, requiring no additional investment in computational resources.

Given the simplicity of the extended physics of failure model, various limitations inherently exist. First and foremost are the reliance on the SCADA data system. As previously explored, this system has the potential to suffer from significant and substantial downtime. As such, when this occurs, it is not possible to utilise this model. Similarly, the model is only explored with respect to three-blade horizontal axis variable pitch ("danish-concept") wind turbines with a three stage planetary gearbox. As such, the applicability to other types of wind turbines (such as vertical axis, two blade or those with direct-drive gearboxes) may vary.

Furthermore, although a proportion of seasonal variation is removed from the time series, some seasonal variation can still be observed in figures 4.8 through to 4.10. Future work would benefit from exploring this limitation and minimising the observed

seasonal variation in the time series. This would assist in the reduction of noise in the model, helping to provide a more accurate reflection of the true condition of the wind turbine gearbox.

4.4   Chapter summary

This chapter has presented two models to enable the creation of a condition index. The first model utilised robust statistical process control charts (namely, exponentially weighted moving average charts). These were employed in order to quantify the quality of 28 time domain statistical features across 5 distinct categories based upon high frequency bearing accelerometer data. The second extended the physics of failure model of Feng et al. (2011) to robustly define a performance and normalisation chart from which to analyse low frequency wind turbine gearbox SCADA data for the purposes of wind turbine gearbox condition assessment.

With regards to the bearing model, simple features such as the RMS and standard deviation were found to be more sensitive and provide earlier detection of degradation than more commonly explored features such as the median and peak-to-average-power ratio. Measures of central tendency were found to typically be poor indicators of degradation (or potential degradation), with the median, Winsorized mean, tri-mean, mean and mode ranking amongst the least suitable features explored. It is interesting to note that the median absolute deviation (MAD) performed well (ranked 4th of 28), perhaps due to the robust nature of this statistic and the manipulation of extreme values in its calculation. This would assist in explaining why features such as the 10th percentile also performed relatively well. Measures of entropy did not perform as well as anticipated.

The developed extended physics of failure model for wind turbine gearbox condition assessment is shown to be capable of identifying a catastrophic gearbox failure without the assistance of prior failure histories. By being able to utilise normal operational behaviour of a typical wind turbine gearbox, accuracy can be increased (as more data is available encapsulating normal operational behaviour than failure data). The original model was extended to take into account the transient nature of wind. Due to the stochastic nature of wind, this is necessary to reduce the potential for false positive and false negative instances of degradation occurring. It was found that the normalisation against external conditions (ambient temperature) – whilst adequate in this case – was not perfect and could be improved in future work. The issues of SCADA data quality were discussed, with the identification of approximately 19 hours of SCADA data being unavailable per month, per wind turbine.

4.5 Research objective 1 (RO1)

This chapter has set out to explore the development of two models to enable the condition assessment of both bearings and gearboxes based upon high and low frequency data (respectively). As such, this chapter provides insight into research objective 1 (RO1) specified in this thesis in section 3.2.1. RO1 is stated as:

> **RO1: Empirical condition assessment models and feature selection**
>
> *To explore if the use of time-domain features are appropriate as leading indicators of degradation with regards to both bearings and gearboxes for either low or high frequency data, utilising either statistical, data-driven or physical models.*

This is further broken down into two research questions (RQ1, RQ2) which will now explore the results of this analysis to answer these questions.

4.5.1 RQ1: High frequency time-domain features for condition assessment

Research question 1 (RQ1) explores the use of time domain features for condition assessment, and as such, corresponds to the research undertaken in section 4.2 of this thesis. RQ1 is:

> **RQ1:** *Are time-domain features beneficial to assisting in the empirical quantification of bearing degradation based upon high frequency accelerometer data? If so, which features differ significantly and substantially as a result of degradation?* (**RO1**)

In order to answer this question, the four sub-questions proposed in section 3.3.1 must be answered.

Initially, in order to identify relevant time-domain statistical features, a systematic review of the literature was undertaken. In total, 28 statistical time-domain features were identified across five distinct categories. Whilst the list of features identified is not exhaustive, it provides a means to perform the analysis to assist in guiding future research as to which time-domain features may be beneficial for their analysis. The 28 identified features were:

- **Measures of central tendency:** Mean, median, mode, trimean, root mean square (RMS) and the Winsorized mean.

- **Measures of variability:** Standard deviation, variance, interquartile range (IQR), range, minimum, maximum and median absolute deviation (MAD).

- **Measures of shape:** Skewness, kurtosis, hyperskewness, hyperflatness, crest factor (CF) and the peak to average power ratio (PAPR).

- **Measures of position:** 10th percentile, 25th percentile, 75th percentile and the 90th percentile.

- **Measures of impurity:** Shannon entropy, Chao-Shen entropy, Miller-Madow entropy, James-Stein style shrinkage estimator of entropy and a Bayesian estimate of entropy (Krichevsky-Trofimov).

Of these features, some were found to be more suited for the early detection of degradation and condition quantification. For instance, the standard deviation and signal RMS were found to be more suitable than the mode or traditional mean in this analysis. The full ranking of these features can be found in section 4.2.5.

Furthermore, of the five categories of feature explored, some provided more insight than others. For instance, measures of variability were found to be more suitable than measures of central tendency (on average). An informal ranking of the categories of features can be presented as:

1. Measures of variability.

2. Measures of position.

3. Measures of shape.

4. Measures of impurity.

5. Measures of central tendency.

However, within these categories, stronger features in a lesser category may outperform weaker features in a higher category. As such, a selection of strong features across multiple categories is recommended for bearing condition assessment.

Finally, with regards to the final sub-question proposed for RQ1, the features which are able to identify degradation at the earliest onset were the standard deviation, variance and the RMS. As the standard deviation is the square root of the variance, these are highly correlated.

As such, in answering RQ1, various time-domain features are beneficial to assisting in the empirical quantification of bearing degradation based upon high frequency accelerometer data as explored in section 4.2. Furthermore, the degree to which various

features differ significantly and substantially as a result of degradation and potential degradation are explored and discussed in section 4.2.8.

4.5.2 RQ2: Low frequency time-domain techniques for condition assessment

Research question 2 (RQ2) explores the currently available models for wind turbine gearbox condition assessment based upon high and low frequency data. As such, this research question corresponds to the research undertaken in section 4.3 of this thesis. Research question 2 is stated in section 3.3.2 as:

RQ2: Is it possible to normalise a non-stationary signal based upon low frequency gearbox data from the influence of loading conditions, transient states and external factors? If so, can this be empirically derived? (**RO1**)

In order to answer this, the three sub-questions proposed in section 3.3.2 must be answered.

The first sub-question is initially answered through the systematic review undertaken in chapter 2. In this, models based upon high frequency techniques were found to be the most prevalent, however, these often employed frequency-domain or time-frequency domain techniques. Further insight is gained when exploring the literature in depth in section 4.3. In this, recently developed physics of failure models are proposed based upon low frequency SCADA data. Of the four potential models, the chosen physical model provides a means to enable time-domain analysis, whilst exploiting the power of physics of failure models.

Following this, the capabilities of the chosen model were explored. All of the models took into consideration the loading of the gearbox (through various techniques). However, only one model (Kusiak and Verma, 2013) identified their limitations with regards to the curtailment and de-rating of the turbine. External factors were considered by two of the models, however this was implicit through the use of a neural network. The model by Kusiak and Verma (2013) did not take external factors into account as it was unnecessary in this case. The model by Feng et al. (2011) explicitly moderated for external conditions. Finally, of all the models, none of the models took into consideration the transient nature of wind.

Of the four models which were explored, as only one of these moderated for external conditions and gearbox loading, this model (Feng et al., 2013) was selected as the basis for the extended physics of failure model developed in this thesis. Although each of the models were based around the identification of normal behaviour of the wind turbine,

only this model was based upon the physics of failure and could easily be extended to incorporate transient states and both curtailment and de-rating of the turbine. As such, this model was selected.

Finally, although three of the four models identified could be explored within the multivariate domain – all except that of Kusiak and Verma (2013) – two of the three remaining models employed neural networks. Whilst this provide a means to enable prognosis and provide non-linear state estimation, these are not transparent and are difficult to verify and validate. As such, the previously identified model was the only transparent physics of failure model could easily be extended into the multivariate domain.

As such, in answering RQ2, it is possible to normalise a non-stationary signal based upon low frequency gearbox data from the influence of loading conditions, transient states and external factors. This is performed throughout section 4.3 and is done utilising empirical data.

4.5.3   RO1: Empirical asset condition assessment models and feature selection

Having explored both RQ1 and RQ2, it is possible to answer research objective 1 (RO1). This states:

> **RO1: Empirical condition assessment models and feature selection**
>
> *To explore if the use of time-domain features are appropriate as leading indicators of degradation with regards to both bearings and gearboxes for either low or high frequency data, utilising either statistical, data-driven or physical models.*

Having explored time-domain features throughout section 4.2, we can state that many of these features are appropriate leading indicators of degradation. Whilst some features are more suited to condition assessment than others, overall, many of these features were capable of identifying degradation prior to the end of the experiment, and thus provide a means to enable the early detection of bearing faults based upon high frequency data.

Similarly, with regards to the gearbox model, the derived time domain features based upon the multivariate extended physics of failure model is able to be employed as a leading indicator of degradation. Various signatures are observed prior to and during the failure of the wind turbine gearbox as explored in section 4.3.

Chapter 5

Robust Multivariate Techniques

The contents of this chapter appear in part in the following peer-reviewed publications:

- Godwin, J. L. & Matthews, P. C. (2014). *Robust Statistical Methods for Rapid Data Labelling.* In Bhatnatgar, V. Editor, *Data Mining and Analysis in the Engineering Field* (pp. 107–141). Hershey, Pennsylvania (USA): IGI Global.

- Jamie L. Godwin & Peter Matthews. (2014, September). *Accurate empirical encapsulation of asset operational behaviour utilizing elitist memetic algorithms for reliability analysis.* In proceedings of the 27th COMADEM congress. pp.1–8. 16-18 September, Brisbane, Australia. **"Best Paper" award recipient.**

This chapter explores the use of distance metrics as a method for outlier analysis in order to identify and quantify anomalous events which can infer the condition (and thus, degradation) of components. Nine metrics are explored for their capability and sensitivity to the early detection of degradation, for inclusion within a health metric.
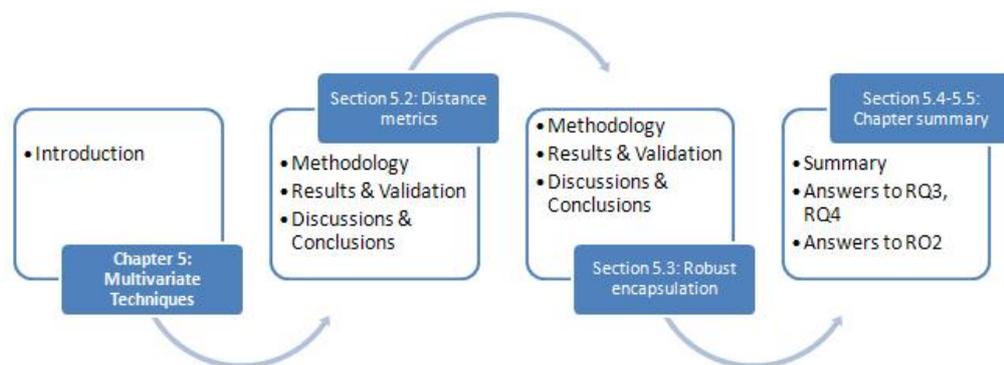
Figure 5.1: Overview of the structure of chapter 5.

5.1   Introduction

Given the extensive array of features which can be explored for the purposes of condition assessment in the time, frequency and time-frequency domain, and also the vast quantity of data often collected from a wide range of sensors; the interactions between these features and the data can be both complicated and complex. As such, substantial benefit can be attained from modelling the interactions of these features and data sources in order to encapsulate the current operational behaviour of an asset into a single univariate feature which is representative of the underlying data. This univariate feature can then be plotted over time, allowing the condition of the component to be observed and also predicted using traditional time series forecasting techniques.

Multivariate distance or similarity metrics provide a means to transform high dimensional data into a univariate feature which can encapsulate the operational behaviour of a component. These are widely utilised in many domains, well researched and often provide a computationally inexpensive means to incorporate high dimensional information into a univariate format. However, whilst these can be utilised within the realm of prognostics and health management, no comparative analysis of distance metrics has been performed previously within the literature. As such, the choice of metric is often subjective and potentially not optimal.

5.2   Outlier analysis as a health metric

Outlier analysis is a natural paradigm in which to explore the notions of failure and asset condition assessment. Failure and degradation – by their nature – are inherently outlying events. If assets were to operate within the limits of normal operational behaviour without failure, there would be no cause for concern. As such, it is deviations from this standard "normal" state which are of interest. The early identification of these deviations and ultimately their prediction is therefore of interest. Thus, techniques and ideas taken from the realm of statistical outlier analysis (in the univariate, bivariate and multivariate domains) can be readily adapted for the assessment of condition of both bearings and gearboxes. This work explores the pre-processing, and processing necessary when utilising multivariate distance and similarity metrics for condition assessment.

5.2.1   Data normalisation

Normalisation of the data is essential in many aspects. Typically, this is done to scale an attribute to ensure that different attributes of different scales can be compared

without bias. Due to the scale variant nature of Minkowski distances, if normalisation is not performed, individual attributes may dominate the derived distance. As such, normalisation ensures that each attribute (feature) contributes equally to the distance.

Normalisation of the data can performed in many ways. Two widely utilised techniques are that of min-max normalisation, also known as "range scaling" (Kohl et al., 2012), and z-score normalisation, also known as autoscaling (Kohl et al., 2012)). Min-max normalisation linearly rescales the data between a predetermined minimum and maximum. This provides definitive bounds for the variable, however, requires a priori knowledge of the data minimum and maximum in order to perform this rescaling. If these values are not known and are estimated, the rescaling has the potential to not be a true representation of the underlying data, as it will not rescale the data correctly. Min-max normalisation is defined by Kumar and Bhandare (2011) as:

$$\hat{\mathbf{x}}_i = \left(\frac{\mathbf{x}_i - \mathbf{X_{min}}}{\mathbf{X}_{max} - \mathbf{X}_{min}} \right) \cdot (\mathbf{K}_{max} - \mathbf{K}_{min}) + \mathbf{K}_{min}) \tag{5.1}$$

Where $\mathbf{x}_i$ represents the vector of current observations, $\mathbf{X}_{min}, \mathbf{X}_{max}$ represent the vectors of minimum and maximum values of the multivariate vector $\mathbf{X}$, and $\mathbf{K}_{min}, \mathbf{K}_{max}$ represent the vector of constant values to bound the new minimum and maximum values by.

Z-score normalisation provides an alternative transform which can similarly be used to rescale data. In this case, rather than a linear re-scaling of the data, data is normalised around zero with a standard deviation of 1. Z-score normalisation is defined by Kumar and Bhandare (2011) as:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mathbf{X}_\mu}{\mathbf{X}_\sigma} \tag{5.2}$$

Where $\mathbf{X}_\mu$ represents the mean vector of the multivariate vector $\mathbf{X}$ and $\mathbf{X}_\sigma$ represents the vector of the standard deviation of $\mathbf{X}$. This normalises the data around zero and is often utilised in artificial neural network sigmoid functions.

Further approaches to normalisation of the data also exist. Due to the nature of these techniques, it would be impossible to present an exhaustive list of techniques. However, for completeness a further two techniques are presented here. These are maximum scaling and profile scaling (Kohl et al., 2012) Maximum scaling is defined by Kohl et al. (2012) as:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\mathbf{X}_{max}} \tag{5.3}$$

Where $\mathbf{x}_i$ represents the vector of current observations and $\mathbf{X}_{max}$ represents the

vector of maximum observed values in the vector $\mathbf{X}$. Profile scaling is defined as per Kohl et al. (2012) as:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\sqrt{\sum_{j=1}^{n}(y_{j,i}^2)}\sqrt{\sum_{j=1}^{D}(\mathbf{x}_i^2)}} \tag{5.4}$$

Where $y_{j,i}$ represents the ith variable of the jth training example. For a review and analysis of the practical implications of these techniques, please refer to Kohl et al. (2012).

Once data has been normalised, it is possible to utilise a variety of distance or similarity metrics to explore various aspects of high dimensional data. These can enable the high level interactions between attributes (features) to be taken into account to determine if the deviation of any feature is of concern, or if it simply natural variation within a process.

5.2.2   Distance metrics

A metric is a non-negative mathematical function which describes the "distance" between two points within a given set. Various conditions must be satisfied by a given function $d(x, y)$ in order to be defined as a metric. These are:

- $d(x, y) > 0$ *(non-negativity)*

- $d(x, y) = d(y, x)$ *(symmetry)*

- $d(x, y) = 0$ *(if and only if $x = y$)*

- $d(x, z) \leq d(x, y) + d(y, z)$ *(triangle inequality)*

Furthermore, a metric can be called scale invariant if the following condition is held:

- $d(x, y) = d(x + z, y + z)$ *(scale invariance)*

For all x, y, z.

Due to the simple definition of a metric, many functions exhibit the require behaviour. Due to this, many metrics exist. As the analysis of all of these metrics is beyond the scope of this work, a small subset are explored here. For a comprehensive listing of all distance metrics which could be employed in this analysis, please refer to the extensive work of Deza and Deza (2009).

Euclidean distance

The Euclidean distance is perhaps the most common of all distance metrics. In practical terms, it represents the "as the crow flies" distance between two points in an n-dimensional space. This is a special case of the Minkowski distance (Deza and Deza, 2009), with the exponent set to 2. As such, the Euclidean distance is defined as per Deza and Deza (2009):

$$D_{Euclidean}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5.5}$$

Where $\mathbf{X}$ and $\mathbf{Y}$ represent two vectors in an $n-$dimensional space. Thus given a point $\mathbf{X}$ (representing the current operational behaviour of the asset, as described by a set of features), and a point $\mathbf{Y}$ (representing empirically defined normal operational behaviour of the asset), it is possible to quantify the current distance between these two points.

Manhattan distance

The Manhattan distance is similar to the notion of Euclidean distance, however, rather describing the distance between two points directly, the Manhattan distance refers to the distance measures between points at right angles. As such, it is named after the grid-like structure of Manhattan island. Similarly to the Euclidean distance, this is likewise a special case of the Minkowski distance, with the exponent set to 1. As such, it is defined as per Deza and Deza (2009) as:

$$D_{Manhattan}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n}|x_i - y_i| \tag{5.6}$$

Chebyshev distance

The final Minkowski distance explored in this chapter is that of the Chebyshev distance. This is the generalised Minkowski distance where the exponent is set to infinity. As such, the Chebyshev distance is defined as per Deza and Deza (2009) as:

$$D_{Chebyshev}(\mathbf{X}, \mathbf{Y}) = \lim_{p \to \infty}(\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}} \tag{5.7}$$

This then represents the absolute maximum value in any dimension between two points. Due to this it is also known as the maximum distance.

Minkowski distances

As specified, the Euclidean, Manhattan and Chebyshev distances are special cases of the Minkowski distance. This is a generalised distance metric which can be applied in the multivariate domain. The Minkowski distance between two n-dimensional vectors, $X = (x_1, x_2 \cdots x_n)$ and $Y = (y_1, y_2 \cdots y_n)$ is defined as per Deza and Deza (2009) as:

$$D_{Minkowski}(\mathbf{X}, \mathbf{Y}, p) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{5.8}$$

For some real-number $p \geq 1$. It can be seen that the Euclidean distance is the Minkowski distance with $p = 2$, the Manhattan distance with $p = 1$ and the Chebyshev distance with $p = \infty$.

Canberra distance

The Canberra distance (Lance and Williams, 1966) is not a member of the Minkowski family. This metric is effectively a weighted adaptation of the Manhattan distance. The Canberra distance is defined as:

$$D_{Canberra}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{5.9}$$

Where $X = (x_1, x_2 \cdots x_n)$ and $Y = (y_1, y_2 \cdots y_n)$.

Cosine distance

The Cosine distance metric further deviates from the Minkowski distances. In this case, the metric provides a similarity rating between 1 (similar) and -1 (not similar). This is defined as per Nguyen and Bai (2011) as:

$$D_{Cosine}(\mathbf{X}, \mathbf{Y}) = \cos(\theta) = \frac{\mathbf{X} \cdot \mathbf{Y}}{||\mathbf{X}|| \, ||\mathbf{Y}||} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}} \tag{5.10}$$

Where $\mathbf{X} \cdot \mathbf{Y}$ represents the dot product of vectors $\mathbf{X}$ and $\mathbf{Y}$, and $||\mathbf{X}||$ represents the magnitude of vector $\mathbf{X}$.

Lance-Williams distance

The Lance-Williams distance (Lance and Williams, 1967) was developed by Lance and Williams in 1967 primarily for the purposes of clustering. By identifying other points

which are not distant, an intuitive cluster of points can be attained. For our purposes however, this distance is not used for clustering, but instead, the quantification of asset behaviour. This is defined as:

$$D_{LW}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)} \tag{5.11}$$

Where n represents the dimensionality of the data, and $|x|$ represents the absolute value of x.

Penrose distance

Given the prior distance metrics, it can be observed that attributes are considered independently of one another. Whilst in many cases it is true that features may be independent, in our case, features cannot be said to be independent (as multiple features are collected from the same bearing). As such, the interactions between variables should be taken into consideration. The Penrose distance extends the prior distance metrics by taking into account the variance of the attributes. By taking attribute variance into consideration, the metric can be made scale invariant. As such, the pre-processing required for the Minkowski, Canberra, Cosine and Lance-Williams metrics explored so far is not required for the Penrose distance.

The Penrose distance is defined as by Penrose (1952) as:

$$D_{Penrose}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \left[\frac{(\mathbf{X} - \mathbf{Y})^2}{n \cdot \mathbf{V}_i} \right] \tag{5.12}$$

Where in this case the vector $\mathbf{Y}$ would typically be defined as $\boldsymbol{\mu}$ representing the $n-$dimensional vector of attribute means and $\mathbf{V}_i$ would represent the variance of attribute i (assumed to be the same across all attributes). By defining $\mathbf{Y}$ as the vector of central points (such as means), deviation from this point in an $n-$dimensional space can determined. As we are interested in normal operational behaviour, it stands to reason that given historical data, the central point will typically represent the desired operational characteristics.

Mahalanobis distance

The Mahalanobis distance (De Maesschalck et al., 2000) further extends the notions presented in the Penrose distance. As the Penrose distance takes attribute variance into consideration when quantifying the distance between two points, the Mahalanobis distance also takes the attribute covariance into consideration when quantifying these distances. This additional information can be utilised to enhance the quality of the

distance metric – and as the behaviour of this metric is well understood – allows the exploitation of various properties of the underlying distribution of the metric value to be exploited for the purposes of early detection, diagnosis and prognosis. The Mahalanobis distance is defined by De Maesschalck et al. (2000) as:

$$D_{Mahalanobis}(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})^T \Sigma^{-1} (\mathbf{X} - \mathbf{Y})} \tag{5.13}$$

Where Σ represents the covariance matrix. Typically, the Mahalanobis distance is written as:

$$D_{Mahalanobis}(\mathbf{X}) = \sqrt{(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})} \tag{5.14}$$

Where $\boldsymbol{\mu}$ represents the $n-$dimensional vector of centres from which we are measuring deviations from. This is as we are often interested in the identification of outlying points (those which are distant from the mean of the data). In the case of condition assessment, we are often interested in quantifying deviations from an anticipated (normal) behaviour which would be set a priori to represent the central vector.

As the Mahalanobis distance takes into consideration both attribute variance and covariance, highly correlated features can be employed without redundancy (unlike the Penrose distance). As such, this metric inherently performs a form of residual analysis by quantifying inter-attribute deviations from empirically anticipated values.

A robust Mahalanobis distance

The final multivariate distance metric explored in this work is that of the robust Mahalanobis distance (RMD). This further builds upon the traditional Mahalanobis distance, however, whilst exploiting robust statistical techniques in order to more accurately define the central vector and also increase metric sensitivity.

The robust Mahalanobis distance is defined in Hardin and Rocke (2005) as:

$$D_{RMD}(\mathbf{X}) = \sqrt{(\mathbf{X} - \boldsymbol{\mu}')^T \Sigma'^{-1} (\mathbf{X} - \boldsymbol{\mu}')} \tag{5.15}$$

Where $\boldsymbol{\mu}'$ and Σ' represent the central vector and covariance as determined by the estimated minimum covariance determinant (MCD) subset (Rousseeuw, 2005). This is a heuristic technique which overcomes the limitations of traditional covariance calculations. As the traditional covariance calculation is highly sensitive to noise (Hardin and Rocke, 2005) and degradation can be thought of as noise, it may potentially be difficult to accurately encapsulate normal operational behaviour using the traditional covariance calculation.

5.2.3 Experimental setup

In order to explore the capability and sensitivity of each identified multivariate metric, three distinct experiments were devised covering the two datasets explored in chapter 4. These are detailed below.

For the purposes of this analysis, data from three bearings were employed from the NASA bearing dataset (Lee et al., 2007). These were:

- **Bearing 1: (Failure history)**: Taken from test 2, bearing 2 of the dataset.

- **Bearing 2: (Suspension history)**: Taken from test 2, bearing 1 of the dataset.

- **Bearing 3: (Suspension history)**: Taken from test 2, bearing 3 of the dataset.

Bearing data

Exploration of metric quality was performed on the high frequency bearing accelerometer data (Lee et al., 2007) by defining three experiments which would assist in determining the most applicable multivariate distance metric. These experiments were:

1. **Experiment 1:** In this experiment, all 3 bearings are employed for the analysis. Of these bearings, 2 do not degrade over the length of the data collection (bearings 2 and 3), whereas the final bearing (bearing 1) did. All 3 bearings are used to determine the multivariate centres. This is done as in practical applications; we may not know the condition of the bearings as data is collected. It is possible that unobservable degradation data is contaminating the (perceived) normal operational behaviour.

2. **Experiment 2:** In this experiment, a bearing (bearing 2) which was known to not degrade substantially over the period to set the multivariate centres of the data. The metrics are then validated on an independent bearing which is known to have degraded (bearing 1) and also one which did not (bearing 3). As failure data is often unavailable in practice, this experiment examines the case when only known normal operational behaviour data is utilised (such as after maintenance, repair or servicing), and explores if it is possible to develop strong techniques to identify artefacts within the data.

3. **Experiment 3:** In this experiment, data from a single bearing is employed for the full analysis; bearing 1. As this bearing failed during the data collection process, the experiment is designed to detail the robustness of each metric. Stronger metrics will be less influenced by the degradation process and as such, should still

provide strong encapsulation of the operational behaviour without requiring additional data – such as normal operational behaviour attained through suspension histories – and without artificially adding noise to the resulting metric.

Due to the multivariate nature of the distance metrics employed, multivariate data must be employed. As such, the multivariate moments model representing the bearing was created based upon the analysis performed in chapter 4. The 4-dimensional model of bearing behaviour was constructed using the moments model (section 4.2) features:

- ***Standard deviation:*** Chosen as a measure of variability due to the strong performance on both datasets employed for the analysis in chapter 4.

- ***Root mean square (RMS):*** Chosen as a measure of central tendency due to the consistently strong performance on both datasets employed for analysis in chapter 4.

- ***Skewness:*** Chosen as a measure of shape due to the ability to quantify the underlying distribution shape in terms of asymmetry.

- ***Kurtosis:*** Chosen as a measure of shape due to the ability to quantify the underlying distribution shape in terms of peakedness.

Gearbox data

Exploration of metric quality was also performed on the low frequency wind turbine SCADA data from the Reliawind consortium (Wilkinson et al., 2010). Three experiments were defined which would assist in determining the most applicable multivariate distance metric for this application. This was done to ensure consistency with the results from the experiments undertaken on the high frequency bearing accelerometer data. As it is possible that certain artefacts exist due to the frequency of data collection, various discrepancies between applications may occur. As such, these experiments were performed in order to minimise this potential whilst validating our findings. These experiments were:

1. ***Experiment 1:*** In this experiment, 3 wind turbines are employed for the analysis. Of these, 1 suffered a catastrophic failure during operation. All 3 wind turbines were used to define the multivariate centres, making this experiment conceptually similar to that as in the NASA bearing experiment 1.

2. ***Experiment 2:*** In this experiment, 1 wind turbine (which did not fail over the period) was employed to set the multivariate centres of the data. The metrics were

then validated on the independent wind turbine which catastrophically failed, as well as the third test turbine. As failure data is often unavailable in practice, this experiment demonstrates the robustness of the approaches under real world conditions of data availability.

3. ***Experiment 3:*** Finally, in the last experiment, only 1 wind turbine is analysed (wind turbine 2). The gearbox of this wind turbine failed during operation. Metric central vectors, variance and covariance should not be influenced by degradation (which will be inherently present in the data), and should simultaneously provide strong encapsulation of the operational behaviour. This experiment looks to explore this.

As the extended physics of failure model developed in section 4.3 utilises both the gearbox oil temperature and planetary gear temperature, a bivariate model can be constructed. These can be employed within the multivariate domain due to the nature of the distance metrics explored. This also provides a naturally extensible framework which can be exploited should other sensors or parameters become available for analysis. As it is anticipated that next generation wind turbines will have over 4 times the number of SCADA channels available, should further gearbox sensors become available, they can easily be incorporated into the developed extended physics of failure model. As such, the extended physics of failure model detailed in section 4.3 based upon a bivariate model of the wind turbine gearbox was constructed utilising:

- Normalised gearbox oil temperature.

- Normalised gearbox planetary gear temperature.

5.2.4 Bearing results

The three proposed experiments to assess the sensitivity and capabilities of each of the nine multivariate metrics were initially performed on the high frequency bearing accelerometer dataset (Lee et al., 2007). As the dataset utilised contains both failure and suspension histories, it is possible to explore many aspects of the multivariate distance metrics for the purposes of condition assessment.

For instance, it is intuitive to explore if the inclusion of additional normal operational behaviour will hinder or improve the capabilities of early detection of a given multivariate distance metric. Similarly, it is also intuitive to explore the alternate case of removing failure data from the analysis. As no failure data is necessary to identify failures within this paradigm, it may be the case that some multivariate distance metrics are more sensitive to this inclusion than others.

Bearing experiment 1

As can be seen in figure 5.2 various artefacts, phenomena and signatures can be observed within the bearing timeseries. Firstly, each of the Minkowski distance metrics are similar. The Chebyshev distance is bounded between 0 and 1, as anticipated (due to the normalisation of the data), which provides a means of quantifying current health as a percentage of the Chebyshev distance. The Euclidean distance and Manhattan distances are bounded higher than the Chebyshev distance as one would expect. All three features show similar artefacts. All three Minkowski metrics are able to quantify degradation on the failed bearing and potential degradation on the healthy bearing. It should be noted that the Chebyshev distance is noisier than both the Euclidean and Manhattan distances. Some healing phenomena (Qiu et al., 2006) can be observed in these time series as discussed in chapter 4. It should also be noted that no benefit is attained with regards to the early detection of degradation due to the application of these multivariate distance metrics; the resultant time series are very closely correlated to the underlying features used in their development. Analysis using Pearsons' product moment correlation coefficient identified strong relationships between the Euclidean distance of the failed bearing and the underlying RMS value ($r(982) = .94$, $p < .01$), the standard deviation ($r(982) = .93$, $p < .01$), and the kurtosis ($r(982) = .83$, $p < .01$). It is interesting to note that a strong negative correlation was found between the Euclidean distance of the failed bearing and the skewness ($r(982) = -.71$, $p < .01$).

The cosine, Canberra and Lance-Williams distances all provide more insight than the Minkowski distances. As figure 5.2 shows, deviation from previous states occurs prior to that of the Minkowski metrics. Whilst initially the cosine distance appears to not correlate with the other metrics, one should be reminded that this is a similarity metric and as such, will decrease due to dis-similarity with the central vector, rather than increase. As such, it is negatively correlated with the other metrics. Whilst the cosine distance is bounded by the metric itself (between -1 and 1), the Canberra distance is not. The Canberra distance in this instance seems more sensitive than the previous distance metrics as it is possible to easily distinguish between both of the bearings immediately. Although the Canberra distance appears slightly noisy for the healthy bearing, a large artefact can be observed at approximately time step 600 which the other metrics did not identify. Similarly, this was also missed by the Lance-Williams distance. Although this artefact was missed, the Lance-Williams metric can be seen to be more sensitive than the Minkowski, cosine and Canberra distances as observed from the magnitude of the change on the failed bearing prior to failure. However, the detection of potential degradation is not as early as with the cosine distance.

With regards to the Penrose distance, substantially more variation can be observed.
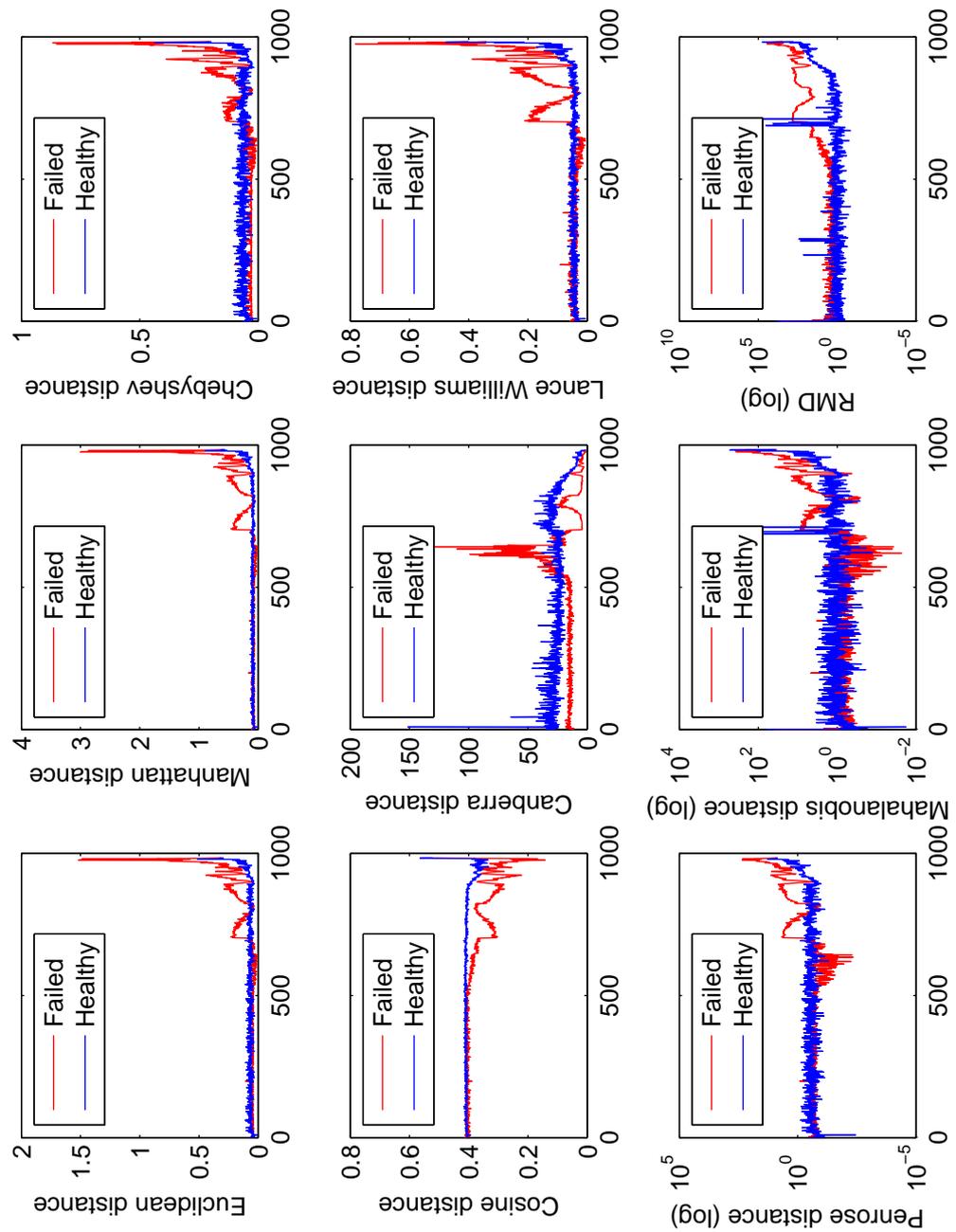
Figure 5.2: Bearing experiment 1 distance metrics.

Due to the large variation between initial conditions and failure with this metric, a logarithmic scale is employed. Deviation from initial conditions are found to be as early as the cosine distance, showing the promise of this metric as a means of early detection of degradation. Furthermore, the artefact identified by the Canberra distance can be observed in this metric as well. The traditional Mahalanobis distance can be identified as being of a similar sensitivity as the Penrose distance to degradation, however, the traditional Mahalanobis distance suffers from substantially more noise than the Penrose distance in this case. Although the traditional Mahalanobis distance is of a similar sensitivity to the Penrose distance, the Penrose distance failed to identify the artefact on the healthy bearing which was identified between time points 600 to 700 by the traditional Mahalanobis distance.

Finally, the robust Mahalanobis distance can be seen to be the most sensitive – varying by multiple orders of magnitude with degradation – whilst containing less noise than both the Penrose distance and the traditional Mahalanobis distance. It is interesting to note that a near linear trend can be observed once the degradation process begins, and continues until failure (the end of the time series). Similar linear trends can be observed on the healthy bearing, which may indicate potential degradation on this bearing as well. It is likely that the linearity of these trends is due to the logarithmic transform applied to the data for display purposes, as this is the natural transform for exponential data.

Bearing experiment 2

As can be seen in figure 5.3, similar patterns and trends can be observed as in figure 5.2. Due to the setup of this experiment, various different signatures can be observed.

Firstly, the Minkowski distances decrease substantially due to degradation. In this paradigm, this would infer that the bearing condition of the failed bearing is increasing (i.e. healing itself) which is not the case. This is observed in all of the Minkowski metrics. Unlike in the previous experiment, in this case, with the assistance of the Manhattan distance, it is possible to separate the bearings from initialisation. This would further suggest that the Manhattan distance is the most sensitive in this context to variation in bearing condition. Some trends continue from the previous experiment, such as the Chebyshev metric being the noisiest of the Minkowski distances. Similar signatures of degradation are observed as in the previous experiment.

The cosine metric provides a clear distinction between the characteristics of each bearing. However, as this metric denotes similarity, the metric should decrease due to degradation. In this case, this does not occur and similarly to the Minkowski metrics, appears to become healthier. This is likewise reflected in the Lance-Williams distance.
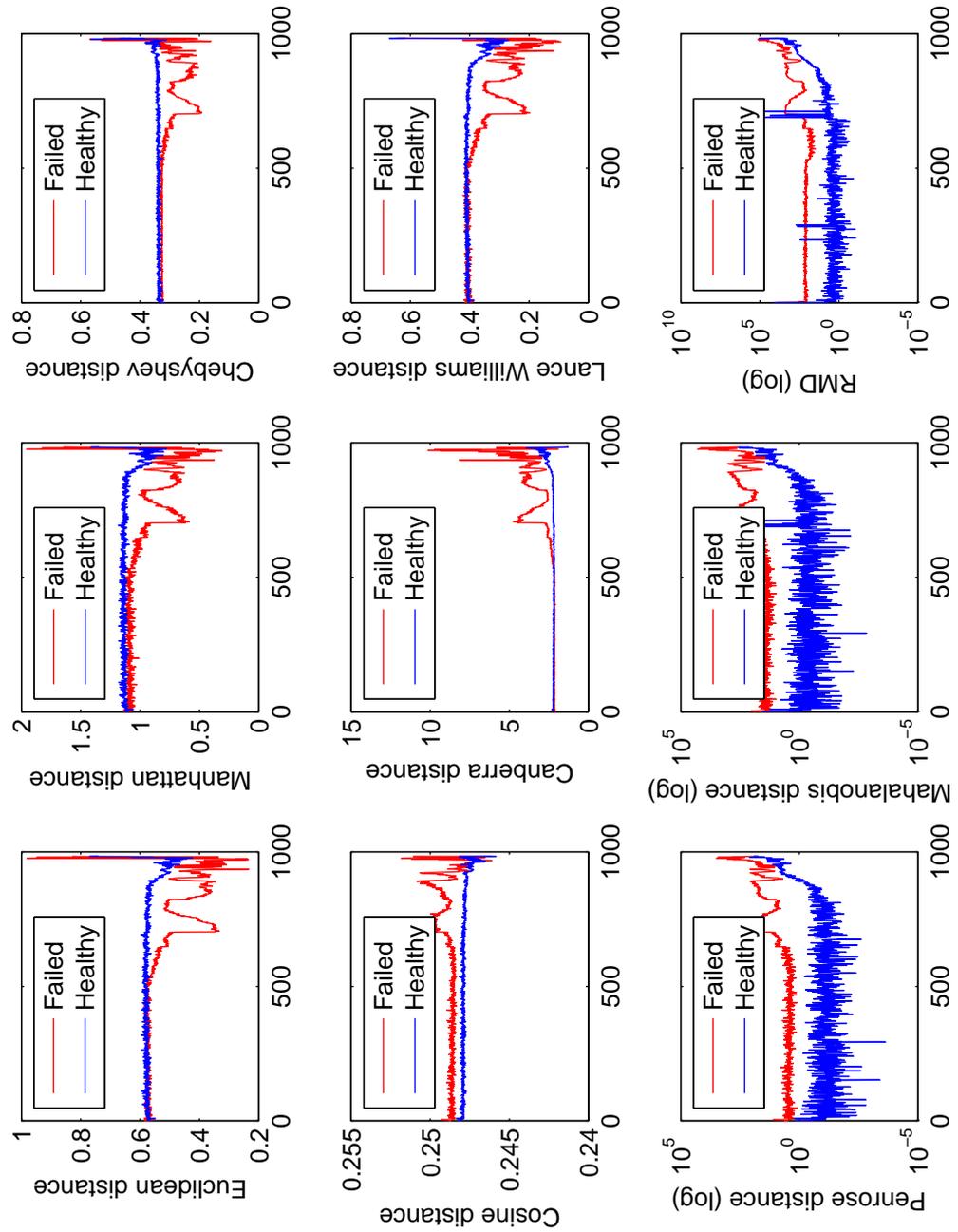
Figure 5.3: Bearing experiment 2 distance metrics.

Whilst in experiment 1 this showed promise due to an increased sensitivity, the sensitivity in this case is comparable to that of the Minkowski distances. The Canberra distance, however, does provide a means to quantify behaviour. This time series is similar to that of the Minkowski distances in figure 5.2; the distances increase as expected, and similar degradation and healing phenomena are observed at the same times as the previous experiment.

Finally, the Penrose, Mahalanobis and robust Mahalanobis distance provide improvement over the other metrics. In each of these cases, both bearings are easily distinguishable, with degradation (and potential degradation) causing an increase in the distance metric. Although the metrics appear noisy, this is primarily due to the logarithmic scale used to present the data. Similarly to the first experiment, the robust Mahalanobis metric has the highest sensitivity (as it is able to identify artefacts within the healthy data which were missed by the other eight metrics). This metric also appears to be less noisy. Due to the sensitivity, the artefacts within the data cause the healthy bearing to appear temporarily degraded. Further analysis reveals these to be outlying data points, potentially due to noise in the data acquisition process. In practice, should a reading like this occur, further data would be acquired in order to verify and validate the change in behaviour. As the values immediately return to anticipated values after these anomalies occur, we can identify these points as noise in this case. This does not occur with the degradation on the failed bearing or potential degradation on the healthy bearing.

Bearing experiment 3

In the final experiment, only the bearing which eventually failed was employed to determine normal operational behaviour for the distance metrics. This is shown in figure 5.4.

Firstly, the Canberra distance in this case did not meet our expectations. Substantial noise is present within the metric (as can be observed in figure 5.4), and degradation causes a decrease in metric value. As such, this metric cannot be deemed suitable for the analysis of bearing degradation. All Minkowski distances perform comparably (as one would expect). Again, the Chebyshev distance contains more noise than either the Manhattan or Euclidean distance. Each of the Minkowski distance metrics increase with degradation, which is promising,

The cosine distance performs better in this experiment than the second, and is comparable to the results attained in the first experiment. As degradation increase, similarity to normal operational behaviour decreases. However, the point at which deviation from prior values occurs provides little to no early detection in comparison
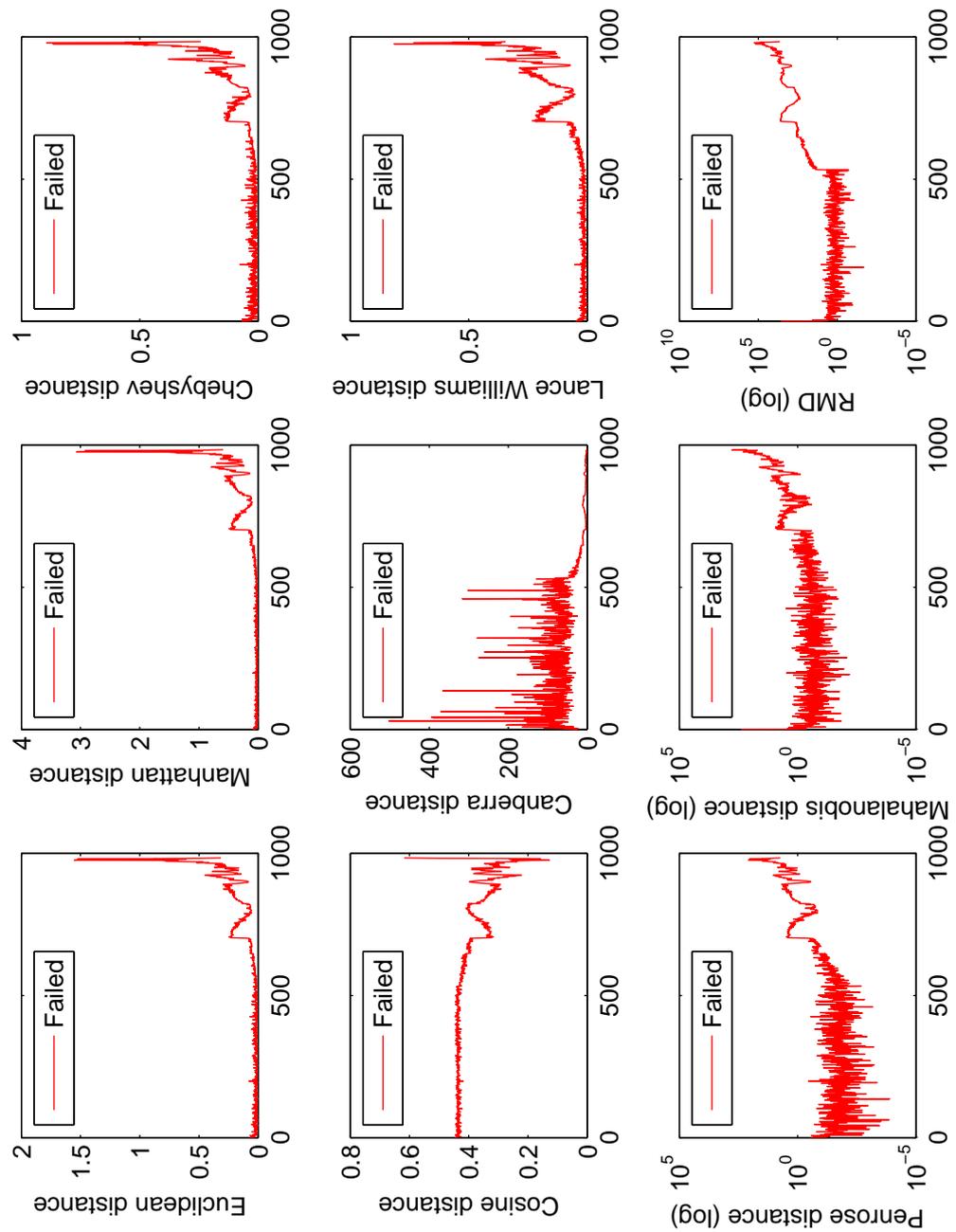
Figure 5.4: Bearing experiment 3 distance metrics.

to other metrics. The Lance-Williams metric is again comparable to the Minkowski distances, however, with increased sensitivity as observed in experiment 1.

Finally, the Penrose, traditional Mahalanobis and robust Mahalanobis distances all out perform the Minkowski, Canberra, cosine and Lance-William metrics. Degradation in all three of these distance functions causes an increase in metric value by orders of magnitude. The Penrose distance contains the most noise, however, the transition between normal operational behaviour and degraded condition is more easily observed in this time series than in that of the traditional Mahalanobis distance. The robust Mahalanobis distance has the least noise, the highest sensitivity and provides the earliest transition from normal operational behaviour to degraded operation.

5.2.5   Gearbox results

The gearbox time series differ substantially from that of the bearing time series in many regards. This is due to many reasons, firstly is the nature of the data itself; the bearing data was accelerometer data, whereas the gearbox data represents temperatures. Secondly, the bearing data is high frequency data (over 20,00 samples per second) which has been sampled to 1 sample per second with the assistance of statistical time domain features (the moments model). In contrast, the wind turbine gearbox SCADA data is sampled once every 10 minutes. Thirdly, the bearing data is collected from a stationary component. Whilst the gearbox model utilises low frequency data, the operational environment (due to its non-stationary operation) is drastically different. Finally, the nature of the model employed effects the time series dramatically. As the gearbox of a wind turbine is subjected to seasonal variation (which we have previously shown in section 4.3 of this thesis that the normalisation does not fully remove), this may be apparent in the time series.

Due to the sample rate and the quantity of data collected from the wind turbines, the multivariate metrics were averaged using the traditional mean on a day-by-day basis in order to quantify the behaviour and condition of a gearbox on a given day. This enables the condition to be tracked over time, whilst assisting in mitigating the influence of noise and anomalous readings from the SCADA system.

Gearbox experiment 1

As seen in figure 5.5, various metrics are better suited for the analysis of wind turbine gearbox condition assessment than others. In this case, it can be seen that both the Euclidean and Manhattan distances are not fit for purpose. With regards to the failed wind turbine which suffered a catastrophic failure between data points 300 and 400;
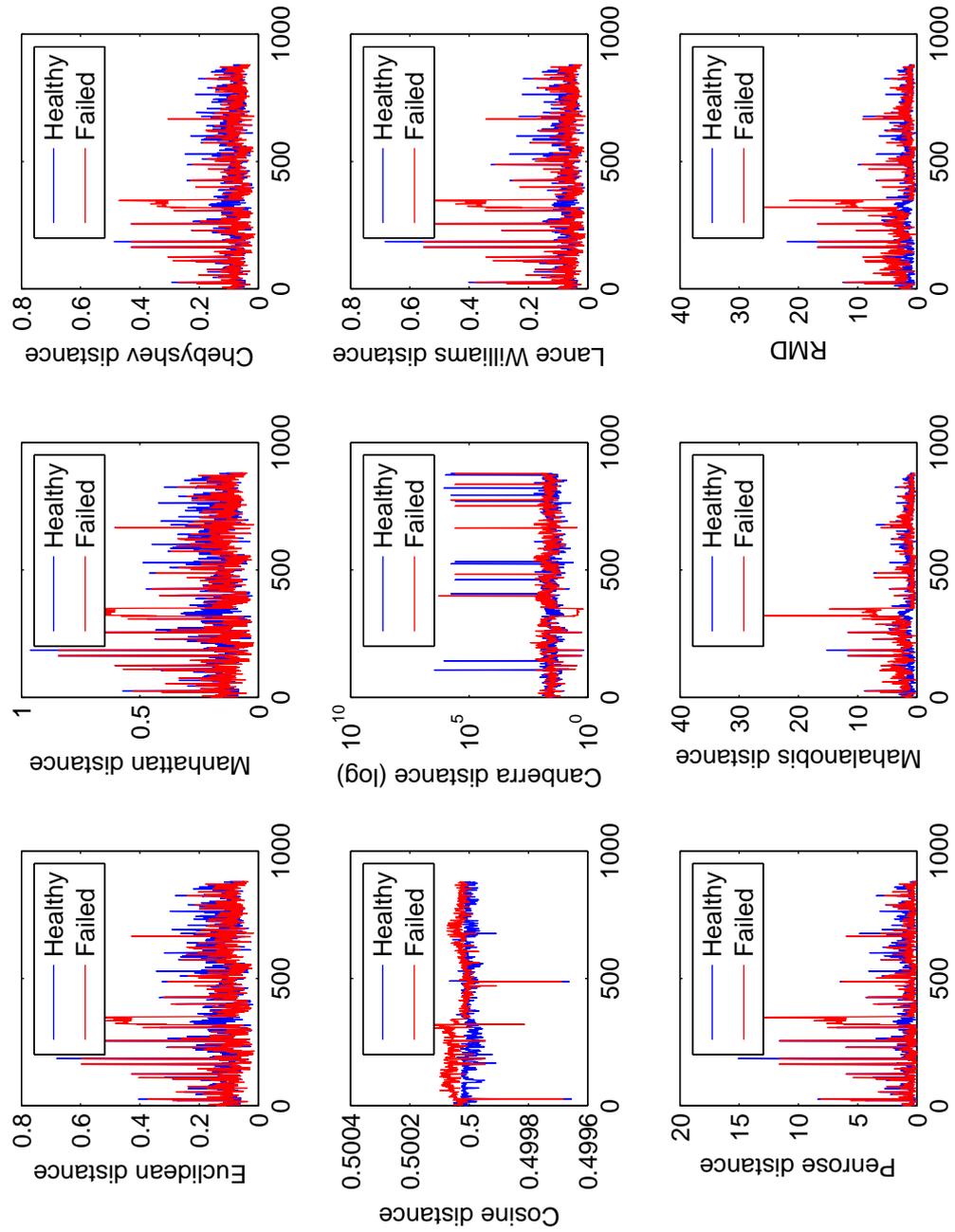
Figure 5.5: Gearbox experiment 1 distance metrics.

although this failure can clearly be observed in both time series, the healthy turbine is given a large distance value at approximately data point 200. This would represent a false positive classification; where degradation does not exist, but the technique is stating that degradation has occurred. In practice, this would warrant an inspection of the gearbox; resulting in lost production and increased maintenance expenditure. This is true for the final Minkowski distance (Chebyshev), however, due to the use of the maximum difference, this is mitigated slightly.

The cosine distance performs poorly. Although deviation of the failed turbine occurs at the point of failure, the signature is extremely weak compared to the Minkowski distances. Similarly, the range of this metric is limited. As such, the use of this metric does not seem justified in this case. The Canberra distance (presented on a log scale here) also does not perform as expected. Although the failure can be identified in the metric, the signature of the failure is weak and is due to a reduction in distance value. It should be noted that many of the large anomalous values in the time series correlate to the missing data as discussed in section 4.3.9 as can be observed in figure 4.11. The Lance-Williams metric is again similar to the Minkowski distances, and offers little else in terms of analysis.

The Penrose distance provides greater sensitivity and a wider range than the other metrics discussed. The artefact of the healthy turbine prior to failure is still present, however, it is possible that this is due to the SCADA system itself. The traditional Mahalanobis distance further improves upon the Penrose distance. In this metric, the catastrophic failure is easily isolated, and the artefact present on the healthy turbine no longer appears more substantial. Finally, the robust Mahalanobis distance (RMD) appears to be more suitable than the prior eight metrics. Higher sensitivity is achieved, the catastrophic failure is easily isolated, the artefact on the healthy turbine is again less substantial than the catastrophic failure and more separation is achieved between "normal" operational behaviour and both the noise and failure in the dataset.

Gearbox experiment 2

In the second experiment, similar results are obtained as in the first experiment. As can be seen in figure 5.6, both the Euclidean and Manhattan distances seem well suited but suffer from the artefact which is present in the healthy turbine time series (leading to a false positive identification of degradation on the gearbox), this is again limited in the Chebyshev distance. Due to utilising the maximum value in the Chebyshev distance, the catastrophic failure which occurred on the failed turbine appears not as substantial as it does in the other Minkowski distances.

Perhaps the most surprising result is that of the Canberra distance. In this case,
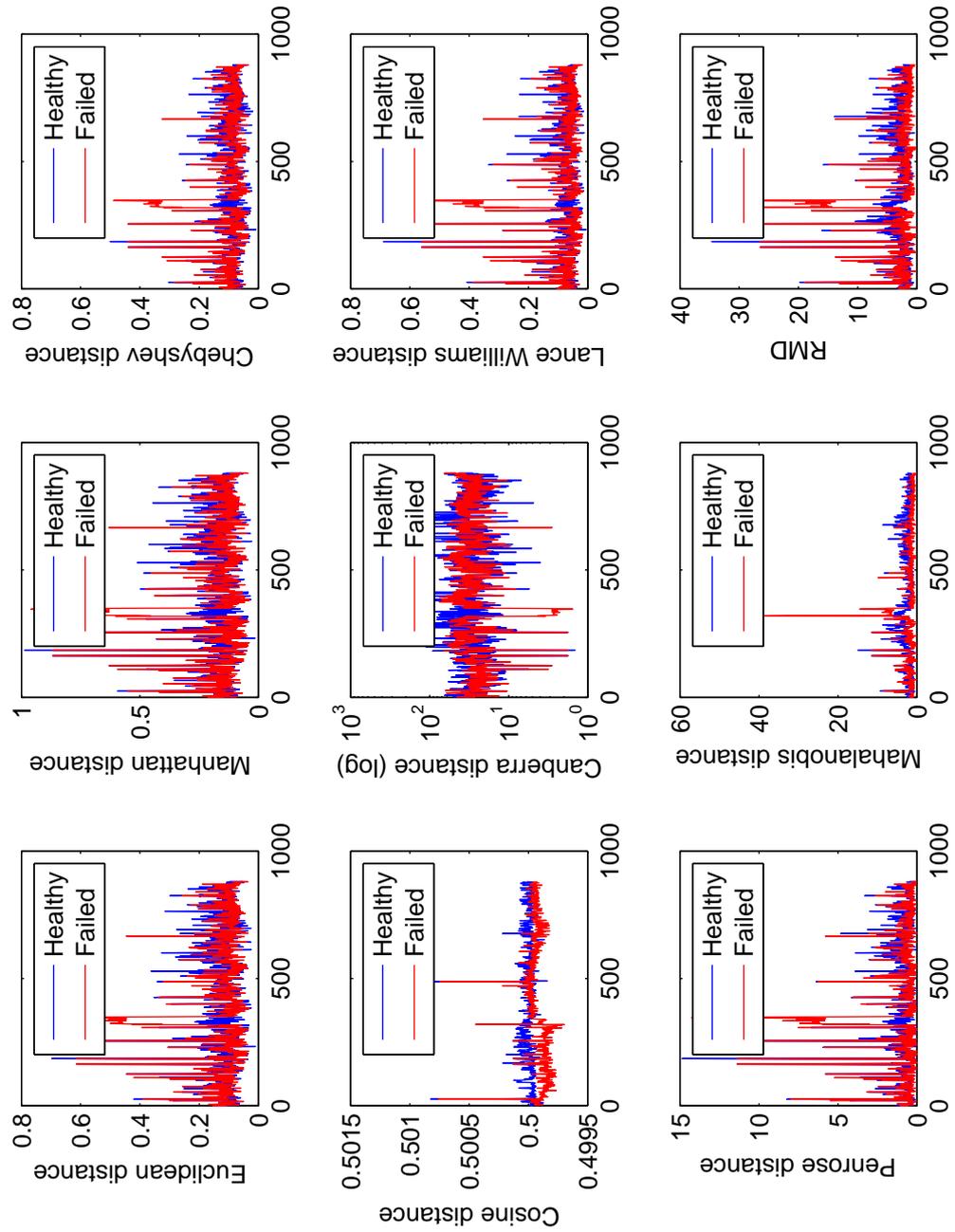
Figure 5.6: Gearbox experiment 2 distance metrics.

the metric does not correctly identify the catastrophic failure which occurred. The healthy turbine in actuality appears more damaged than the turbine which failed. This is unacceptable as a false negative classification would incur substantially more losses (and maintenance expenditure) than a false positive classification.

Similarly, the cosine distance performs as poorly in this case as it did in the first experiment. Although a larger shift can be observed around the time of failure, the separation between the healthy and failed turbines is minimal, and there is little to no trend or pattern regarding the failure. Immediately after the failure the two time series are overlapping; one would anticipate that this means both turbines are equally healthy (or damaged). However, as can be observed in the other metrics, this is not the case. Again, the Lance-Williams metric performs comparably to the Minkowski distances.

The Penrose distance performs similarly to in the first experiment, with a strong signature occurring when the gearbox fails. This likewise suffers from the same false-positive artefact in the healthy gearbox as the other metrics. The traditional Maha-lanobis distance however, does not suffer from this artefact, and is more sensitive to the gearbox failure. Similarly, the robust Mahalanobis distance is similarly sensitive and whilst the artefact in the healthy time series is somewhat substantial, more artefacts can be observed in the time series of the failed gearbox prior to failure, potentially assisting in the early identification of damage.

Gearbox experiment 3

In this experiment, as only the failure gearbox data was utilised to determine the multivariate metric centres, only a single time series is presented here. This experiment explores how sensitive the metric is to the inclusion of degraded data. Similar results are achieved as in the prior two experiments, as can be seen in figure 5.7. Each of the Minkowski distances perform similarly, with the Chebyshev distance containing a little more noise.

Both the cosine and Canberra distances under-perform and do not meet expecta-tions: no failure can be ascertained in the cosine distance, and the failure again causes a reduction in the Canberra distance. Although the failure can be seen in the Canberra distance, given the large quantity of noise in the metric, it is difficult to distinguish. The Lance-Williams metric is similar to the Minkowski distances, as in previous exper-iments. Of the final three metrics, the robust Mahalanobis metric once again performs better than its peers. Higher sensitivity is achieved in comparison to both the Penrose distance and the traditional Mahalanobis distance. The sensitivity in this case is an artefact of estimating the minimum covariance determinant (MCD) subset. Due to the large quantity of degraded data present, and the accurate encapsulation of behaviour as
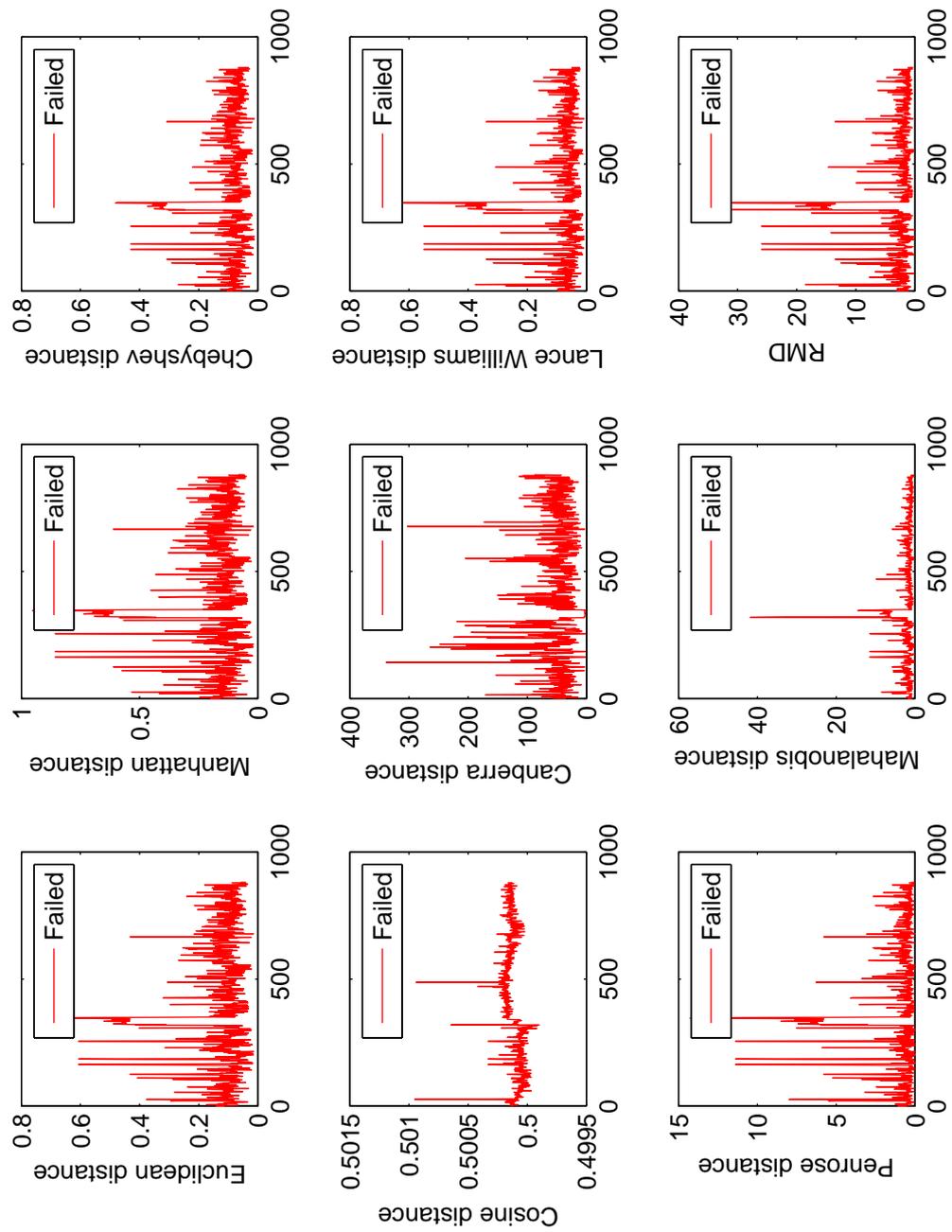
Figure 5.7: Gearbox experiment 3 distance metrics.

determined by the minimum covariance determinant algorithm, slight deviations from this are given large distances. As the traditional covariance calculation is highly sensitive to noise, many of these 'peaks' in the robust Mahalanobis distance are suppressed as they are deemed to be within normal limits (which is not always the case).

5.2.6   Discussion

As explored in sections 5.2.4 and 5.2.5, different multivariate distance metrics were anticipated to perform better in different cases. This was not the case. As can be seen repeatedly, the robust Mahalanobis distance (RMD) performs consistently better than any of the other multivariate metrics explored in this work.

This is due to the encapsulation of normal behaviour with the assistance of an estimated minimum covariance determinant (MCD) subset (Rousseeuw and Driessen, 1999). Although this technique is computationally more expensive, it has been utilised to create a metric which is more sensitive to both bearing and gearbox degradation.

The Minkowski distances all perform similarly in this case. It is anticipated that in models containing higher dimensionality, larger differences would be seen between these metrics. However for small dimensional models – such as the bearing moments model (4) and extended physics of failure wind turbine SCADA gearbox model (2) – these differences in this case are minimal.

The cosine distance performed reasonably well on the bearing dataset, however, performed poorly for the analysis of the gearbox. As such, it may be the case that this metric in particular is suited to particular applications. As this is is a similarity metric, rather than a distance metric, this assist in explaining the poor result attained for the gearbox data. Similarly, the Canberra distance did not perform well on the gearbox data. This is interesting as the Minkowski distances, Penrose and both the traditional and robust Mahalanobis distances performed in comparable manners across both datasets. Due to the behaviour of the cosine distance in the second bearing experiment, it was felt that this metric should be avoided. The Lance-Williams metric appeared to be very closely correlated to the Minkowski distances; this metric performed better on the bearing data than the gearbox data. This is perhaps due to the higher dimensionality of the bearing moments model than the extended physics of failure gearbox model.

The Penrose distance typically performs well, however, it should be pointed out that highly correlated variables will influence the metric in a potentially biased manner. For instance, in an 8 dimensional model, if 7 of these attributes are highly correlated with each other, these will dominate a substantial contribution to the distance. If the values they are measuring are not directly associated with degradation, this may lead to

potentially false-positive readings as there are only (in reality) 2 unique attributes in the model, with 1 attribute accounting for seven eighths (87.5%) of the contribution to the distance metric. As the Mahalanobis distances takes these interactions into consideration, highly correlated variables can be employed within this metric. This also holds true for the robust Mahalanobis distance.

Overall, the metrics can be summarised as follows:

- **_Minkowski distances:_** Typically these metrics performed better than the cosine and Canberra metrics, and should be explored initially to ascertain if distance metrics are a viable technique to be employed on the dataset in question.

- **_Cosine distance:_** Due to the spurious results attained in the second bearing experiment, it is recommended that this metric is avoided for bearing and gearbox health assessment.

- **_Canberra distance:_** As this metric failed to identify the gearbox failure in the second gearbox experiment, it is recommended that this metric be explored in detail prior to its use in practice, or avoided entirely.

- **_Lance-Williams distance:_** Typically, this metric performed similarly to the Minkowski metrics. As such, it is recommended to utilise the Minkowski metrics first before exploring the Lance-Williams metric.

- **_Penrose distance:_** Although this performs well in both cases, as high correlated variables have the potential to influence the metric, it is recommended that this metric is not employed when this is the case.

- **_Traditional Mahalanobis distance:_** This out performs all of the previous metrics, however, has the potential to lack sensitivity due to the calculation of the traditional covariance matrix employed.

- **_Robust Mahalanobis distance:_** This is the best performing metric explored in both datasets, and has higher sensitivity than the traditional Mahalanobis distance due to the use of the minimum covariance determinant subset (Rousseeuw and Driessen, 1999) employed for covariance calculation.

5.2.7 Conclusion

This section has explored the use of 8 multivariate distance metrics and 1 multivariate similarity metric in 3 independent experiments across 2 datasets. The robust Mahalanobis distance (RMD) was found to consistently outperform the other metrics explored in this work, and to be applicable on both the bearing and gearbox datasets.

This metric was found to be the most sensitive to degradation, and provide a means of early detection due to this sensitivity. Whilst the other candidate distance and similarity metrics explored in this work are viable for use in a PHM solution, the quality of the solution would be reduced if alternative distance metrics were employed.

With regards to the high frequency bearing accelerometer dataset, the three experiments show that the Minkowski distances all performed comparably. Similar trends and patterns are observed amongst each of these three metrics. Both the cosine and Lance-Williams metrics also performed similarly to the Minkowski distances. The Canberra distance metric differs from these, however, its use as a metric for condition assessment is limited. Both the Penrose distance and traditional Mahalanobis distance are well suited, with the traditional Mahalanobis distance being slightly more suitable for condition assessment due to both its sensitivity and also inherently taking attribute covariance into consideration.

With regards to the low frequency wind turbine SCADA gearbox dataset, similar results are found. This suggests that the data acquisition frequency may be of limited concern when employing these techniques. The Minkowski distances are noisy and provide minimal insight into the underlying condition of the gearbox. This trend continues with the Lance-William metric. In this case, the cosine distance differs substantially from the previous metrics, but is also limited in it's application. The Canberra distance is an improvement over this, however, the failure causes an unexpected reduction in the metric value, again limiting the value of the metric in this case. The Penrose and traditional Mahalanobis distances both prove suitable in this case, however, the traditional Mahalanobis distance is slightly more sensitive.

The robust Mahalanobis distance (RMD) metric out performs the other metrics in this case due to the selection of a strong subset of data from which to base the estimates of centres (location) and covariance (scatter). This is due to utilising techniques to assist in identifying the minimum covariance determinant (MCD) subset. In this work, the Fast-MCD algorithm was employed (Rousseeuw and Driessen, 1999). This is a heuristic technique and is not as accurate as exact methods. It is anticipated that stronger encapsulations of this subset will further increase the sensitivity of the robust Mahalanobis distance, it is of interest to explore the techniques which can identify or estimate the minimum covariance determinant subset.

As such, this thesis will now explore the influence of the effect of the minimum covariance determinant (MCD) subset on the resulting robust Mahalanobis distance (RMD) for the purposes of condition assessment.

5.3   Robust estimation of the covariance matrix

5.3.1   Introduction

As hinted upon in section 5.2.6, the calculation of the covariance metric strongly effects the calculation of the Mahalanobis distance. This is due to the sensitivity of the traditional covariance calculation to both outliers and noise. This section of the thesis explores the robust calculation of the covariance matrix for use in the Mahalanobis distance. By limiting the influence of outlying data, data which is regarded as noise and also degradation data, a more accurate representation of the true condition of an asset can be achieved.

This chapter explores the calculation of the covariance matrix with respect to the Mahalanobis distance. The traditional covariance calculation is presented, and shown to be inferior in comparison to the minimum covariance determinant (MCD) subset for the purposes of condition assessment. Following this, exact algorithms for the calculation of the minimum covariance determinant (MCD) subset are explored. Heuristic techniques are then presented, followed by the development of an elitist memetic algorithm (EMA) in order to overcome the limitations of both the exact and heuristic methods.

5.3.2   Traditional covariance estimation

The covariance matrix encapsulates both the variance of attributes and also the co-variance of attributes in the multivariate domain. Typically, the covariance matrix is denoted Σ although it is often referred to in the literature as S. In this matrix, the variance of the attributes populates the diagonals of the matrix, with the covariances between attributes populating the off-diagonal positions. The notion of covariance is closely linked to that or correlation, with correlation between two variables being the normalised covariance.

Consider a matrix X, with rows of this matrix denoted $\mathbf{X_i}$. The covariance matrix is defined as:

$$\Sigma = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})^T \tag{5.16}$$

Where $\bar{\mathbf{X}}$ traditionally represents the sample mean. As in practice the majority of data collected from assets will pertain to normal (or near-normal) operational behaviour, the covariance matrix encapsulates the interactions between the model variables based upon their typical behaviour. If one was to utilise degraded behaviour to derive the covariance matrix, it would be possible to analyse discrepancies between the

covariance matrix representing normal operational behaviour and also the covariance matrix representing degraded behaviour in order to gain insight into the behaviour of the degradation in this case.

However, this calculation of the covariance matrix is heavily influenced by noise. In practice, this means that if traditional covariance calculations are used when employing the Mahalanobis distance for the purposes of health assessment, the sensitivity of the metric will be reduced. This is due to the degradation, erroneous measurements and also noise contaminating the dataset. If data containing degradation is utilised to derive the covariance matrix, it would be difficult for the resulting distance metric to be able to identify future cases of degradation at the level which was present within the data.

This phenomena is known as masking and swamping, and defined by Hardin and Rocke (2005) as:

- **Masking:** Masking occurs when a group of outlying points (degraded asset behaviour) skews the mean and covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

- **Swamping:** Swamping occurs when a group of outlying points (degraded asset behaviour) skews the mean and covariance estimates toward it and away from other inlying points and the resulting distance from the inlying points to the mean is large.

Thus, masking causes degradation to be "hidden" as it is given a smaller Mahalanobis distance; effectively causing degradation to be classed as normal operational behaviour (false negatives). Similarly, swamping causes normal operational behaviour to be given a higher Mahalanobis distance as it appears more degraded than it actually is (false positives). As such, it is essential that the data is cleansed and processed in such a way to minimise these phenomena and ensure the covariance matrix truly encapsulates the behaviour of the component.

Within the literature various techniques to robustly estimate the covariance matrix exist. These range from exact techniques to heuristics employed in order to ease the computational complexity of the calculation.

5.3.3   Robust estimation

Robust estimation of the covariance matrix aims to identify outlying points prior to the calculation of the covariance matrix in order to determine a more accurate representation of the interaction between model attributes. Within the realm of univariate data, manual inspection of box-plots can be performed to identify outlying points. In

the case of bivariate data, scatter plots can be inspected to perform this identification. However, in the multivariate domain, it becomes impractical to inspect all possible combinations of attribute scatter plots. As stated in Robson (2003), an outlier containing mild but systematic errors in all of its components (dimensions) will remain hidden from manual inspection of the scatter plots.

As is demonstrated by Rousseeuw and Driessen (1999) and also shown in figure 5.8, traditional covariance estimation cannot accurately encapsulate behaviour even in bivariate cases. In their work, the Fast-MCD algorithm (Rousseeuw and Driessen, 1999) is utilised to robustly estimate the covariance matrix.

For the purposes of robust estimation one assumes a certain quantity of contamination of the dataset. This enables the robust calculation of the covariance matrix to be modelled as a subset selection problem (Robson, 2003). The problem then becomes one of identifying the most suitable h of N points which encapsulates the true behaviour of the multivariate dataset. In practice this is a fair assumption: techniques such as the minimum covariance determinant (MCD) have a high breakdown point (51% with regards to covariance estimation) due to their robust nature. Thus, as long as the majority of the data ($\geq 51\%$) pertains to "normal" operational behaviour, limiting our subset of points to some h of N will not disregard potentially useful information.

In order to quantify the quality of a given subset of points, a metric is required. The MCD technique employs the determinant of the inverse of the covariance matrix. This can effectively be thought of as being proportional to the volume of the m dimensional ellipse covered by the covariance matrix. Thus, the subset of points which minimise the
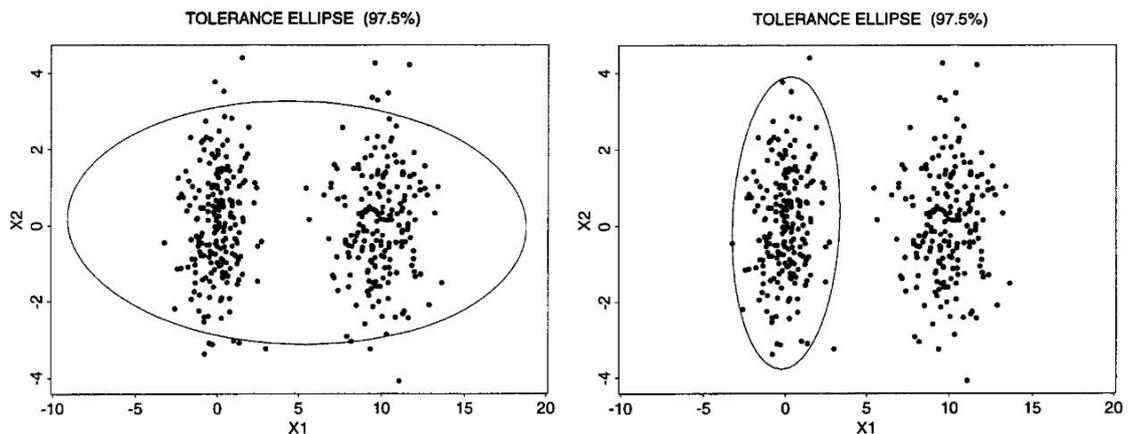
(a) Traditional covariance calculation (bivariate) tolerance ellipse (97.5%) (Rousseeuw and Driessen, 1999).

(b) Robust covariance calculation (bivariate) tolerance ellipse (97.5%) (Rousseeuw and Driessen, 1999).

Figure 5.8: Estimation of the covariance matrix.

determinant of the inverse of the covariance matrix given h of N points most accurately encapsulates the behaviour of the data. The minimum covariance determinant can be formulated as per Hardin and Rocke (2005):

$$
\begin{aligned}
MCD &= (\bar{X}_j^*, S_j^*) \\
j &= \text{set of } h \text{ points } |S_j^*| < |S_k^*| \\
\forall \text{ sets } k \text{ s.t. } &\#|K| = h \\
\bar{X}_j^* &= \frac{1}{h} \sum_{i \in j} x_i \\
S_j^* &= \frac{1}{h} \sum_{i \in j} (x_i - \bar{X}_J^*)(x_i - \bar{X}_J^*)^T
\end{aligned}
\tag{5.17}
$$

The MCD technique is superior to other robust estimators such as the minimum volume ellipsoid (MVE) due to many reasons, for instance, as suggested by Rousseeuw and Driessen (1999):

- A higher breakdown point.

- Higher statistical efficiency.

- Higher accuracy.

- Robust distances based upon the MCD are more precise than those based upon the MVE.

- The MCD is commonly used in linear discriminant analysis, and so is more widely recognised.

As such, the use of the MCD subset is preferred over other methods (such as the MVE) and is utilised in this thesis. Although the Fast-MCD algorithm (Rousseeuw and Driessen, 1999) has been mentioned previously in this thesis, this is an estimate of the true MCD subset that is calculated based upon a heuristic. It is possible to determine the true MCD subset through the use of exact algorithms which are sound and complete, and provide a means to establish a better encapsulation of typical asset or component behaviour than estimates of the MCD subset can provide.

Exact algorithms

Given the requirement to identify a subset of points in order to determine normal operational behaviour, various exact and heuristic techniques exist. Exact algorithms

provide a comprehensive search of the search space in order to effectively evaluate the quality of every possible MCD subset, ensuring that the minimum is found. Four algorithms are presented here, as explored in the work of Robson (2003).

Branch and bound

The branch and bound algorithm is an exact algorithm presented by Robson (2003) for the calculation of the true MCD subset. In this algorithm, a subset tree is constructed of depth h. Every possible subset is constructed and evaluated to determine the determinant of the inverse covariance matrix. This is an exponential time algorithm which has $\binom{h}{N}$ combinations. Given the substantial number of subsets to be evaluated (even on smaller samples), it is clear that this algorithm is not feasible in practice.

For instance, given a dataset of size $N = 1000$, $h = 750$, there are over 4.82×10^{242} combinations. This cannot be computed in real-time (or at all in practice). In order to reduce the number of subset evaluations performed, pruning of the subset tree is performed. Pruning is performed once it is known that a branch of the subset tree cannot contain the MCD subset. Although this reduces the quantity of points requiring evaluation substantially (in the case of Robson (2003), by over 99.99%), the number of remaining points is still too large to evaluated effectively in practice.

Sweepline algorithm

The sweepline algorithm (Pesch, 2000) achieves improvements upon the traditional branch and bound algorithm by utilising the notion that the MCD subset is convex complete. That is, the convex hull formed by the MCD subset must only contain points which are part of this subset. In the work of both Pesch (2000) and Robson (2003), exploration of this algorithm was only undertaken for bivariate data. Although $m-$dimensional convex hulls can be computed, the computational intensity of the calculation increases with the dimensionality.

In their analysis, Robson (2003) found that of the potential MCD subsets, only 1.5 in 1 billion (on average) were convex complete. In practice, this was found to equate to an improvement over the branch and bound algorithm by two orders of magnitude.

Extreme algorithm

The final algorithm presented by Robson (2003) is referred to as the Extreme algorithm. This extends the sweepline algorithm by focusing on the points which make up the convex hull. As the subset of points in this hull by definition is smaller than the convex

complete subset, identification of the points which would construct this convex hull would enable the identification of the MCD subset.

In the analysis of this technique presented in Robson (2003), this technique was found to outperform the sweepline algorithm slightly. It is interesting to note that this technique is computationally more effective when the subset size is small ($h = \frac{1}{4} \cdot n$). Again, as this exploration was performed in the bivariate case, the influence of the dimensionality of the data was not taken into consideration. Given the complexity of the current algorithm, one can assume that complexity will increase with the dimensionality of the data.

Bernholt and Fischers algorithm

Bernholt and Fischers algotithm (Bernholt and Fischer, 2004) further refines the concepts identified in the extreme algorithm of Pesch (2000). In this case, a more specific condition is employed in order to reduce the search space and assist in determining the MCD subset.

The condition exploited is that the MCD subset must be selectable by an ellipsoid (Bernholt and Fischer, 2004). This is summarised by Robson (2003) as follows: as the ellipsoid is convex, the subset is convex complete ... as the convex hull is the boundary of the smallest convex region containing these points, the convex hull lies within the ellipsoid. Furthermore, Robson (2003) goes on to state that the reverse is not true and that a convex complete subset may not necessarily be selectable by an ellipsoid. Thus, they suggest that algorithms utilising ellipsoids may be faster than those utilising convex hulls.

In the bivariate exploration of the algorithm in Robson (2003), the redundancy of the solutions presented by this algorithm is noted. In their case, five points were utilised to define a quadratic to define the scope of the ellipsoid in the bivariate domain. It was found that solving the derived linear system often resulted in many redundant subsets (as the ellipsoid often covered the same points). The authors of Robson (2003) remedy this with the implementation of a binary search tree to ensure that subsets are not evaluated redundantly. However, the added computational complexity of this negates the benefit achieved through the algorithm. The algorithm is shown to be an improvement over the previous techniques. However the running time – as per the former algorithms – is again not practical with a running time of $\mathcal{O}(N^{d(d+3)/2})$ (Bernholt and Fischer, 2004) (where d represents the dimensionality of the data).

Heuristic algorithms

Identification of the MCD subset has been shown to be $NP-$complete by reduction to a maximal clique problem by Bernholt and Fischer (2004). Given the difficulty of exact calculation of the MCD subset, various heuristics and approximation techniques are employed in practice to assist in a suitable subset of points. As any improvement over utilising the traditional covariance matrix will translate to more practical distance metrics, the exact MCD subset is not required. Rather, any subset which can be shown to approximate the true MCD subset can be employed.

Fast-MCD

The Fast-MCD algorithm (Rousseeuw and Driessen, 1999) is the most common of all MCD subset algorithms. This is a heuristic technique which iteratively refines the selected subset until a robust estimate is found. The algorithm is defined in two parts, one for $N < 600$ and one for $N \geq 600$. The first is defined as follows (Rousseeuw and Driessen, 1999):

1. Define $h = .75 \cdot N$

2. Generate x candidate solutions:

   (a) Create a sample H_1 either randomly of size h, or of $p+1$ (iteratively adding random points if the covariance matrix is singular)

   (b) Carry out two "C-steps"

3. For the 10 of x candidate solutions with the lowest determinants:

   (a) Carry out "C-steps" until convergence

4. Report the solution with the lowest determinant

In the case that $N \geq 600$, the algorithm is defined as (Rousseeuw and Driessen, 1999)

1. Construct up to five disjoint random subsets of size n_{sub}

2. For each subset, repeat x times:

   (a) Construct an initial sample $h_{sub} = .75 \cdot n_{sub}$

   (b) Carry out two "C-steps"

   (c) Keep the 10 of x solutions with the lowest determinant

3. Merge the disjoint subsets

4. Repeat on the merged subset y times:

   (a) Carry out two "C-steps"

   (b) Keep the best 10 of y results

5. In the full dataset, for each result:

   (a) Perform "C-steps" until convergence

   (b) Keep the subset with the lowest determinant

6. Return the best identified subset

There are a few interesting notions to discuss here. Firstly, is that of candidate initialisation, and secondly, that of a "C-step".

With regards to candidate initialisation, if minimal contamination is within the dataset, it make no difference if the h points are drawn randomly (Rousseeuw and Driessen, 1999). If the dataset is highly contaminated, $p + 1$ datapoints are selected, with a "C-step" performed to identify the remaining points to add to the subset.

As defined by Rousseeuw and Driessen (1999), a "C-step" is a concentration step and is at the heart of the algorithm. This is performed by calculating the central vector and covariance matrix of the currently selected points within the subset, and then calculating the Mahalanobis distance of each point of the dataset. These points are then sorted with the smallest h being identified as a subset which is equal to or better than the current subset. This is proven to always be the case, and is shown in Rousseeuw and Driessen (1999). Thus, by randomly selecting a subset of points and iteratively refining these, a practical approximation of the MCD can be obtained.

Other heuristic MCD estimators

Given the simplistic nature of the Fast-MCD estimator, various other (similar) estimators have appeared within the literature. For instance, the DET-MCD (deterministic-MCD) estimator (Hubert et al., 2012). In this algorithm, seven estimates of the central vector and covariance are computed in order to determine the preliminary subset for use with the Fast-MCD C-step procedure.

5.3.4   Characteristics of the MCD subset

Utilising the true (or estimated) MCD subset changes the underlying distribution of the eventual distance metric employed (in the case of this thesis, the Mahalanobis distance). There are three cases to consider:

Firstly, when the true central point and covariance matrix are known – and given that the data is multivariate normal – the Mahalanobis distance is distributed according to (Mardia et al., 1979):

$$MD_{\Sigma}^2(\mathbf{x_i}, \boldsymbol{\mu}) \sim \chi 2_p \tag{5.18}$$

In the second case, the ordinary mean and covariance estimates are utilised. This leads to an exact Beta distribution. Given the mean ($\bar{\mathbf{X}}$) and covariance (S) as per equation 5.16, the distribution is given by Gnanadesikan and Kettenring (1972) as:

$$\frac{(n-1)^2}{n} MD_S^2(\mathbf{x_i}, \bar{\mathbf{X}}) \sim Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right) \tag{5.19}$$

Finally, in the third case, the estimate of S is based on independent samples of $\mathbf{x_i}$. Two distributions are possible depending upon the location parameter. In the case that $\boldsymbol{\mu}$ is used, an exact $F-$ distribution with parameters (Hardin and Rocke, 2005)

$$\frac{np}{n-p} MD_S^2(\mathbf{x_i}, \boldsymbol{\mu}) \sim F_{p,n-p} \tag{5.20}$$

In the alternate case that $\bar{\mathbf{X}}$ is used, an approximate $F-$ distribution with the same parameters is found (Hardin and Rocke, 2005)

$$\frac{np}{n-p} MD_S^2(x_i, \bar{X}) \dot{\sim} F_{p,n-p} \tag{5.21}$$

Utilising the true or estimated MCD subset for the calculation of the Mahalanobis distance utilises multiple cases of these distributions. In their work, Hardin and Rocke (2005) show that points utilised to determine the MCD subset behave similar to the χ^2 or Beta distribution, whereas those beyond the MCD subset (the extreme points) behave like that of the $F-$ distribution. They go further to show that since S_X^* does not have a Wishart distribution, the previously detailed $F-$ distribution does not fit well. Proof is given for a replacement distribution to be employed in the case of the MCD being employed to robustly calculate distances, namely (Hardin and Rocke, 2005):

$$\frac{c(m-p+1)}{pm} MD_{S_X^*}^2(\mathbf{X_i}, \bar{\mathbf{X}}^*) \sim F_{p,m-p+1} \tag{5.22}$$

Where c, m can be estimated through Monte-Carlo simulation or asymptotically as discussed in detail in Hardin and Rocke (2005).

As such, by understanding the distribution of the robust Mahalanobis distance function, we can exploit these findings for the purposes of condition assessment. For instance, the use of these distributions makes it possible to identify outlying behaviour

within high dimensional models based upon sound statistical thresholds, thereby removing the requirement of a subjective (and potentially inaccurate or biased) threshold limit. However, these limits will only be as strong as the underlying model itself. Thus, it is necessary to explore the potential limitations of all aspects of the model, including the Fast-MCD algorithm.

5.3.5 Limitations of the Fast-MCD algorithm

Although the Fast-MCD algorithm (Rousseeuw and Driessen, 1999) provides an estimation of the MCD subset, there are inherent limitations to its use in practice.

Firstly, in its purest form, the Fast-MCD algorithm performs sub-sampling of the data. For $n > 600$, data is sampled into smaller groups and estimates are produced. These estimates are then tested on larger samples to see if they are of sufficient quality. This is done in order to increase the practicality of the algorithm. However, by ignoring data points, potentially useful information is inherently discarded. For instance, if we are exploring asset condition data, this sub-sampling has the (albeit, unlikely) potential to increase the proportion of degraded data utilised in the estimation of the MCD subset; leading to the incorrect identification of degradation as normal operational behaviour. This reduces the quality of the subset found by the algorithm.

Secondly, the majority of proposed subsets are not iterated through until convergence. Whilst this has advantages in terms of computational complexity, the potential exists for high quality solutions to be ignored due to poor initial conditions. Although empirical evidence presented in Rousseeuw and Driessen (1999) suggests that robust subsets converge quicker than non-robust subsets, no definitive proof is provided. As such, whilst this may hold in the cases explored in their work, it may not hold for other cases (such as with regards to asset condition data).

Finally, the number of iterations of the Fast-MCD algorithms required increases dramatically as the dimensionality of the data increases. The number of restarts – F-MCD$_{iteration}$ – required by the algorithm to have a $(1 - \alpha) \cdot 100\%$ confidence that the true MCD subset is found is:

$$\text{F-MCD}_{iteration} \geq \frac{\log(\alpha)}{\log(1 - p(1 - k)^d)} \tag{5.23}$$

Where d represents the dimensionality of the data, α represents the probability requirement, p the global convergence probability and k the number of outliers within the data.

As such, in order to overcome the limitations of the Fast-MCD heuristic, artificially intelligent search techniques can potentially be explored in order to provide an alter-

native means of identifying a suitable MCD subset which is of comparable (or better) quality than that provided by the traditionally utilised Fast-MCD algorithm.

5.3.6 Memetic algorithm for the estimation of the MCD subset

Genetic algorithms (Fonseca and Fleming, 1995) are bio-inspired probabilistic search algorithms based upon survival of the fittest and Darwin's theory of natural selection. The underlying concept of these algorithms is to hold a population of potential solutions which improve and evolve over time (generations) through both inter-population competition and also controlled manipulation. This class of algorithms are able to efficiently find strong solutions for complex problems, and unlike many numerical methods, are both highly robust and able to identify global optima (Grady et al., 2005).

Memetic algorithms further extend this notion by the inclusion of a local optimisation procedure. This simulates the experience of an individual within the population (Elbeltagi et al., 2005). By employing a local search, the solution space can be reduced by ensuring that only locally optimal individuals are presented for selection within the genetic algorithm. This enables the memetic algorithm to outperform traditional genetic algorithms by quickly generating a strong set of individuals, which can then be exploited to move away from the local optima, towards the global optima (Grady et al., 2005).

As the calculation of the MCD subset is NP-complete (Bernholt and Fischer, 2004), enumeration of the search space is implausible in practice, and as random search algorithms do not perform better than enumerative methods in practice (Grady et al., 2005), evolutionary algorithms such as the genetic and memetic algorithms (amongst others) provide a viable alternative to exhaustive search techniques.

In this analysis, a binary coded elitist memetic algorithm (EMA) is employed to assist in the identification of a suitable MCD subset. In this case, an individual is represented by a string of 1s and 0s, with 1 representing inclusion into the MCD subset, and 0 representing exclusion from this subset. A population is defined as a set of individuals.

A step-wise representation of a memetic algorithm can be provided as follows:

1. ***Initialisation:*** Generate initial population of size N and calculate the fitness of each individual.

2. ***Memetic search:*** For each individual, optimise based upon individual local search.

3. ***Elitist crossover:*** Construct a new population, compromising the fittest h individuals of the previous generation, and the product of the crossover of randomly

selected individuals from the previous generation.

4. **Genetic mutation:** Mutate each individual in the new population. If the mutation is beneficial after local optimisation, keep this mutation, otherwise, keep the original individual.

5. **Iteration:** Go to step 2, repeat until the fitness of the best individual is acceptable, or a predetermined number generations (or length of time) has passed.

As can be observed, memetic algorithms rely heavily on the genetic operations of crossover and mutation. Whilst this would typically limit the capabilities of the technique to those of the genetic algorithm, this is not the case. Due to the use of a local optimisation procedure, the limitations of the genetic algorithm can be overcome. Specifically, the response times for solution quality for a given genetic algorithm are typically erratic. It is uncertain if and when an improvement will be made, and unknown how substantial that improvement will be. In contrast, techniques such as gradient descent (Snyman, 2005) will typically offer continual (but often small) improvements to the solution. Utilising the local search enables continual, small improvements to be made to the solution, overcoming this limitation of the genetic algorithm, whilst providing all of the benefits which are associated with the use of genetic algorithms.

In the construction of the elitist memetic algorithm, five factors must be considered. They are:

1. **The fitness function:** To quantify solution quality.

2. **The crossover function:** In order to genetically evolve individual solutions.

3. **The mutation function:** In order to assist in escaping local optima.

4. **Solution validation:** To ensure the MCD subset constraints are met.

5. **Local search procedure:** To increase the quality of the candidate solutions.

6. **Elitist tournament selection:** To ensure strong solutions are retained.

The fitness function will be application specific. In this case, the objective function of the algorithm is to minimize the determinant of the inverse of the covariance matrix produced by the h points selected by the genetic algorithm; as such this will define the fitness function. Formally, this is:

$$Cost(x) = \det(\Sigma_h^{-1}) \qquad (5.24)$$

Where $\det(x)$ represents the determinant of x, and Σ_h represents the covariance matrix calculated on the h points selected by the EMA.

The crossover function is utilised to synthesise a new individual based upon the genetics of two parents. In this case, genes are randomly selected from each parent with equal probability. This enables new individuals to be generated, whilst retaining the characteristics of the parents gene sequences which achieved the high level of fitness associated with that parent. This is demonstrated in table 5.1.

Table 5.1: Example of a binary coded memetic algorithm crossover function.

Parent 1	011111010111
Parent 2	110011011111
Child 1	011011010111

Randomly selecting each gene from the parents will assist in keeping high quality gene sequences within the population whilst allowing for local optima to be escaped. As this is utilised within an elitist memetic algorithm (as opposed to a traditional genetic algorithm), various benefits are achieved over utilising a point-based crossover function. Specifically, due to the local optimisation performed, by randomly selecting genes from parents, if both parents agree on a gene, this is guaranteed to be included. As such, stronger solutions will naturally remain within the population. Although this has the potential for getting caught within local optima, the assistance of the mutation function helps ensure that this is not the case.

Mutation is applied to each individual of the new population. Mutation is performed in a binary coded genetic algorithm (as in this case), by having a fixed probability to randomly change a gene from 0 to 1, or from 1 to 0. In this case, this probability is defined as 1.5%. This enables populations which have converged to a local optima to more easily identify global optima. By randomly including (or excluding) certain points, many potential candidates can be explored, with the best of these candidates being utilised to create new solutions based upon the crossover function detailed above. An example of a potential mutation of a binary coded solution is given in table 5.2.

Table 5.2: Example of a binary coded memetic algorithm mutation function.

Candidate	000000000000
Mutation	000100000100

Due to the crossover and mutation functions, it is possible to generate individuals which may violate the constraints imposed by the MCD algorithm, specifically, having exactly h of n points within the subset. In order to ensure the validity of each individual,

each individual is verified. This is done by establishing the size of the subset for each individual. Three cases must be considered:

- **The selected subset size is correct:** If the subset size is correct, the individual is left unchanged.

- **The selected subset size is too large:** If too many points are included in the MCD subset, points are randomly removed until it is sound (mimicking the mutation function).

- **The selected subset size is too small:** If too few points are included in the MCD subset, points are randomly added until it is sound (mimicking the mutation function).

As the memetic algorithm incorporates a local search function to optimise individual fitness, there is a necessity to employ a heuristic or exact search function for this process. In this instance, the local search method employed is similar to that of the Fast-MCD algorithm. C-steps (Rousseeuw and Driessen, 1999) are performed until convergence to ensure that local optima are reached. Although this adds computational complexity to the search procedure, it has been shown in many cases that adding this local search enables the memetic algorithm to significantly outperform the traditional genetic algorithm (Elbeltagi et al., 2005).

It is interesting to note the behaviour of the "C-step" function and the associated behaviour of the memetic algorithm. It was noted by Rousseeuw and Driessen (1999) that strong estimates of the MCD subset will require fewer C-steps in order to converge. Whilst no proof was given, empirical evidence to suggest this was provided. This behaviour assists the elitist memetic algorithm substantially. As stronger candidates are generated, fewer C-steps will be performed. Thus, the local search component of the algorithm will require less computation after weaker candidates have been discounted throughout the search procedure. As weaker candidates can take up to 8 times as long as strong candidates to converge through C-steps (Rousseeuw and Driessen, 1999), emergent behaviour appears where fewer C-steps are required as the search continues.

This behaviour is further compounded by the utilisation of an elitist tournament selection process for the selection of individuals for genetic crossover. In this, the fittest j individuals are chosen to initially populate the following generation of candidate solutions, with the remaining individuals created through random genetic crossover of these j fittest individuals from the previous generation. This ensures that high quality MCD subsets are provided as seeds to each generation of the algorithm. Furthermore, this takes advantage of the emergent behaviour described above (as these candidates

have already been iterated through until convergence) and also the the mutation rate (as this attempts to escape the local optima potentially encountered in the C-step process).

5.3.7 Results

In order to determine the quality of the proposed elitist memetic algorithm, it is essential that the MCD subset determined by the algorithm is bench marked against both the state-of-the-art Fast-MCD algorithm, and other evolutionary techniques (such as a traditional genetic algorithm). This enables both the quality of the algorithm and the speed of each of the algorithms to be explored. To perform the analysis, implementation was performed in Java 7, on a 2.4 GHz Intel i5 machine with 4GB of memory running Windows 7. C-steps (Rousseeuw and Driessen, 1999) were performed on each Fast-MCD subset until convergence, and the elitist memetic algorithm population base was determined to be 5.

Table 5.3 shows the fitness function (the MCD subset determinant) for the elitist memetic algorithm, a traditional genetic algorithm, the Fast-MCD algorithm and the random search algorithm over a period of 1 hour. In this case, the dataset employed covered 100,000 data points of wind turbine SCADA gearbox data as discussed in both sections 4.3 and 5.2.5, and processed into the bivariate condition assessment model as per the extended physics of failure model in section 4.3. As a benchmark for comparison, in this case, the traditional covariance calculation would return a fitness of $\det(\Sigma^{-1}) = 126.28$.

Over the course of this experiment, the elitist memetic algorithm routinely out performs the Fast-MCD heuristic. The strongest subset returned by the Fast-MCD algorithm had a fitness of 20.89, which was reported 57 minutes, 15 seconds into the search. In contrast to this, the elitist memetic algorithm had found a superior solution within 15 seconds of search initialisation. This corresponds to 6311 iterations of the

Table 5.3: Comparison of search algorithms over a 1 hour search with $N = 100,000$.

Time elapsed (Min)	Random search	Fast-MCD heuristic	Genetic algorithm	Elitist memetic algorithm
Initialisation	123.97	41.36	122.02	124.70
10	123.86	21.01	117.34	20.88
20	123.78	20.94	114.20	20.88
30	123.66	20.92	112.78	20.88
40	123.61	20.91	111.95	20.88
50	123.49	20.90	110.93	20.88
60	123.47	20.89	108.80	20.88

Fast-MCD algorithm, in comparison to 6 generations of the elitist memetic algorithm. The traditional genetic algorithm outperformed the random search, however, was unable to match either the Fast-MCD heuristic or the elitist memetic algorithm. Whilst this was expected, the poor quality of the final result was unexpected, and it had been anticipated that further improvements would have been achieved in the following 60 minutes that the experiment was ran. For comparison, over the hour, 710 generations of the genetic algorithm were created.

As the probability of finding the true MCD subset with the Fast-MCD algorithm increases with the number of iterations performed as per Eq. 5.23, and given that the number of iterations are directly proportional to time, it is worth investigating whether or not the elitist memetic algorithm out performs the Fast-MCD for a fixed time period of a shorter duration. Table 5.4 presents the averages of the fitness functions returned by each of the algorithms for 40 searches for the duration of of 1 minute for 100, 1,000, 10,000, and 100,000 data points. In each case, both the elitist memetic algorithm and Fast-MCD heuristic outperform both the random search algorithm (as anticipated) and also the traditional genetic algorithm.

It is interesting to note that for each of the 4 sample sizes evaluated (100, 1,000, 10,000 and 100,000 data points), the worst subset found by the elitist memetic algorithm, was better than or equal to the best subset found by the Fast-MCD heuristic. For $N = 100$, both the elitist memetic algorithm and the Fast-MCD heuristic returned the same subset in each of the 40 cases. For $N = 1,000$, the 40 results returned by the memetic algorithm were identical, however, this subset was only found twice by the Fast-MCD heuristic (2 cases of 40, or 5%). For $N = 10,000$, the elitist memetic

Table 5.4: Average fitness of search algorithms (40 samples over 1 minute).

		N			
		100	1,000	10,000	100,000
	Min	36.85	95.20	116.65	122.87
Random	Mean	40.42	100.79	119.26	123.61
	Max	43.60	104.82	121.46	123.97
	Min	15.29	36.56	83.67	119.44
Genetic	Mean	15.29	38.88	86.31	121.13
	Max	15.29	41.14	89.11	122.55
	Min	15.29	20.76	20.51	20.91
Fast-MCD	Mean	15.29	20.77	20.57	21.11
	Max	15.29	20.80	20.62	21.38
	Min	15.29	20.76	20.50	20.88
EMA	Mean	15.29	20.76	20.50	20.88
	Max	15.29	20.76	20.50	20.89
Traditional		129.50	139.29	124.22	126.28

algorithm outperformed the Fast-MCD heuristic in all cases, and was able to find 18 unique, superior solutions which were not found al all by the Fast-MCD heuristic. This trend continues with $N = 100,000$, where all 35 unique solutions of the 40 solutions found by the elitist memetic algorithm outperformed the Fast-MCD heuristic.

In order to directly assess the quality of the elitist memetic algorithm, two-tailed t-tests were performed to see if the elitist memetic algorithm consistently outperformed the Fast-MCD heuristic for the given sample sizes. The case of $N = 100$ was not evaluated as both algorithms returned the same subset in each of the 40 iterations. As such, in this case it can only be stated that the elitist memetic algorithm performed equal to that of the heuristic for this dataset and sample size. For $N = 1,000$, significant differences between the population fitness of the elitist memetic algorithm ($M = 20.761$, $SD = 0$) and the Fast-MCD heuristic ($M = 20.773$, $SD = .001$) were found ($t(39) = 7.929$, $p < .001$).

A similar result was found for the case of $N = 10,000$, with the elitist memetic algorithm ($M = 20.509$, $SD < .001$) outperforming the Fast-MCD heuristic ($M = 20.570$, $SD = .260$) significantly ($t(39) = 14.301$, $p < .001$). Finally, the result held for N=$100,000$ with the elitist memetic algorithm ($M = 20.888$, $SD < .001$) again out performing the Fast-MCD heuristic ($M = 21.112$, $SD = .261$) at a significant level ($t(39) = 11.859$, $p < .001$).

Statistical analysis between the elitist memetic algorithm and either the traditional genetic algorithm or random search algorithm was not performed due to the poor quality of the MCD subset returned by the genetic algorithm on the larger samples ($N > 100$).

To further understand the variation between the elitist memetic algorithm and the Fast-MCD heuristic, the differences between the best subsets found over the 1 minute experiment were explored. Tables 5.5, 5.6 and 5.7 show the elitist memetic algorithm compared to the random search, the genetic algorithm and the Fast-MCD heuristic (respectively). It can be seen that identical subsets are returned for smaller samples (namely $N = 100$ and $N = 1,000$) by the elitist memetic algorithm and the Fast-MCD heuristic. For larger samples, there is considerable agreement between the chosen subsets, with $< 1\%$ of the data pertaining to the benefit provided by the elitist memetic algorithm. There is larger deviation between the genetic algorithm and the elitist memetic algorithm in comparison to that of the Fast-MCD heuristic and the elitist memetic algorithm. This suggests that the genetic algorithm without the local optimisation (as per the elitist memetic algorithm) is not well suited to this application.

Table 5.5: Comparison of random search against the elitist memetic algorithm (EMA).

N	Elitist memetic search	Random algorithm	Agreement	Longest matched group	Longest different group	Avg. matched length	Avg. different length
100	15.29	30.70	82	12	3	5.20	1.20
1,000	20.76	95.20	670	14	7	2.92	1.44
10,000	20.50	115.56	6274	18	8	2.68	1.59
100,000	20.88	123.55	62915	25	13	2.68	1.58

5.3.8   Discussion

For both the $N = 100$ and $N = 1,000$ cases, identical subsets are returned by the elitist memetic algorithm and Fast-MCD algorithm (in all 40 of 40 cases for $N = 100$, and 2 of 40 cases for $N = 1,000$). Due to the size of the search space, it is impossible to know if this is the true MCD subset, or if both algorithms were caught in within local optima.

In all cases evaluated, the elitist memetic algorithm was able to return the same subset as the Fast-MCD heuristic, or was able to provide a superior MCD subset. Whilst this does not prove that the elitist memetic algorithm consistently outperforms the Fast-MCD heuristic, it does provide some empirical evidence to suggest that for this dataset in particular, some benefit is gained through the application of the elitist memetic algorithm in this case.

The quality of the result generated by the elitist memetic algorithm is down to the unique interaction between the local search heuristic, crossover functions and the mutation function which are employed for this purpose. Due to the Fast-MCD heuristic being employed for the local search optimisation within the elitist memetic algorithm, local optima are encountered quickly. This is a limitation of the Fast-MCD heuristic and can be exploited by the application of genetic operators. Due to the nature of the crossover function, if both parents agree on the inclusion of a point within the MCD subset, this point is guaranteed to be included. As such, this process retains the

Table 5.6: Comparison of the traditional genetic algorithm against the elitist memetic algorithm (EMA).

N	Elitist memetic Algorithm	Genetic Algorithm	Agreement	Longest matched group	Longest different group	Average matched length	Average different length
100	15.29	15.29	100	100	0	100	0
1,000	20.76	38.88	834	28	4	6.17	1.22
10,000	20.50	86.31	6796	22	6	3.15	1.49
100,000	20.88	121.13	63078	30	14	2.72	1.59

Table 5.7: Comparison of Fast-MCD heuristic algorithm against the elitist memetic algorithm (EMA).

N	Elitist memetic algorithm	Fast-MCD heuristic	Agreement	Longest matched group	Longest different group	Average matched length	Average different length
100	15.29	15.30	100	100	0	100	0
1,000	20.76	20.76	1000	1000	0	1000	0
10,000	20.50	20.52	9932	643	1	145.82	1.00
100,000	20.88	20.92	99021	571	3	102.05	1.01

strongest genes, whilst effectively searching through those which remain.

Whilst this would typically be detrimental to the algorithm (due to the high probability of simply returning the Fast-MCD subset), the mutation function ensures that this is not the case. The mutation function allows individual solutions and the population in general to escape a local optimum, enabling stronger results than the Fast-MCD algorithm alone. A mutation rate of 1.5% represents an expectation of 1,500 points of the 100,000 in the largest dataset explored to be mutated for each individual, each generation. As this is an elitist algorithm, the best results are kept. Thus, even a slight improvement over previous results is utilised as the basis within the next generation of candidate solutions.

5.3.9 Conclusion

In this section, an elitist memetic algorithm was compared to a random search, a traditional genetic algorithm and the Fast-MCD heuristic for determining the minimum covariance determinant (MCD) of four datasets. Calculation of this value is essential to accurately encapsulate the normal operational behaviour of assets in the multivariate domain when employing the robust Mahalanobis distance (RMD).

As the elitist memetic algorithm is shown to be able to outperform the Fast-MCD heuristic, the use of the elitist memetic algorithm should be used in practice to effectively determine the MCD subset in cases where it is impractical to perform exact enumeration. It is recommended that the use of the elitist memetic algorithm is employed when $N > 100$. This figure is chosen as both the elitist memetic algorithm and Fast-MCD heuristic repeatedly find the same subset for small values of N, however, the elitist memetic algorithm in this case always returned a better or equal to subset for $N > 1,000$, and thus, should be applied in practical applications where N exceeds this value.

Although both the random search algorithm and traditional genetic algorithm out perform the traditional covariance calculation, given the practicality of utilising either

the Fast-MCD heuristic or the elitist memetic algorithm as an estimator for the MCD subset, these techniques should be avoided.

5.4 Chapter summary

In this chapter we have explored the potential of utilising multivariate distance metrics as a means of condition assessment. Both high frequency and low frequency data sources were explored (namely; high frequency bearing accelerometer data and low frequency wind turbine gearbox SCADA data). In total, 8 multivariate distance metrics and 1 multivariate similarity metric were comparatively evaluated for their sensitivity and robustness. It was found that the Mahalanobis distance, and specifically the robust Mahalanobis distance based upon the minimum covariance determinant (MCD) subset was the most sensitive and provided the greatest insight in this case.

Due to this, exploration of the characteristics of the MCD subset was performed. Exact algorithms were shown to not be feasible in practice, and alternative estimations were sought. Due to the NP-complete nature of the problem, AI search techniques were utilised to overcome the limitations of exact algorithms. The traditional covariance calculation was comparatively evaluated to four techniques. Namely, a random search, a genetic algorithm, the Fast-MCD heuristic (Rousseeuw and Driessen, 1999) and a novel elitist memetic algorithm. The elitist memetic algorithm was found in this case to be comparable to the Fast-MCD heuristic in all cases, returning an equivalent or superior solution.

As the elitist memetic algorithm is able to determine a strong subset of points for the calculation of the MCD subset, this assists in providing a more accurate and robust representation of the Mahalanobis distance. Having a more sensitive distance metric – as has already been shown in this chapter – is essential when applying these techniques for the purpose of condition assessment. As such, by enabling the distance metric to be more sensitive by utilising a strong encapsulation of a desired behaviour (such as normal operational behaviour), this can be exploited within an enterprise to provide actionable intelligence to an organisation in terms of maintenance decisions, planning and scheduling. However, quantifying asset or component condition alone does not provide realised benefits to an organisation or enterprise. The knowledge provided by this quantification process must be exploited by the organisation or enterprise in order begin to realise the benefits which can be attained through a CBM strategy or PHM system. As such, the exploitation of this derived metric for the purposes of condition assessment is performed in the following chapter, chapter 6.

5.5   Research objective 2 (RO2)

This chapter has explored the use of multivariate distance and similarity metrics and the robust encapsulation of asset behaviour through exploiting artificially intelligent search techniques. As such, this chapter provides insight into research objective 2 (RO2) specified in this thesis in section 3.2.2. RO2 is stated as:

> **RO2: Robust multivariate distance metrics for condition assessment**
>
> *To explore the use of traditional and robust multivariate statistical distance metrics to quantify asset condition.*

This is further broken down into two research questions (RQ3, RQ4). The results of the analysis performed in this chapter will now be utilised to answer this research objective and the associated research questions.

5.5.1   RQ3: Multivariate distance metrics for asset degradation quantification

Research question 3 (RQ3) explores the use of multivariate distance metrics to quantify asset or component condition. As such, this research question corresponds to the research undertaken in section 5.2 of this thesis. RQ3 is:

> **RQ3:** *Can the use of multivariate distance or similarity metrics be exploited to empirically derive a metric for condition assessment? If so, which metrics are the most sensitive to changes in asset health?* **(RO2)**

In order to answer this question, the six sub-questions proposed in section 3.3.3 must be answered.

Initially, the identification of multivariate distance metrics was undertaken in the systematic literature review undertaken in chapter 2. This identified a substantial number of distance metrics which could be explored. The work of Deza and Deza (2009) provides an extensive enumeration of all relevant metrics for this purpose. From this work, a selection of applicable metrics were chosen based upon their potential to be incorporated into a condition metric and also their popularity within the literature.

In total, 9 multivariate metrics were selected. Of these, the three Minkowski distances along with the Mahalanobis distance have previously been explored with regards to condition monitoring. However, cosine similarity, the Lance-Williams distance, the Penrose distance and also the robust Mahalanobis distance had not previously been

widely explored. Whilst it cannot be stated for certain why these metrics have not been utilised for condition assessment, it is likely due to the limited body of research employing multivariate metrics within the realm of PHM systems.

It does not seem that the data acquisition rate effects the choice of multivariate distance metric in this case. It appears that the more information the metric exploits with regards to the underlying data, the better suited the metric is to condition assessment. This can be seen in the differences between the Minkowski distances, the Penrose distance, the Mahalanobis distance and finally the robust Mahalanobis distance.

Of all the metrics, the most computationally simple is that of the Manhattan distance. This has the fewest computational operations, however, the capabilities of this metric were shown to be limited. All of the distance metrics aside from the robust Mahalanobis distance can be computed in real time. The robust Mahalanobis distance can be computed in real time in cases where effective algorithms are employed to determine the minimum covariance determinant subset for the purposes of defining the attribute centres and covariance.

In section 5.2.6, the robust Mahalanobis distance is found to be the most sensitive to degradation. This is perhaps due to the robust nature of this metric. Following this, the traditional Mahalanobis and then the Penrose distance are found to be somewhat sensitive, but not as sensitive as the robust Mahalanobis distance.

As section 5.2 builds upon the analysis performed in chapter 4 it should be apparent that the quality of the derived condition index is directly dependent upon the quality of the features which are employed as a basis to create the index. Furthermore, the quality of the derived metric is also highly dependent upon the fixing of a point which accurately represents normal operational behaviour. In the case that poor quality features are employed a long with a poor encapsulation of normal asset behaviour, then it is unlikely that analysis employing multivariate distance metrics would provide any meaningful insight.

As such, in answering RQ3, multivariate distance or similarity metrics can be exploited to empirically derive a metric for condition assessment. Furthermore, of the 9 multivariate metrics explored, the robust Mahalanobis distance is shown to be the most sensitive of the metrics explored in section 5.2.

5.5.2 RQ4: Encapsulation of normal operational behaviour

Research question 4 (RQ4) explores the possibility of utilising artificially intelligent search techniques to encapsulate asset behaviour. Research question 4 is stated in section 3.3.4 as:

RQ4: Can the use of AI search techniques be employed to assist in deriving a robust encapsulation of the data which can be employed in conjunction with distance metrics? If so, how do these techniques compare to current exact or heuristic techniques? (RO2)

In order to answer this, the four sub-questions proposed in section 3.3.4 must be answered.

The first sub-question is again answered initially through the systematic review of the literature undertaken in chapter 2. Typically, encapsulation of normal behaviour is performed in data-driven techniques (such as neural networks) based upon model residuals. Alternatively, this encapsulation is performed when physics of failure models are presented. Typically, these approaches model normal behaviour as a fixed point, rather than a range of behaviours. Statistical techniques often model normal behaviour as a distribution of behaviours. Although all of the techniques were extensible into the multivariate domain, some techniques (such as neural networks) are less extensible than others (due to their computational complexity).

Typically, data-driven techniques are the most common technique to encapsulate behaviour, followed by statistical techniques and finally physics of failure. Whilst an almost infinite number of encapsulation techniques exist, the minimum covariance determinant (MCD) subset in conjunction with the Mahalanobis distance has only been explored in a limited number of works in the realm of PHM.

When employing the minimum covariance determinant, due to the NP-complete nature of this problem, it is possible to reduce this to an alternative NP-complete problem which is well suited to artificially intelligent search techniques (such as the travelling salesman problem, or a maximal clique). It is known that exhaustive search techniques (for instance, via breadth or depth first search) are not well suited to these types of problems. As such, alternative artificially intelligent techniques are required. Genetic algorithms provide a means to effectively search the search space, however, these often get trapped in local optima. As such, the inclusion of a heuristic local search is recommended to increase the quality of the search. It is anticipated that most evolutionary search algorithms would provide improvement over traditional techniques, with some (such as in this case with the elitist memetic algorithm) outperforming the Fast-MCD heuristic.

Due to the NP-complete nature of the subset selection problem, exhaustive techniques are not feasible on real-world condition monitoring data. As such, heuristic techniques are often employed such as the Fast-MCD algorithm (Rousseeuw and Driessen, 1999). In section 5.3, an elitist memetic algorithm was utilised incorporating the key elements of the Fast-MCD algorithm and is shown to provide a more precise encapsu-

lation of the data in this case. Due to limitations imposed in practice, this outperforms random search, genetic search and the Fast-MCD heuristic.

As such, in answering RQ4, it is possible for AI search techniques to be employed to assist in deriving a robust encapsulation of the data. Furthermore, the identified technique (the elitist memetic algorithm) can be employed in conjunction with multivariate metrics. The elitist memetic algorithm can be positively compared to both current exact and heuristic techniques.

5.5.3   RO2: Robust multivariate distance metrics for asset condition assessment

Having explored both RQ3 and RQ4, it is possible to answer research objective 2 (RO2). This states:

RO2: Robust multivariate distance metrics for asset condition assessment

> *To explore the use of traditional and robust multivariate statistical distance metrics to quantify asset condition*

Having explored an extensive array of multivariate distance metrics in section 5.2, many of these have been shown to be viable metrics for the purposes of condition assessment. Overall, the robust Mahalanobis distance was the most sensitive and the most robust to both noise and degradation. Although the Minkowski distances are potentially useful, the simple nature of these metrics limits their applicability. The cosine distance, Lance-Williams distance and Canberra distance are all viable candidates, however, may not perform as well as the Penrose and traditional Mahalanobis distances.

Section 5.3 explored the use of artificially intelligent search techniques with regards to the estimation of the robust Mahalanobis distance parameters. The use of an elitist memetic algorithm was found to be capable of encapsulating normal operational behaviour and found to perform favourably to the state-of-the-art technique commonly employed; the Fast-MCD heuristic (Rousseeuw and Driessen, 1999).

As such, RO2 has been answered throughout this chapter. The robust Mahalanobis distance (RMD) has been shown to be effective as a metric for the purposes of condition assessment, and the encapsulation of normal operational behaviour has been shown to be of a higher quality when this is utilised in conjunction with artificially intelligent search techniques.

Chapter 6

Knowledge Exploitation

Having derived a robust metric for the purpose of condition assessment, this can be exploited to realise many of the ambitions of a condition based (proactive) maintenance strategy. However, merely having a quantification of the condition of an asset is not enough to realise these goals. For instance, it is useful to estimate (or predict) when a component will fail, to estimate the condition of an asset in the future, the likely failure mode (predictive diagnostics) and also to understand the conditions which caused the failure to occur.

This chapter explores the potential of employing the derived robust Mahalanobis distance metric for condition assessment to perform these analyses. In section 6.2, tacit knowledge is extracted to assist in understanding why a failure occurred. In section 6.3, the potential for predictive diagnosis is explored. Finally, in section 6.4, the remaining useful life of components is estimated.
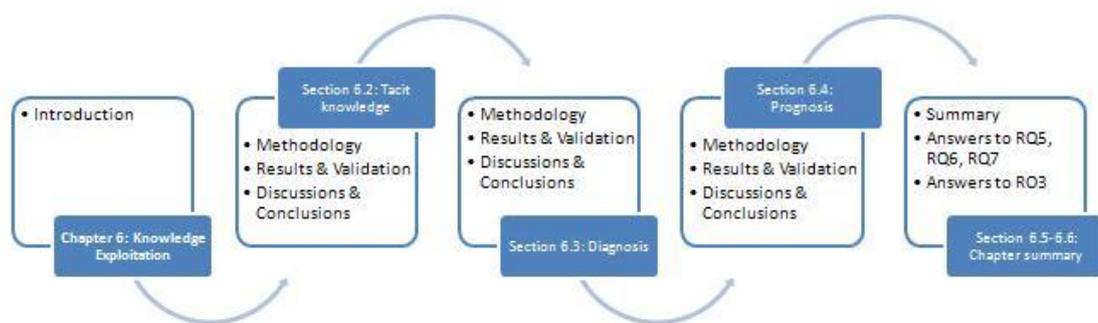
Figure 6.1: Overview of the structure of chapter 6.

6.1   Introduction

Throughout this thesis, the development of a novel methodology has been presented to enable the use of a condition metric within a PHM system on both bearings and gearboxes. In the case of this thesis, a PHM methodology has been specifically developed for the purposes of wind turbine gearbox condition assessment. In chapter 4, high and low frequency models were built based upon multivariate models. These multivariate models were utilised in conjunction with multivariate distance metrics in chapter 5 to create a univariate time series which provided insight into asset condition. However, whilst this can be utilised in practice to assist in the development of a condition based maintenance (or proactive maintenance) strategy, more information and knowledge can be extracted from the developed condition metric and associated techniques.

Due to the nature of distance metrics, these functions are unbounded. Within this work, this is not an accurate representation of degradation. A new component will have little degradation, however, the amount the component can potentially degrade is limited (the asset cannot degrade indefinitely, as it must fail at some point). As such, exploiting the distribution of the distance it is possible to objectively limit the bounds of the distance distribution when performing condition assessment. By exploiting this knowledge, a means to rapidly label vast repositories of data is also presented. This provides the capability to exploit known data mining and machine learning techniques to develop expert systems in order to assist in maintenance planning and scheduling.

The notion of distance and similarity is a continuing theme throughout this thesis and is explored in further detail throughout this chapter. For instance, it is intuitive to explore the notion of similarity with regards to failure data rather than normal operational behaviour. This enables the potential for predictive diagnostics to be undertaken, assisting in the planning and scheduling of maintenance.

Furthermore, it is of interest to take the derived health metric and explore the capabilities of asset condition prediction and remaining useful life (RUL) estimation. Given the current condition of an asset, it is natural to ask what the condition will be like at a given point in the future. In order to realise substantial maintenance savings, accurate remaining useful life estimation is essential. If the developed technique under estimates the RUL, over-maintenance will occur. Similarly, if the developed technique over estimates the RUL, failure will occur and unscheduled maintenance must be performed at an increased cost.

6.2 Automated knowledge extraction

The contents of this section appear in part in the following peer-reviewed publication:

- Godwin, J. L., & Matthews, P. (2014, January). *Rapid labelling of SCADA data to extract transparent rules using RIPPER*. In Reliability and Maintainability Symposium (RAMS), 2014 Annual (pp. 1-7). IEEE, Colorado Springs, CO, USA. **IEEE T.L. Fagan Award Recipient**

6.2.1 Introduction

Having created a statistically sound and robust metric in chapter 5 which can be used to quantify the degradation of assets, this can be exploited for retrospective analysis over vast quantities of historical data in order to extract tacit knowledge. Tacit knowledge is often defined as knowledge which is difficult to transfer or transmit to another person either written or orally (Collins, 2001). For instance, recognising a familiar face amongst many is often trivial, yet it is difficult to communicate how this is done.

As such, data repositories such as SCADA databases often contain many tacit elements which encapsulate various aspects of the enterprise, for instance:

- Data pertaining to decisions made about maintenance actions.

- Data pertaining to pre-cursors to failure (such as operational behaviour and external conditions).

- Data pertaining to standard operating procedures (SOPs) of the enterprise.

For instance, if a maintenance operator decides to service a component, the history of this component is available and can be analysed to determine not only if this action was justified, but also to extract the tacit knowledge of the operator to assist in understanding why this decision was made. It may be the case that standard operating procedures do not cover all eventualities (i.e. "service the gearbox every 12 months") when in reality, the maintenance operator understands implicitly the precursors to the gearbox failure. Thus, extracting this tacit knowledge would enable increased availability whilst reducing costs (through improved maintenance policy) within the enterprise.

This section exploits the metric derived throughout this thesis in order to label data for processing with traditional data-mining techniques for the development of expert system classifiers and tacit knowledge extraction. Given a set of prior examples (training data), data mining techniques enable the creation of decision trees, tables, and knowledge bases of propositional rules amongst others (Hall et al., 2009). This allows the automated classification of new data based on observations made regarding previous data.

6.2.2   Tacit knowledge and rule extraction

Although many data mining techniques exist to serve a multitude of purposes, for this work the codification of tacit knowledge through rule extraction is explored. This is because of the transparent nature of these techniques and their often simple computation. In contrast, techniques such as artificial neural networks (ANN) and support vector machines (SVM) are difficult to understand due to their "black-box" nature. Although work has been done previously in extracting rules from neural networks, such as in Baesens et al. (2003), this is not explored here and is beyond the scope of this thesis.

In this work, it is of interest to extract meaningful knowledge from historical data. To do this, the use of propositional rule learners, decision trees or decision tables can be employed. As a motivating example, figure 6.2 and tables 6.1 to 6.2 provide hypothetical examples of decision trees (such as from the C4.5 algorithm in figure 6.2), decision tables (such as from the decision table algorithm in table 6.1) and also propositional rules (such as from the RIPPER propositional rule learner algorithm in table 6.2), based upon the IRIS dataset (Fisher, 1936).

For the purposes of this work, these three algorithms are explored for their potential to extract and codify tacit knowledge based upon the condition metric which has been developed and explored throughout this thesis. The RIPPER algorithm, C4.5 and the decision table algorithm are explained in detail below.
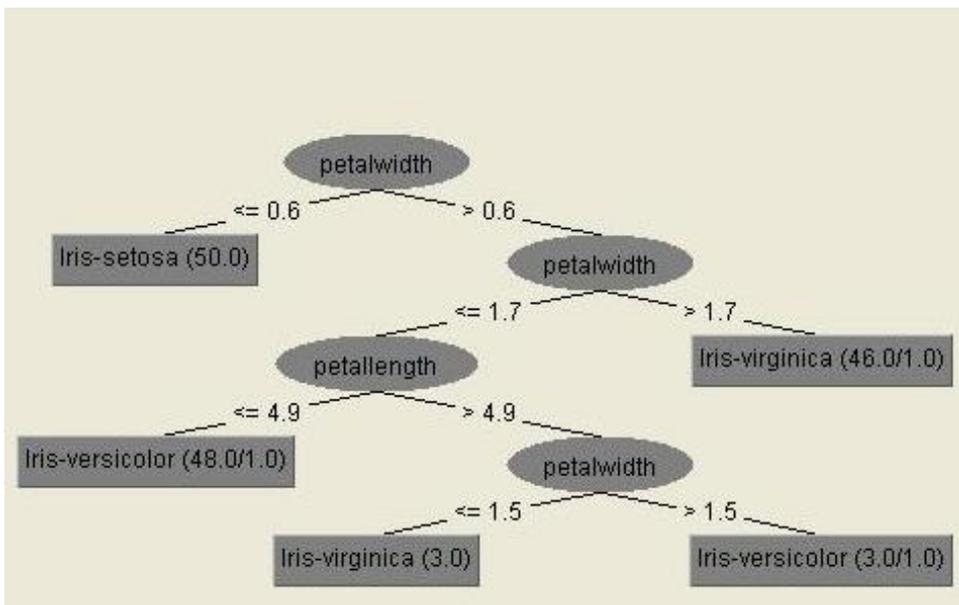
Figure 6.2: An example decision tree (C4.5) from the IRIS dataset (Fisher, 1936).

Table 6.1: An example decision table (C4.5) from the IRIS dataset (Fisher, 1936).

Petal Width	Class
(1.75 to ∞)	Iris-virginica
(0.8 to 1.75)	Iris-versicolor
(-∞ to 0.8)	Iris-setosa

Table 6.2: An example rule base (RIPPER) from the IRIS dataset (Fisher, 1936).

Rule Number	Rule	Class
1	petal length $\geq$ 3.3 & petal width $\leq$ 1.6 & petal length $\leq$ 4.9	Iris-versicolor
2	petal length $\leq$ 1.9	Iris-setosa
3	Else	Iris-virginica

Ripper

The RIPPER algorithm is an extension of the IREP algorithm which was utilized in prior generation decision tree algorithms. Data is randomly split into two sets, namely the growing set (for creating rules) and pruning set (for removing rules), typically at a 2:1 ratio (but can be at N:1). Conditions are greedily added to a rule until it is 100% accurate by adding conditions which maximize Foils information gain criteria (Cohen and Singer, 1999). This is defined as:

$$Foil(L, R) = t \left\{ \log_2 \frac{p_1}{p_1 + n_1} - \log_2 \frac{p_0}{p_0 + n_0} \right\} \tag{6.1}$$

Where L is the potential candidate (the literal condition) to be added to the rule R, p_0 is the number of positive instances (accurate classifications) for rule R, n_0 is the number of negative instances (inaccurate classifications) of R, p_1 is the number of positive instances of $R + L$, n_1 is the number of negative instances of $R + L$, and t is the number of positive instances of R covered by $R + L$.

The grown rule is then immediately pruned by removing conditions which decrease the accuracy of the rule on the pruning data partition. Rules are continually grown until all instances are covered. Once the rule base has been created, optimization of these rules is then performed. This is done as follows: for each rule created by the algorithm, two new variants are produced: the revision and the replacement. The revision is created by greedily adding conditions to the rule and then pruning, aiming to maximize the accuracy of the entire rule base. The replacement which is created is a new rule which has been grown and pruned which also aims to maximize the accuracy of the entire rule base. Pseudo-code for this algorithm, with further details can be found in Cohen (1995b) and Cohen and Singer (1999).

C4.5

The C4.5 algorithm (Quinlan, 1993) builds upon the ID3 (Iterative Dichtomiser) algorithm. In this, information (Shannon) entropy is used as a means of quantifying the quality of a decision to assist in building a decision tree. The ID3 algorithm is as follows:

1. Calculate the entropy of each attribute (feature) in the dataset.

2. Split the dataset based upon the attribute which has the lowest entropy (i.e. lowest uncertainty, or highest information gain).

3. Add a decision tree node based upon the chosen attribute.

4. Recurse on subsets for all remaining attributes.

However, as the ID3 algorithm uses a greedy approach (step 2), often the algorithm gets stuck in local optima. Similarly, the technique can also over fit to the data (over training) and as such, requires careful preparation of the data in order to avoid this. The C4.5 algorithm (Quinlan, 1993) fixes many of these issues, and is as follows:

1. For each feature in the dataset:

   (a) Calculate the information gain (difference in entropy) when splitting on this feature.

2. Create a decision node that splits on the best attribute identified in the previous step.

3. Recurse on subsets obtain from this splitting and add nodes as a children of this node.

Due to the changes, this algorithm does not suffer the same limitations of ID3 (Quinlan, 1986) and can handle missing, continuous and discrete attributes, and can handle attributes with different weights.

Decision table

Decision tables provide a means to codify knowledge in a schema (a set of features included in the table) and a body (a list of labelled instances). Building of the schema and body is done through induction with the aim of minimising classification error.

In the case of decision tables, as the number of feature subsets is large (2^n, where n represents the number of features in the dataset), the problem is modelled as a search

problem. For inductive decision tables, identification of the appropriate features to employ typically utilise a best-first search as per Kohavi (1995).

Once a set of features has been identified, induction is essentially used to create a decision table majority (DTM) where the default rule represents the majority class. A subset of features can then be identified as before, with a new DTM created as a part of this DTM, hence the inductive nature of this approach. Full details of this algorithm are given in Kohavi (1995).

6.2.3 Methodology

The extended physics of failure model for wind turbine gearbox condition assessment based upon low frequency SCADA data presented in section 4.3 was utilized in conjunction with the distance metric chosen in chapter 5 to create univariate time series incorporating both gearbox oil temperature and planetary gear temperature. Data from 6 wind turbines was collected over 28 months, with samples taken every 10 minutes. Three of these turbines were utilized to determine attribute covariance and central points for the parameters in robust Mahalanobis distance as in chapter 5. The RMD values for the remaining 3 turbines (referred to as test turbines 1–3, as in section 4.3) were then calculated. In total 380,593 instances were utilized for training data with the remaining 380,592 available instances employed for test data. Missing data was removed as it does not encapsulate the current operating condition of the gearbox. However, duplicate, erroneous and implausible values (Sainz et al., 2009) were retained in the data. Due to the scarcity of catastrophic gearbox failures, only 1 failure was present within the data. This was placed in the test set so that the identification of this fault would assist in validating the technique as no failure data had been utilized in training.

In this case, the parameterised approximate $F-$distribution discussed in section 5.3.4 could not be utilised as the extended physics of failure model in this case is not multivariate normal. This was assessed though four multivariate normality statistics, inspection of the scatter plot (due to the low dimensionality of this model in particular) and also inspection of the QQ plot. In this case, Royston's $H = 47.71, p < .001$ (Royston, 1983), Mardia's Skewness (Mardia, 1970) $= 90.95, p < .001$, Mardia's Kurtosis (Mardia, 1970) $= 6.13, p < .001$ and Henze-Zirkler's $T = 3.80, p < .001$ (Henze and Zirkler, 1990). Thus, the data (in this case) cannot be said to be multivariate normal.

Inspection of the scatter plot provides some insight into why the data is not multivariate normal. As can be seen in figure 6.3, to the right hand side, a cluster of high distances are found. This contamination potentially relates to the degradation of the gearbox, however, without an additional gearbox failure to explore, this cannot be
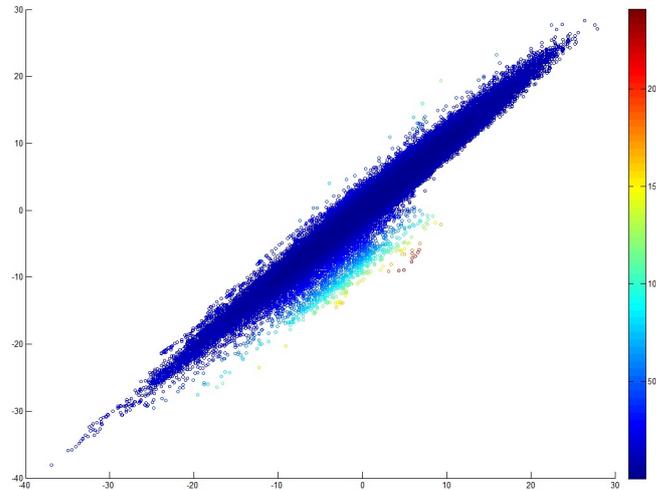
Figure 6.3: Scatter plot of bivariate extended physics of failure gear box model showing the robust Mahalanobis distance (RMD).

determined for certain. As such, normality of the data cannot be assumed, and the parametrised $F-$distribution in this case cannot be utilised. As such, the traditional χ_2^2 distribution was utilised as an alternative. Although this is not ideal, the low dimensionality of this data (m) and high sample size (n) mitigate the issues of exploiting this distribution as discussed by Hardin and Rocke (2005).

For this analysis, and as the $F-$distribution cannot be employed, a $\chi_2^2 = 13.82$, $p = .001$ critical value was then utilized for determining potential damage to the gearbox, whereas a critical value of $\chi_2^2 = 9.214$, $p = .01$ was employed for recommending inspection levels. Values below this represent normal operating conditions. Although the $F-$distribution could not be employed in this case, given the large number of instances ($n = 380,593$) and small dimensionality ($m = 2$) of the data, the impact of the non-normality is minimized as these are the two mitigating factors as expressed in Hardin and Rocke (2005).

Condition index

Figures 6.4, 6.5, and 6.6 show the derived condition index on the three turbines taken from the test set (referred to as test turbines 1, 2 and 3 respectively). Test turbines 4 though 6 were employed to derive the MCD subset for encapsulating normal behaviour. To reduce noise within the prognostic, the RMD value has been aggregated daily; reducing the number of data points displayed on each time series from 126,864 to 881. An abnormal event can clearly be seen occurring in figure 6.5. This corresponds to the

catastrophic gearbox failure which occurred on the turbine. Whilst noise exists in the other two wind turbine time series, at no point does it remain consistently higher than the $\chi^2_2 = 13.82$, $p = .001$ threshold specified a priori for damage to the gearbox.

Figures 6.4 and 6.6 do not contain failures. Although noise is present within the time series, a substantial quantity in this case is due to missing data as explored in section 4.3.9. Figure 6.4 contains 4 points at which potential degradation is identified by the distance metric. In practice, it would be possible to perform in depth analysis of the on-board high frequency data to verify that this is the case, and also to correlate this to missing data within the SCADA system itself. Similarly, figure 6.6 contains 12 points which indicate potential damage to the gearbox. Many of these (50%) occur towards the end of the time series. It may be possible that degradation was starting to occur on this gearbox, however, it may simply be noise due to ambient temperature fluctuations or poor quality data from the SCADA system. Due to the limited data available, it was not possible to identify the cause of this noise in this case.

Table 6.3 presents a breakdown of the quantity of instances for each of the test-set turbines, representing normal operating behaviour, recommended inspection levels and potential damage levels. On average, the wind turbine gearboxes exhibit normal operation behaviour 96.02% of the time. Whilst this is to be expected, it is slightly lower than anticipated. Inspections were recommended on average 2.39% of the time and damage was potentially identified 1.60% of the time. It should be noted that the turbine which suffered the catastrophic failure operated in the damaged state for 2.95% of the monitored period; higher than the 1.02% and 2.27% of the other wind turbines in the validation set. This is expected given the catastrophic failure of the turbine. The standard deviation of the RMD values is also higher for the failed turbine – 4.11 – as opposed to 2.68 and 3.52. Similarly, the maximum value reported for the failed turbine is substantially higher ($RMD = 48.57$) than the two other validation turbines ($RMD = 26.23$ and $RMD = 33.69$).

Prior to failure, the failed turbine exhibited traits of potential damage ($RMD \geq 13.82$) on 10 separate occasions. The first instance of this is 5 months before the failure, with regular signatures up until failure. For 4 consecutive months operators would have had the opportunity and capability to perform a thorough inspection of the gearbox

Table 6.3: Breakdown of data-mining labels (per turbine) in test set.

Label	Test turbine 1	Test turbine 2	Test turbine 3	Average
Normal operation	860 (97.62%)	821 (93.12%)	829 (94.10%)	846 (96.03%)
Inspection suggested	12 (1.36%)	32 (3.86%)	32 (3.63%)	21 (2.39%)
Potential damage	9 (1.02%)	26 (2.95%)	20 (2.27%)	14 (1.59%)
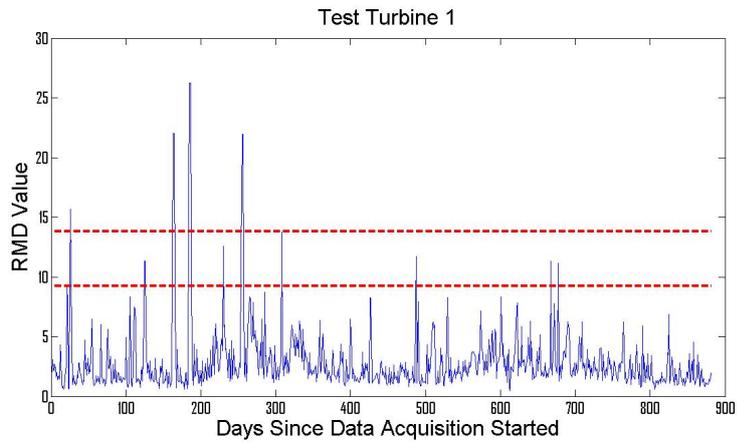
Figure 6.4: Thresholds for test turbine 1 plotted against health index over time.
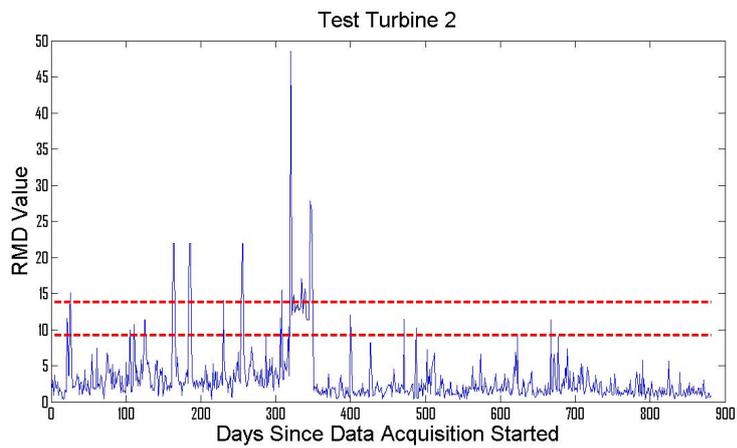
Figure 6.5: Thresholds for test turbine 2 plotted against health index over time.
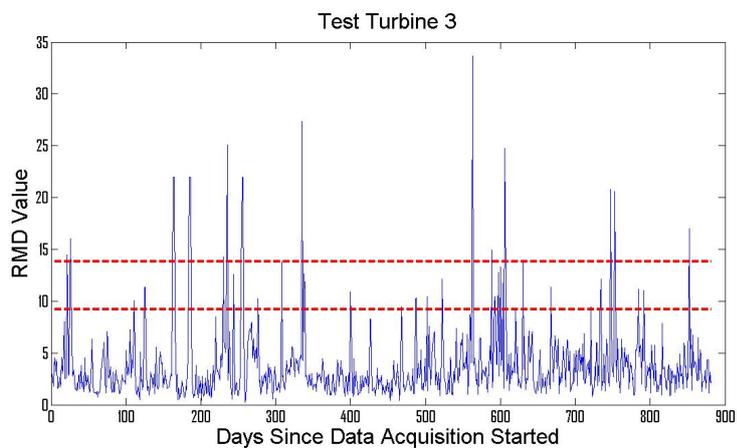
Figure 6.6: Thresholds for test turbine 3 plotted against health index over time.

during low wind (to mitigate lost production) in order to determine if any further maintenance action was required. After maintenance, no further damage is reported on the wind turbine. This is reflected in the inspection condition also; 13 events (which correlate to the damage condition) occur before failure, with 4 events occurring after maintenance has been performed. The condition index can also provide quantification of maintenance effectiveness: prior to failure, the 2 month RMD average value for the failed turbine was 6.56; higher than the average of 4.51 (providing additional context). However, after maintenance, the 2 month average dropped to 1.82; slightly lower than the 2.25 average for the other turbines. Due to the strength of the obtained condition index created utilizing the developed methodology, rule extraction was then performed on the labelled data. The three labels employed were: "normal operating behaviour", "inspection recommended" and "potential damage". This enabled transparent, human-readable rules to be extract which could express the conditions encoded within the developed prognostic index on different wind turbine SCADA attributes.

Data balancing

Due to the inherent bias placed upon the majority class (in our case, data representing normal operating condition), balancing of the classes is required. An imbalance of up to 1:1000 (representing 1 abnormal instance for each 1000 normal instances) can exist in wind turbine data (Verma and Kusiak, 2011). Work done by Godwin and Matthews (2013) shows an imbalance of between 1:125 to 1:380 with SCADA data for wind turbine pitch faults. As such, repeated random sampling with replication was utilized to balance the dataset. To do this, 20 instances were sampled with each attribute averaged to create a new data instance. This process ensures classes within the data contain the same amount of information, so that any bias apparent from class sizes in the data mining algorithms are removed.

In order to extract meaningful, transparent, human-readable rules, effective attribute selection is essential. After balancing the data using repeated random sampling, feature selection was performed through attribute selection techniques in conjunction with an independent domain expert to reduce the attributes within the model from the original 190 SCADA channels of data. Attributes were ranked by the information gain provided (in terms of entropy). There was strong consensus amongst chosen attributes. The attributes placed within the model were:

- Oil temperature.

- Gear temperature.

- Rotor speed.

- Ambient temperature.

- Energy generated.

- Wind speed.

Rules were then extracted from the developed model using the RIPPER propositional rule learner (Cohen, 1995b), the decision table algorithm (Kohavi, 1995) and C4.5 (Quinlan, 1993). In order to ensure that no over-training took place, the data was further split into two sets; training data (for growing and pruning rules) and a test set (to independently verify model accuracy). In total, 3 training and test sets were created, each holding out data from one of the three turbines which had been labelled, in order to demonstrate robustness to selection of training data.

6.2.4   Results

For each of the three algorithms, three experiments were performed. Each experiment held out one turbine for the purposes of validation. As such, two turbines were used to train each algorithm, with the final turbine data being utilised to quantify the accuracy of the developed model. In this case, the descriptive statistics employed to quantify the quality of the developed models are:

- **_Accuracy:_** This is defined as $\frac{TP+TN}{N}$ and corresponds to the percentage of correctly classified instances.

- **_Recall:_** This is defined as $\frac{TP}{TP+FN}$ and measures how good a system is at detecting positive instances. It is also known as sensitivity.

- **_Precision:_** This is defined as $\frac{TP}{TP+FP}$ and measures how good a system is at detecting negative instances. It is also known as specificity.

- **_F-measure:_** This is defined as $\frac{2TP}{2TP+FP+FN}$ and is the harmonic mean of the precision and recall.

- **_ROC-area:_** The area under the curve plotted over the true positive versus false positive instances.

Where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, FN represents the number of false negatives and N is the sample size. The analysis of these results is presented in detail below. An overview of the three experiments showing the accuracy and associated κ-statistic (Cohen, 1960) (the agreement between the model created by the data-mining techniques and the underlying condition metric) is given in table 6.4.

Table 6.4: Accuracy and κ-statistics gained for each algorithm across each experiment.

		κ-statistic	Accuracy
RIPPER	Experiment 1	96.10%	97.40%
	Experiment 2	93.84%	95.89%
	Experiment 3	92.56%	95.04%
C 4.5	Experiment 1	96.77%	97.85%
	Experiment 2	95.99%	97.32%
	Experiment 3	74.32%	82.88%
Decision table	Experiment 1	95.36%	96.90%
	Experiment 2	84.67%	89.78%
	Experiment 3	56.37%	70.92%

RIPPER

With regards to the RIPPER algorithm, a mean classification accuracy of 96.11% (95.04% - 97.40%, $SD = 1.20\%$) was achieved, along with a mean κ-statistic of 94.16% (92.56% - 96.10%, $SD = 1.79\%$). This shows that the methodology is well suited for extracting rules from this model in this manner, and over fitting is not present.

Table 6.7 presents the descriptive statistics of the first of the created models. As can be seen in table 6.8, the classification matrix produced by the RIPPER algorithm on the first independent validation set provides strong evidence that the model is accurately representing each of the classes (normal operation, inspection recommended and potential damage). The normal operation class is represented with over 98% accuracy; with 2890 of 2939 instances within the independent validation set correctly identified. The damage state is represented with 99% accuracy (2938 instances correctly classified out of 2939). Accuracy for the inspection state is lower at 93% (2760 instances correctly classified out of 2939). Although the gearbox failure was included in the training set for the purposes of rule extraction, the high accuracy attained along with the high κ-statistic shows that the technique is robust to this. This perhaps due to the balancing which was undertaken prior to the analysis.

Table 6.9 presents the descriptive statistics of the second of the created models. As can be seen in table 6.10, the classification matrix produced by the RIPPER algorithm on the second independent validation set provides strong evidence that the model is accurately representing each of the classes (normal operation, inspection recommended and potential damage). In this case, the turbine which was held out from the training dataset was that which suffered the gearbox failure. The normal operation class is represented with over 99% accuracy; with 2938 of 2939 instances within the independent validation set correctly identified. The damage state is represented with 98% accuracy (2904 instances correctly classified out of 2939). Accuracy for the inspection state is once again lower at 88% (2613 instances correctly classified out of 2939). Given that

the gearbox failure was not included within the training set for the purposes of rule extraction, the high accuracy attained along with the high κ-statistic shows that the technique is once again robust, and can be used in cases where failure data is not available.

Table 6.11 presents the descriptive statistics of the third of the created models. As can be seen in table 6.12, the classification matrix produced by the RIPPER algorithm on the final independent validation set provides strong evidence that the model is accurately representing each of the classes (normal operation, inspection recommended and potential damage). The normal operation class is represented with over 98% accuracy; with 3017 of 3051 instances within the independent validation set correctly identified. The damage state is represented with 98% accuracy (3000 instances correctly classified out of 3051). Accuracy for the inspection state is again lower at 87% (2682 instances correctly classified out of 3051). Although this test had the lowest total accuracy (95.04%) and $\kappa-$statistic (92.56%) of all experiments performed with the RIPPER algorithm, this can still be regarded as a highly accurate classifier.

For completeness, the rules generated by the second model are presented in table 6.5. In this case, "gear temperature" refers to the temperature of the planetary gears. In total, 12 rules are extracted which encapsulate the data in transparent human readable rules, the underlying classifier (in this case, the robust Mahalanobis distance). These are discussed in section 6.2.5.

Table 6.5: Extracted RIPPER rules encapsulating gearbox failure.

Rule number	Rule	Class
1	Rotor Speed $\leq$ 1 RPM	Potential Damage.
2	Rotor Speed $\geq$ 17 RPM	Potential Damage.
3	Oil Temperature $\geq$ 48.4 Degrees, and Gear Temperature $\leq$ 47.79 Degrees	Potential Damage.
4	Energy Generated $\leq$ 0.25 MW and Rotor Speed $\geq$ 6 RPM	Potential Damage.
5	Ambient Temperature $\geq$ 25.2 Degrees	Potential Damage.
6	Rotor Speed $\geq$ 14 RPM and Energy Generated $\leq$ 0.92 MW and Gear Temperature$\leq$ 47.79 Degree	Potential Damage.
7	Gear Temperature $\geq$ 51.20 Degrees	Potential Damage.
8	Energy Generated $\leq$ 0.33 MW Rotor Speed $\geq$ 11 RPM	Potential Damage.
9	Energy Generated $\leq$ 0 MW	Potential Damage.
10	Ambient Temperature $\leq$ 5.1 Degrees and Oil Temperature $\geq$ 47.5 Degrees	Potential Damage.
11	Gear Temperature $\leq$ 41.17 Degrees	Recommend Inspection.
12	Else	Normal Operation

C4.5

With regards to the C4.5 algorithm, a mean classification accuracy of 92.68% (82.88% - 97.85%, $SD = 8.49\%$) was achieved, along with a mean κ-statistic of 89.03% (74.32% - 96.77%), $SD = 12.74\%$. This shows that whilst the methodology is well suited for extracting rules from this model in this manner, the technique (on average), does not perform as well as the RIPPER algorithm.

Table 6.13 presents the descriptive statistics of the first of the created models. As can be seen in table 6.14, the classification matrix produced by the C4.5 algorithm on the first independent validation set provides strong evidence that the model is accurately representing each of the classes. The normal operation class is represented with over 99% accuracy; with 2921 of 2939 instances within the independent validation set correctly identified. The damage state is strongly encapsulated with 99% accuracy (2938 instances correctly classified out of 2939). Accuracy for the inspection state is lower (as was the case with the RIPPER algorithm) at 94% (2768 instances correctly classified out of 2939). This is comparable to the results attained through the RIPPER algorithm.

Table 6.15 presents the descriptive statistics of the second of the created models. As can be seen in table 6.16, the classification matrix produced by the C4.5 algorithm on the second independent validation set provides strong evidence that the model is accurately representing each of the classes. As before, the turbine which was held out from this training dataset was that which suffered the gearbox failure. The normal operation class is represented with 100% accuracy; with 2939 of 2939 instances within the independent validation set correctly identified. Again, the damage state is encapsulated with 99% accuracy (2916 instances correctly classified out of 2939). Accuracy for the inspection state classification is once again lower at 92% (2726 instances correctly classified out of 2939), but is higher than previous results.

Table 6.17 presents the descriptive statistics of the third of the created models. As can be seen in table 6.18, the classification matrix produced by the C4.5 algorithm on the final independent validation set provides strong evidence that the model is accurately representing each of the classes. The normal operation class is represented with over 99% accuracy; with 3040 of 3051 instances within the independent validation set correctly identified. The damage state is represented with 98% accuracy (2997 instances correctly classified out of 3051). Accuracy for the inspection state is poor at just over 50% (1549 instances correctly classified out of 3051). This test had the lowest total accuracy (82.89%) and $\kappa-$statistic of all experiments performed with the C4.5 algorithm (74.32%), suggesting the true performance of this technique may be lower than the experiments presented here.

For completeness, the tree generated by the second model is shown in figure 6.7. As can be seen, although the accuracy attained is comparable to that of the RIPPER algorithm above, the trade off in this case is the size of the derived decision tree. In this case, an increase of 1.5% accuracy is attributed to the 33 nodes and 17 branches on the tree.

Decision table

With regards to the decision table algorithm, a mean classification accuracy of 85.87% (70.92% - 96.90%, $SD = 13.42\%$) was achieved, along with a mean κ-statistic of 78.80% (56.37% - 95.36%, $SD = 20.15\%$). This shows that whilst the methodology is well suited for extracting rules from this model in this manner, the technique (on average), does not perform as well as either the RIPPER or C4.5 algorithm.

Table 6.19 presents the descriptive statistics of the first of the created models. As can be seen in table 6.20, the classification matrix produced by the decision table algorithm on the first independent validation set provides strong evidence that the model is accurately representing each of the classes. The normal operation class is represented with over 99% accuracy; with 2938 of 2939 instances within the independent validation set correctly identified. The damage state is encapsulated with 98% accuracy (2888 instances correctly classified out of 2939). Accuracy for the inspection state is lower (as was the case with both the RIPPER and C4.5 algorithms) at 92% (2718 instances correctly classified out of 2939). This is lower than the results attained through the both previous experiments performed on this dataset.
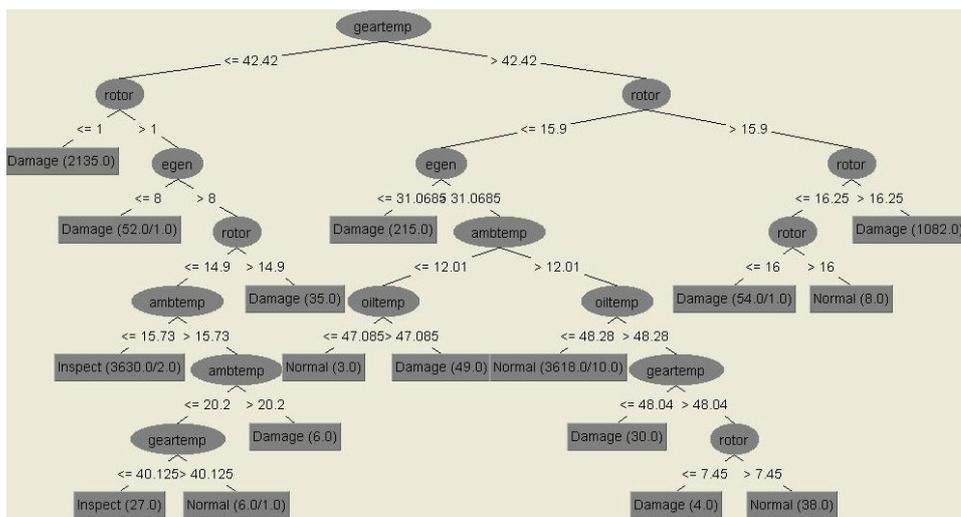
Figure 6.7: An example decision tree (C4.5) from the second test set.

Table 6.21 presents the descriptive statistics of the second of the created models. As can be seen in table 6.22, the classification matrix produced by the decision table algorithm on the second independent validation set provides strong evidence that the model is accurately representing each of the classes. As in both previous cases, the turbine which was held out from this training dataset was that which suffered the gearbox failure. The normal operation class is represented with 99% accuracy; with 2914 of 2939 instances within the independent validation set correctly identified. Again, the damage state is encapsulated with close to 98% accuracy (2880 instances correctly classified out of 2939). Accuracy for the inspection state classification is relatively poor at 72% (2122 instances correctly classified out of 2939).

Table 6.23 presents the descriptive statistics of the third of the created models. As can be seen in table 6.24, the classification matrix produced by the decision table algorithm on the final independent validation set provides strong evidence that the model is accurately representing each of the classes. The normal operation class is represented with over 99% accuracy; with 3025 of 3051 instances within the independent validation set correctly identified. The damage state is represented with 95% accuracy (2900 instances correctly classified out of 3051). Accuracy for the inspection is exceptionally poor at just over 18% (just 566 instances correctly classified out of 3051). In this case, this test had the lowest total accuracy (70.92%) and $\kappa-$statistic (56.37%) of all experiments performed with the decision table algorithm, suggesting this algorithm may not be suitable for this problem in particular.

For completeness, a subset of rules generated by the second model are presented in table 6.6. In this case, a total of 412 rules were generated by the system. As such, a subset of 10 rules are given in the table.

Table 6.6: Subset of rules extracted decision table from second test set (encapsulating gearbox failure).

Oil temperature (°C)	Rotor Speed	Class
31.1025 - 33.40	'(16.63$-\infty$)'	Damage
36.3975 - 36.40	'(16.63$-\infty$)'	Normal
40.01 - 42.03	'(16.63$-\infty$)'	Damage
37.30 - 39.98	'(16.63$-\infty$)'	Damage
47.80 - 47.80	'(16.63$-\infty$)'	Damage
42.03 - 43.80	'(16.63$-\infty$)'	Damage
48.00 - 48.00	'(16.63$-\infty$)'	Damage
48.20 - 48.20	'(16.63$-\infty$)'	Damage
48.29 - 48.30	'(16.63$-\infty$)'	Damage
14.42 - 31.10	'(16.63$-\infty$)'	Damage

6.2.5 Discussion

Due to the size of the generated rule bases for the decision table algorithm (540, 412 and 176 rules for the three test cases, respectively), it is not recommended that this technique is utilised. Similarly, this algorithm performed worse than both the RIPPER algorithm and also the C4.5 algorithm. Although the C4.5 algorithm does perform well as a classifier, it is not as consistent as the RIPPER algorithm in this case. Given that the RIPPER algorithm has the highest consistency, the highest average accuracy and the highest average $\kappa-$statistic, it is recommended that the RIPPER algorithm be employed for the purpose of rule extraction. Furthermore, given the substantially smaller knowledge base produced by the RIPPER algorithm (10, 12 and 13 rules for the three test cases, respectively), dissemination of the codified knowledge is potentially easier.

In discussing the extracted RIPPER rules with an independent domain expert, 3 distinct comments were made. Firstly, it was noted that having normal operation behaviour as the default case is intuitive, as the turbine operates in this state for a significant portion of its life. Secondly, favourable comments regarding the size of the rule base were put across; 12 rules were deemed to provide adequate encapsulation of the developed condition metric without being too cumbersome or providing little additional value. Finally, the human-readability and transparency of the rules was praised, with 8 of the 12 rules (66.67%) deemed highly intuitive and were immediately understandable. Given the simple nature of this technique, it is envisaged that more complex models which do not rely on physics of failure models will uncover novel degradation characteristics or behaviours, allowing for these characteristics to be exploited in the future.

Exploitation of this knowledge is straight forward. For instance, in table 6.5, rule 2 states:

> "**IF** the wind turbine rotor speed is greater than or equal to 17 RPM, **THEN** potential gearbox damage exists."

This is because limited instances in the training dataset existed where this was the case. Furthermore, given the strong encapsulation of the potential damage condition (over 98% accuracy) on the test set (which contained the failed gearbox), we can suggest that this alone presents a means for identifying degradation of the gearbox.

6.2.6 Conclusion

In this work, a novel robust methodology utilizing suspension and failure histories for extracting transparent, human-readable rules based on a condition index in any domain

was presented. For illustration and consistency throughout this thesis, the technique was applied for the health assessment of a wind turbine gearbox. A condition index was developed based upon the extended physics of failure model developed in section 4.3, which is employed in conjunction with a robustly defined Mahalanobis distance to rapidly label historical SCADA data allowing for rule extraction. Data from three independent turbines provided model attribute covariance and centres, with no failure data present in the training data. A historical failure is accurately identified within the test set, with strong signatures of condition present potentially as early as 5 months before catastrophic failure occurred. In addition, maintenance effectiveness is also reflected in the reduction of the RMD value after the gearbox has been replaced.

Utilizing outlier analysis performed in the χ^2 domain due to the non-normality of the multivariate model (as assessed through four independent multivariate normality tests) transparent rules were extracted to represent three condition levels: normal operation, inspection recommended and potential damage. Attribute selection by a domain expert and automated techniques (information gain) were found to be in agreement. Three models were built and assessed over three algorithms, with a mean accuracy of 96.11% (SD = 0.98%) attained by the RIPPER algorithm when validated against an independent validation set with data from a separate turbine. A mean $\kappa-$statistic of 94.17% (SD = 1.46%) is found for the RIPPER algorithm; showing strong agreement between the extracted model and the underlying condition index. Of the 12 extracted rules presented to an expert, 8 were found to be highly intuitive, representing known failure conditions.

A mean accuracy of 92.68% (SD = 6.93%) was attained by the C4.5 algorithm when validated against the three independent validation sets. A mean $\kappa-$statistic of 89.03% (SD = 10.40%) is found for the C4.5 algorithm; showing weaker accuracy and agreement between the extracted model and the underlying condition index than the RIPPER algorithm. Similarly, with regards to the decision table algorithm, a mean classification accuracy of 85.87% (SD = 10.96%) is attained with a mean $\kappa-$statistic of 78.80% (SD = 16.45%). As such, the decision table algorithm is the least suitable of these three techniques explored in this case for the purposes of codifying tacit knowledge.

Table 6.7: Descriptive statistics of test set 1 produced via RIPPER.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	0.983	0	1	0.983	0.992	0.996
Inspect	0.939	0	1	0.939	0.968	0.969
Damage	1	0.039	0.928	1	0.962	0.981
Weighted Avg.	0.982	0.974	0.013	0.976	0.974	0.974

Table 6.8: Confusion matrix (test set 1) produced via RIPPER.

		Predicted		
		Normal	Inspect	Damage
	Normal	2890	0	49
Actual	Inspect	0	2760	179
	Damage	0	1	2938

Table 6.9: Descriptive statistics of test set 2 produced via RIPPER.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	1	0.003	0.995	1	0.997	0.998
Inspect	0.928	0.001	0.997	0.928	0.961	0.968
Damage	0.992	0.036	0.932	0.992	0.961	0.984
Weighted Avg.	0.983	0.973	0.013	0.975	0.973	0.973

Table 6.10: Confusion matrix (test set 2) produced via RIPPER.

		Predicted		
		Normal	Inspect	Damage
	Normal	2938	0	1
Actual	Inspect	0	2613	326
	Damage	30	5	2904

Table 6.11: Descriptive statistics of test set 3 produced via RIPPER.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	0.989	0.061	0.89	0.989	0.937	0.967
Inspect	0.879	0.002	0.996	0.879	0.934	0.938
Damage	0.983	0.012	0.977	0.983	0.98	0.992
Weighted Avg.	0.966	0.95	0.025	0.954	0.95	0.95

Table 6.12: Confusion matrix (test set 3) produced via RIPPER.

		Predicted		
		Normal	Inspect	Damage
	Normal	3017	0	34
Actual	Inspect	311	2682	38
	Damage	40	11	3000

Table 6.13: Descriptive statistics of test set 1 produced via C4.5.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	0.994	0	1	0.994	0.997	0.995
Inspect	0.942	0	1	0.942	0.97	0.96
Damage	1	0.032	0.94	1	0.969	0.984
Weighted Avg.	0.98	0.978	0.011	0.98	0.978	0.978

Table 6.14: Confusion matrix (test set 1) produced via C4.5.

		Predicted		
		Normal	Inspect	Damage
Actual	Normal	2921	0	18
	Inspect	0	2768	171
	Damage	1	0	2938

Table 6.15: Descriptive statistics of test set 2 produced via C4.5.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	1	0.003	0.995	1	0.997	0.998
Inspect	0.928	0.001	0.997	0.928	0.961	0.968
Damage	0.992	0.036	0.932	0.992	0.961	0.984
Weighted Avg.	0.983	0.973	0.013	0.975	0.973	0.973

Table 6.16: Confusion matrix (test set 2) produced via C4.5.

		Predicted		
		Normal	Inspect	Damage
Actual	Normal	2939	0	0
	Inspect	0	2726	213
	Damage	16	7	2916

Table 6.17: Descriptive statistics of test set 3 produced via C4.5.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	0.996	0.216	0.697	0.996	0.821	0.976
Inspect	0.508	0.002	0.994	0.508	0.672	0.667
Damage	0.982	0.039	0.926	0.982	0.954	0.975
Weighted Avg.	0.873	0.829	0.086	0.872	0.829	0.815

Table 6.18: Confusion matrix (test set 3) produced via C4.5.

		Predicted		
		Normal	Inspect	Damage
Actual	Normal	3040	2	9
	Inspect	1273	1549	229
	Damage	46	8	2997

Table 6.19: Descriptive statistics of test set 1 produced via decision table.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	1	0.02	0.962	1	0.98	0.999
Inspect	0.925	0.002	0.997	0.925	0.959	0.971
Damage	0.983	0.025	0.951	0.983	0.967	0.988
Weighted Avg.	0.986	0.969	0.015	0.97	0.969	0.969

Table 6.20: Confusion matrix (test set 1) produced via decision table.

		Predicted		
		Normal	Inspect	Damage
	Normal	2938	0	1
Actual	Inspect	74	2718	147
	Damage	42	9	2888

Table 6.21: Descriptive statistics of test set 2 produced via decision table.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	0.991	0.119	0.806	0.991	0.889	0.99
Inspect	0.722	0.001	0.997	0.722	0.838	0.952
Damage	0.98	0.033	0.937	0.98	0.958	0.981
Weighted Avg.	0.974	0.898	0.051	0.913	0.898	0.895

Table 6.22: Confusion matrix (test set 2) produced via decision table.

		Predicted		
		Normal	Inspect	Damage
	Normal	2914	0	25
Actual	Inspect	647	2122	170
	Damage	53	6	2880

Table 6.23: Descriptive statistics of test set 3 produced via decision table.

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Normal	0.991	0.432	0.535	0.991	0.695	0.858
Inspect	0.186	0.002	0.976	0.186	0.312	0.762
Damage	0.951	0.002	0.995	0.951	0.972	0.999
Weighted Avg.	0.873	0.709	0.145	0.835	0.709	0.66

Table 6.24: Confusion matrix (test set 3) produced via decision table.

		Predicted		
		Normal	Inspect	Damage
	Normal	3025	13	13
Actual	Inspect	2484	566	1
	Damage	150	1	2900

6.3   Fault mode identification

The contents of this section appear in part in the following peer-reviewed publication:

- Godwin, J. L., Matthews, P. C. & Watson, C. (2014) "Robust Multivariate Statistical Ensembles for Bearing Fault Detection and Identification". In proceedings of the IEEE Prognostics and health management (PHM) conference. 22 - 24 June, Spokane, Washington State, USA

6.3.1   Introduction

Having identified a subset of time-domain features (the 'Moments model') for the purposes of condition assessment in section 4.2, and incorporated these into a robust metric which can be utilised to infer asset health in section 5.2, we can further refine the notion of distance and exploit the ideas presented so far throughout this thesis.

Up until this point, training of the multivariate centres has focused on the identification of normal operational behaviour. In practice, this enables more accurate and precise identification of current asset condition, which in turn enables proactive and predictive maintenance actions to be undertaken. This process has largely been an unsupervised process due to the majority class representing normal operational behaviour. In these instances, utilising a robust calculation of the covariance matrix will encapsulate normal behaviour due to the underlying distribution of the data.

However, it is possible to explore the notion of supervised training in order to encapsulate specific behaviours of interest. By encapsulating known behaviours, we can exploit the derived distance metric as a notion of similarity between current operational behaviour and a previously known condition. This work explores the notion of supervised learning with respect to the robust Mahalanobis distance for the purposes of fault diagnosis. In this case, the known behaviours explored are bearing failure modes. This enables the use of distance metrics to explore the distance (or similarity) between current operational behaviour and previous known failures.

As such, utilising known failure data to determine attribute centres and covariance based upon the moments model developed in section 4.2 will enable a distance metric to encapsulate this behaviour. Thus, by exploring the exploitation of multiple multivariate distance metrics, which have each been trained to encapsulate different failure modes (or normal operational behaviour), we can potentially perform fault diagnosis. Thus, exploiting ensembles of multivariate distance metrics for the purposes of fault diagnosis potentially enables more information regarding the underlying asset condition to be presented to maintenance managers or decision makers in order to choose the most beneficial maintenance action to perform.

6.3.2   Bearing failure modes

In order to explore the notion of ensembles of techniques for the purposes of fault diagnosis, it is essential that the failure modes of a bearing are understood. As such, four common bearing failures are of interest and are discussed here. Although many failure mechanisms exist for bearings (Evans, 2011), for the purposes of this work, only surface fatigue failures are considered; namely, pitting and spalling. Pitting is the formation of surface defects through fatigue whereas spalling is characterised by deep cracking of the component due to fatigue of sub-surface defects.

We are only interested in these surface fatigue behaviours due to the data available for this analysis. Given the large number of failures modes, and varying degrees of severities of failures across the four components within a bearing (inner race, outer race, roller elements and cage), it would not be practical to perform run-to-failure testing until adequate sample sizes of each degradation mode across each severity across each component were collected.

Inner race fault

Surface fatigue of the inner race, as shown in figure 6.8 occurs when the inner race of the bearing (the race which contacts the rotating axis) degrades. Early identification of inner race faults enables only the inner race component to be replaced as degradation will not have spread to the cage or ball (roller elements). The inner race frequency is:

$$BPFI(Hz) = S\frac{N}{2}\left(1 + \frac{B}{P}\cos\phi\right) \tag{6.2}$$

Figure 6.8: An example bearing race suffering from surface fatigue of the inner race Stevens (2011).

Outer race fault

Surface fatigue of the outer race, as shown in figure 6.9 occurs when the outer race of the bearing (the stationary race which contacts the balls or roller elements) degrades. Early identification of outer race faults enables the preventive scheduling to replace the outer race, or complete bearing if necessary. The outer race frequency is:

$$BPFO(Hz) = S\frac{N}{2}\left(1 - \frac{B}{P}\cos\phi\right) \tag{6.3}$$
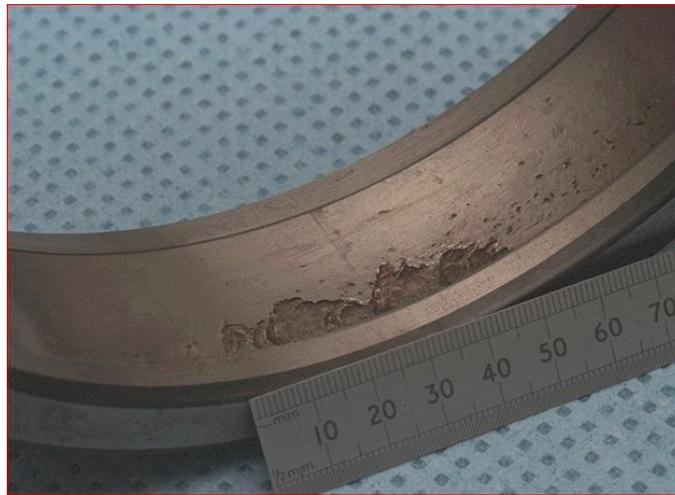
Figure 6.9: An example bearing race suffering from surface fatigue of the outer race Stevens (2011).

Roller element fault

Surface fatigue of the roller elements (or ball), as shown in figure 6.10 occurs when the roller elements of the bearing (the rotating elements held in place between the races which contact both the inner and outer races) degrades. Early identification of roller element faults enables defective elements to replaced, whilst saving the rest of the bearing. The roller element (ball) frequency is:

$$BSF(Hz) = S\frac{P}{2B}\left(1 - \frac{B^2}{P^2}\cos^2\phi\right) \tag{6.4}$$

Figure 6.10: An example roller element (ball) suffering from surface fatigue Stevens (2011).

Cage fault

Surface fatigue of the roller element cage, as shown in figure 6.11 occurs when the cage which holds the roller elements in place degrades. Cage faults are not common and are illustrated here for completeness. Early identification of cage faults enables a replacement cage to be fitted whilst minimising damage to the roller elements and races. The cage frequency is:

$$FTF(Hz) = S\frac{1}{2}\left(1 - \frac{B}{P}\cos\phi\right) \tag{6.5}$$

6.3.3   Dataset employed

In order to encapsulate the various failure modes of bearings, fault data was necessary. As the use of multivariate distance metrics are employed to define these behaviours, these metrics must be trained to identify the failure modes of interest.

Due to this, the bearing dataset made available through the Case Western Reserve University (2011) was employed. This dataset contains high frequency bearing data which has been seeded with faults through electro-discharge machining. Full details of

Figure 6.11: An example bearing cage suffering from degradation due to wear Stevens (2011)

this dataset are available in section 3.4.3. In total, data from 4 bearings was employed, with each pertaining to an operational mode of interest. Faults of size .0007 inches are employed for this analysis. In this case, the modes of operation are:

- Normal operational behaviour.

- Inner race fault.

- Outer race fault.

- Roller element (ball) fault.

In total, 10 seconds of data (sampled at 48 kHz) were taken from each operational mode and processed using the bearing moments model created in section 4.2. The 4-dimensional model thus contained 40 data points (1 data point per second), with 10 data points pertaining to each operational mode of the bearing.

6.3.4 Encapsulation of failure behaviour

To demonstrate the validity of the model, traditional univariate statistical analytical techniques were employed to elaborate on the variance of each of the attributes for each fault.

The attributes in the model should differ significantly between the various faults so that this knowledge can be exploited for both fault identification and fault diagnosis. Each attribute was analysed individually to determine their ability to discriminate differing fault modes.

It should be noted that whilst skewness, kurtosis, RMS and the standard deviation have previously been utilised in the analysis of high frequency bearing data, there is limited application of these techniques within the multivariate domain. Although

individually each feature potentially enables early detection and diagnosis, without the additional context provided within the multivariate domain, faults may not readily be identified. For instance, skewness and kurtosis can remain stable even though the underlying data distribution changes due to possible degradation. Similarly, the signal RMS and standard deviation can remain constant whilst the skewness and kurtosis of the signal changes significantly. As such, by performing analysis in the multivariate domain, we can mitigate these flaws by providing the additional context of the other features employed.

Statistical analysis of these features was performed independently to determine the capabilities of each feature in terms of both fault detection and also fault identification (diagnosis). If these features cannot detect deviations from normal operational behaviour or accurately discriminate between fault modes in the univariate domain, no diagnostic power is achieved. As such, ensuring these features are capable of performing fault detection and identification must be performed.

Skewness

The 10 data points representing normal operation behaviour had an average skewness of -0.17 ($SD = 0.02$), the 30 data points representing the inner race fault, outer race fault and roller element faults had an average skewness of -0.9 ($SD = 0.01$), 0.08 ($SD = 0.01$) and 0.01 ($SD = 0.02$) (respectively). One way analysis of variance (ANOVA) showed that the skewness varied significantly between failure modes, $F(3, 36) = 510.67$, $p < .01$. The samples are known to be independent and the assumption of normality was held as assessed by the Shapiro-Wilk test; given that $p = .36$ for the inner race fault, $p = .41$ for the outer race fault, $p = .07$ for the normal condition and $p = .13$ for the roller element fault, we can conclude the various faults identified by the signal skewness are normally distributed. Homogeneity of variance was tested by employing Levenes' F statistic, and found to be significant $F(3, 36) = 4.15$, $p < .05$, indicating the assumption of equal variances does not hold. As such, an adjusted F statistic is employed. In this case, Welchs' F statistic is employed, and found to be significant $F(3, 17.58) = 777.15$, $p < .01$. For completeness, it should be mentioned that the Brown-Forsythe F statistic was also significant ($F(3, 26.97) = 510.68$, $p < .01$).

As the analysis of variance only makes it known that the various faults differ significantly, post-hoc analysis was performed in order to determine significant differences of signal skewness between fault types. This analysis was carried out as per the Games-Howell procedure due to the assumption of homogeneity of variance not holding for this attribute in particular. It was found that the normal operational behaviour of the bearing had significantly different skewness from that of an inner race fault (-0.74, $p < .01$),

an outer race fault (-0.24, $p < .01$) and a roller element fault (-0.17, $p < .01$). As such, the skewness was deemed to be beneficial in assisting bearing fault identification as it is able to effectively discriminate between between normal operational behaviour and each of the varying fault types. In order to assess the quality of signal skewness as a diagnostic tool, the three further comparisons which were required to distinguish the fault mode were performed. Statistically significant differences in skewness were found between the inner race fault and outer race fault (-0.17, $p < .01$), the inner race fault and roller element fault (-0.10, $p < .01$) and also between the outer race fault and roller element fault (0.07, $p < .01$); showing the benefit of employing skewness in the diagnostic process.

Kurtosis

Following a similar analysis to that which was performed with regards to skewness, kurtosis was also analysed. In this case, the 10 data points representing normal operation behaviour had an average kurtosis of 2.92 ($SD = 0.03$), the 30 data points representing the inner race fault, outer race fault and roller element faults had an average kurtosis of 7.91 ($SD = 0.09$), 6.94 ($SD = 0.09$) and 3.08 ($SD = 0.09$), respectively. One way analysis of variance showed that the kurtosis varied significantly between failure modes, $F(3, 36) = 9522.948$, $p < .01$. The samples are known to be independent and the assumption of normality was held as assessed by the Shapiro-Wilk test; given that $p = .81$ for the inner race fault, $p = .27$ for the outer race fault, $p = .05$ for the normal condition and $p = .29$ for the roller element fault, we can conclude the various faults identified by the kurtosis are normally distributed. Homogeneity of variance was tested by employing Levenes' F statistic, and found not to be significant $F(3, 36) = 2.14$, $p > .05$, indicating the assumption of equal variances is held. As such, neither an adjusted F statistic (as was required previously) or the Kruskal-Wallis test is required in this case.

As done previously, post-hoc analysis was performed in order to asses the significant differences in kurtosis between the different fault types present on the bearing. This analysis was carried out as per Tukeys' procedure, as unlike previously, the assumption of homogeneity of variance held. Normal operational behaviour of the bearing was found to be significantly different in terms of kurtosis from inner race faults (-4.59, $p < .01$), outer race faults (-4.01, $p < .01$) and roller element faults (-1.56, $p < .01$). As such, kurtosis can be demonstrated to be a useful tool in fault detection for these modes on a bearing. In order to determine the quality of kurtosis for fault identification, significant difference should be found between each fault mode. Significant differences in kurtosis were found between inner race faults and outer race faults (0.58, $p < .01$),

inner race faults and roller element faults (4.44, $p < .01$) and also outer race faults and roller element faults (3.86, $p < .01$). Thus, similarly to skewness, kurtosis can be effectively employed for both fault detection and identification.

Root Mean Square (RMS)

Next, the RMS was analysed for its effectiveness at both fault detection and identification. The 10 data points representing normal operation behaviour had an average signal RMS of 0.64 ($SD = 0.002$), the 30 data points representing the inner race fault, outer race fault and roller element faults had an average RMS of 0.28 ($SD = 0.001$), 1.02 ($SD = 0.004$) and 0.14 ($SD = 0.003$), respectively. One way analysis of variance showed that the signal RMS varied significantly between bearing faults, $F(3, 36) = 193858.90$, $p < .01$. The samples are known to be independent and the assumption of normality was held as assessed by the Shapiro-Wilk test; given that $p = .32$ for the inner race fault, $p = .44$ for the outer race fault, $p = .78$ for the normal condition and $p = .94$ for the roller element fault, we can conclude the various fault types described by the RMS are normally distributed. Homogeneity of variance was tested by employing Levenes' F statistic, and found to be significant $F(3, 36) = 7.14$, $p < .05$, indicating the assumption of equal variances does not hold in this case. As such, an adjusted F statistic is employed. In this case, Welchs' F statistic is employed, and found to be significant $F(3, 15.46) = 214685.56$, $p < .01$. For completeness, it should be mentioned that the Brown-Forsythe F statistic was also significant ($F(3, 18.40) = 193858.90$, $p < .01$).

Post-hoc analysis was performed in order to determine significant differences of RMS between fault types. This analysis was carried out as per the Games-Howell procedure due to the assumption of homogeneity of variance did not hold for the signal RMS. Again, as per both skewness and kurtosis, statistically significant differences between normal operational behaviour in terms of the mean RMS were found with inner race faults (-0.22, $p < .01$), outer race faults (-0.96, $p < .01$) and roller element faults (-0.8, $p < .01$). This shows the capability of the signal RMS for fault detection and as such justifies its incorporation into the model. In order to assess the performance of the RMS for fault identification, three further post-hoc comparisons were made. The signal RMS was found to be significantly different between inner race faults and outer race faults (-0.74, $p < .01$), inner race faults and roller element faults (0.14, $p < .01$) and also outer race faults and roller element faults (0.88, $p < .01$).

Standard Deviation

Finally, the standard deviation of the signal was analysed for its effectiveness at both fault detection and identification. The data points representing normal operation behaviour had an average standard deviation of 0.63 ($SD < 0.001$), the 30 data points representing the inner race fault, outer race fault and roller element faults had an average standard deviation of 0.28 ($SD < 0.002$), 1.02 ($SD < 0.005$) and 0.14 ($SD < 0.004$) respectively. The samples are known to be independent, and the assumption of normality was held for each group as assessed by the Shapiro-Wilk test; given that $p = .08$ for the normal condition, $p = .32$ for the inner race fault, $p = .45$ for the outer race fault, and $p = .93$ for the roller element fault. As such, it is possible to perform traditional analysis of variance in this case without the assistance of a modified F statistic. Thus, we can conclude that the various fault types described by the standard deviation are normally distributed. Homogeneity of variance was tested by employing Levenes' F statistic, and found to be significant $F(3, 36) = 6.75$, $p < .001$, indicating the assumption of equal variances does not hold. As such, an adjusted F statistic is once again employed. In this case, Welchs' F statistic is employed, and found to be significant $F(3, 15.78) = 210223.49$, $p < .001$. For completeness, it should be mentioned that the Brown-Forsythe F statistic was also significant ($F(3, 18.40) = 194254.99$, $p < .001$).

Post-hoc analysis was performed in order to determine significant differences of standard deviation between fault types. This analysis was carried out as per the Games-Howell procedure due to the assumption of homogeneity of variance not holding for the standard deviation. Again, as per both skewness and kurtosis, statistically significant differences between normal operational behaviour in terms of the mean standard deviation were found with inner race faults (-0.22, $p < .001$), outer race faults (-0.96, $p < .001$) and roller element faults (-0.80, $p < .001$). This shows the capability of the standard deviation for fault detection and as such justifies the incorporation of this feature into the model. In order to assess the performance of the standard deviation for fault identification (diagnosis), three further post-hoc comparisons were made. The standard deviation was found to be significantly different between inner race faults and outer race faults (-0.74, $p < .001$), inner race faults and roller element faults (0.14, $p < .001$) and also outer race faults and roller element faults (0.88, $p < .001$).

6.3.5 Ensemble diagnosis

Due to the availability of high frequency bearing data with seeded faults, the accurate encapsulation of the four bearing conditions identified in the informal FMEA was possible. As we can determine the multivariate centres and attribute covariances for each

fault mode, the possibility arises for an ensemble of condition metrics to be employed to quantify current bearing condition and provide insights for both fault detection and identification. Although these factors will be application, bearing and load specific, it provides a means to validate the fault identification approach established.

Since the robust Mahalanobis distance quantifies the deviations from the vector of defined centres - $\hat{\mu}$ - whilst taking into account the attribute covariance - Σ - smaller values represent a greater degree of similarity to the data which derived the centres and covariance. As such, we can specify the current likely behaviour of the bearing by initially defining 4 health metrics:

$$RMD_{Norm} = \sqrt{(\mathbf{X} - \hat{\boldsymbol{\mu}}_{Norm})^T \hat{\Sigma}_{Norm}^{-1} (\mathbf{X} - \hat{\boldsymbol{\mu}}_{Norm})} \tag{6.6}$$

$$RMD_{Inner} = \sqrt{(\mathbf{X} - \hat{\boldsymbol{\mu}}_{Inner})^T \hat{\Sigma}_{Inner}^{-1} (\mathbf{X} - \hat{\boldsymbol{\mu}}_{Inner})} \tag{6.7}$$

$$RMD_{Outer} = \sqrt{(\mathbf{X} - \hat{\boldsymbol{\mu}}_{Outer})^T \hat{\Sigma}_{Outer}^{-1} (\mathbf{X} - \hat{\boldsymbol{\mu}}_{Outer})} \tag{6.8}$$

$$RMD_{Ball} = \sqrt{(\mathbf{X} - \hat{\boldsymbol{\mu}}_{Ball})^T \hat{\Sigma}_{Ball}^{-1} (\mathbf{X} - \hat{\boldsymbol{\mu}}_{Ball})} \tag{6.9}$$

With RMD_{Norm} representing normal operational behaviour of the bearing (that is, the covariance and means determined robustly through data with normal behaviour), RMD_{Inner} representing behaviour of the inner race fault, RMD_{Outer} representing the behaviour of the outer race fault and finally RMD_{Ball} representing the roller element fault behaviour. Given the 4-tuple $\boldsymbol{\varrho} = \{RMD_{Norm}, RMD_{Inner}, RMD_{Outer}, RMD_{Ball}\}$ we have:

$$Bearing(\mathbf{X}) =$$

$$\begin{cases} Normal & if \min(\{\boldsymbol{\varrho}\}) = RMD_{Norm} \\ Inner\,Race\,Fault & if \min(\{\boldsymbol{\varrho}\}) = RMD_{Inner} \\ Outer\,Race\,Fault & if \min(\{\boldsymbol{\varrho}\}) = RMD_{Outer} \\ Roller\,Element\,Fault & if \min(\{\boldsymbol{\varrho}\}) = RMD_{Ball} \end{cases} \tag{6.10}$$

6.3.6   Result and discussion

It is essential to verify and validate that the developed bearing condition moments model and developed expert diagnostic system accurately detect and identify the seeded faults in the initial bearing dataset. Figure 6.12 shows the distances gained when
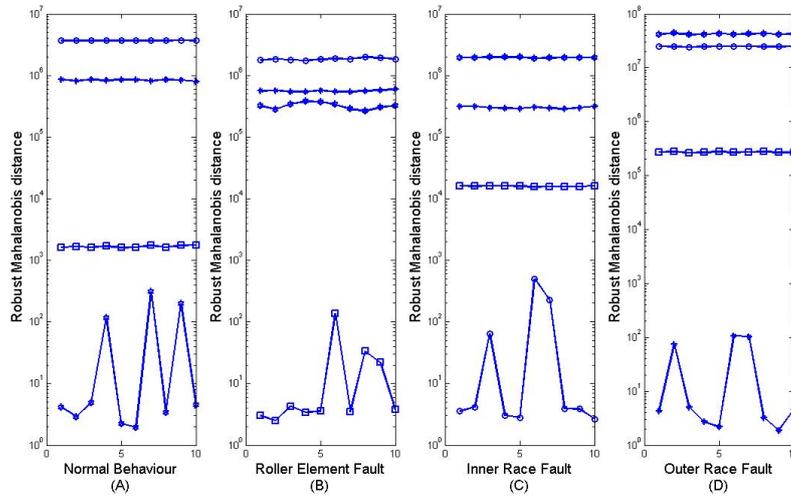
Figure 6.12: An ensemble of Robust Mahalanobis distances trained on normal bearing data (A), roller element fault data (B), inner race fault data (C) and outer race fault data (D). Normal behaviour is represented by $\star$, roller element behaviour by $\square$, inner race behaviour by $\circ$ and outer race behaviour by $\diamond$.

comparing different fault types trained on the seeded fault data. It can be seen that in each of the 10 data points for the fault in question, the Robust Mahalanobis distance accurately detects the fault and performs correct identification. This represents a 100% fault detection and identification rate on the initial validation dataset, thus validating the approach. It is interesting to note that there are orders of magnitude between the various faults; an artefact present due to careful selection of attributes which were statistically relevant for this task, in conjunction with a highly sensitive distance metric.

In order to demonstrate the alternate case where an inferior distance metric and feature set are employed, a set of inferior features were utilised as a bearing condition model. In this case, a 4-dimensional model made up of:

- The mean.

- The mode.

- The maximum.

- The peak to average power ratio (PAPR).

The traditional Mahalanobis distance was then utilised to determine the bearing condition index (utilising the traditional covariance calculation). The results of this can be seen in figure 6.13. As can be seen, although some diagnosis is possible (i.e it
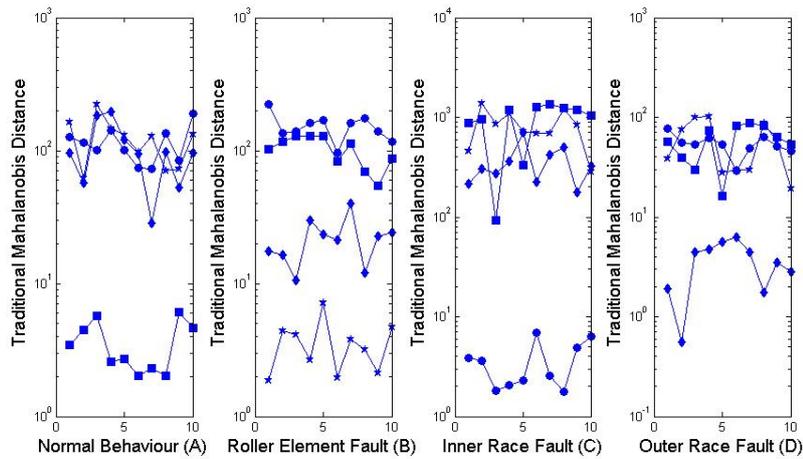
Figure 6.13: An ensemble of traditional Mahalanobis distances trained on normal bearing data using inferior features showing (A), roller element fault data (B), inner race fault data (C) and outer race fault data (D). Normal behaviour is represented by □, roller element behaviour by ⋆, inner race behaviour by ∘ and outer race behaviour by ◇.

is possible to know if we do not have normal operation, it is possible to distinguish between roller element faults and inner race faults, but not outer race faults).

Thus, by utilising a superior set of features in conjunction with a superior distance function, accurate diagnosis is possible in this case. However, in order to validate this approach, the technique must be shown to be practical on an independent dataset.

NASA Dataset

Figure 6.14 shows both the normal behaviour index and the outer race index as applied to independent dataset for the purposes of validating the approach. In this case, the dataset was supplied from the NASA bearing dataset (Lee et al., 2007), freely available through their prognostics centre of excellence. Each data point represents a single second snapshot of the bearing. The data is pre-processed with the moments model described in section 4.2, from 20 kHz to 1 Hz. As each of the Robust Mahalanobis distances encapsulate these behaviours, the minimal value represents the current operating mode. The centres utilised for figure 6.14 are based upon the robust centres derived from each of the 10 samples from the Case Western Reserve University (2011) dataset employed in section 6.3.6 for each of the four operational modes of interest. In this case, the failed bearing (bearing 1) from the 2nd test of the NASA dataset was employed for validation. This bearing suffered an outer race fault at the end of its useful life.
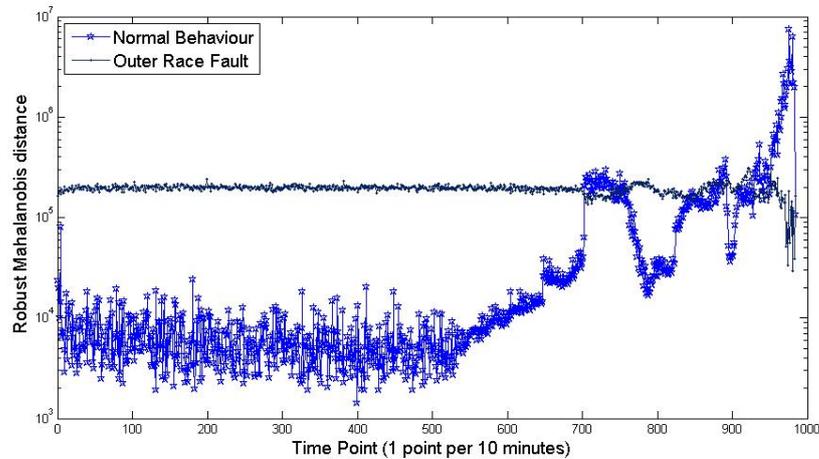
Figure 6.14: Validation of ensemble distances for diagnosis on the NASA bearing dataset (Lee et al., 2007).

Upon initial inspection, it can be seen in figure 6.14 that normal operational behaviour is the prevalent mode of operation for the majority of the life of the bearing. This it to be expected as typically the bearing will not be subjected to a fault. It can be seen that at approximately data point 700, some form of shock is presented to the system, and as such, the outer race fault briefly becomes more likely to have occurred than normal behaviour. After this point, although the bearing recovers slightly, after time step 900, there is no doubt to the existence of an outer race fault. This provides confidence in the system, as the fault is identified on an independent system.

However, once the metrics for both the roller element (ball) fault and the inner race fault are also added into the analysis, issues occur with this approach. As can be seen in figure 6.15, no longer is the most recognised failure mode that of the inner race fault. Furthermore, the bearing is deemed to have a roller element (ball) fault rather than operating normally. This false positive ultimately reduces the accuracy of the resulting diagnosis.

The poor quality diagnosis performed in this case can be attributed to many reasons. It is likely that these are due to the significant and substantial differences between the two datasets employed. For instance, these differences include (but are not limited to):

- The data collection frequency.

- The bearing RPM.

- The bearing loading conditions.

- The size of the fault.
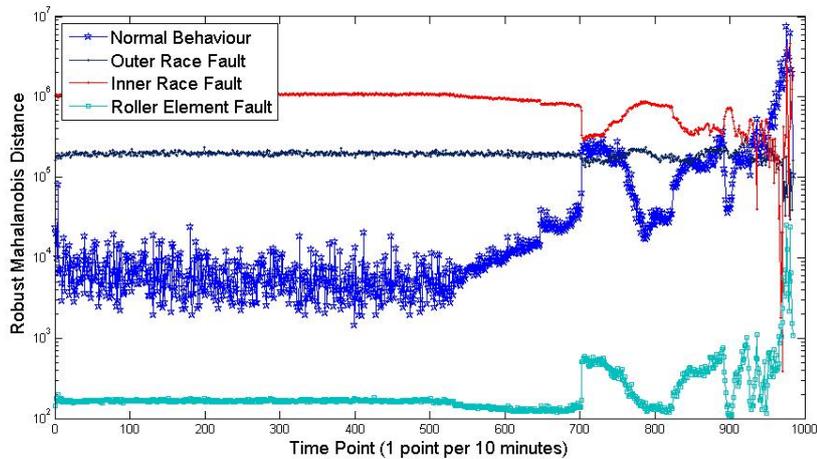
Figure 6.15: Additional validation of ensemble distances for diagnosis on the NASA bearing dataset (Lee et al., 2007) with encapsulation from the (Case Western Reserve University, 2011) dataset.

- The type of bearing employed.

Firstly, the frequency of data collection is different. The dataset used for verification was recorded at 48 kHz, with the validation dataset (NASA bearing data) recorded at 20 kHz. This disparity should not cause many issues due to use of statistical features to encapsulate behaviour, but this should be noted.

Secondly, the RPM of the drive shaft in both cases was different; the verification dataset (Case Western Reserve University, 2011) was collected at 1750 RPM, with the NASA bearing data (Lee et al., 2007) recorded at 2000 RPM. This is thought to introduce some noise into the system as no normalisation has yet been performed to mitigate this problem.

Next, the test set-up was different for the validation set and the verification set. The NASA data had a 6000 lbs radial load applied; it is likely that this will effect the data and must be normalised for in practical applications.

Following this, fault size is not taken into account with the system trained only on 0.007 inch faults; their effect on the statistical features in the model should be assessed and normalised for.

Both datasets employed different bearings for the analysis. As such, slight differences in the manufacture and use of these bearings will alter the results of this approach. Normalisation of some features for bearing conditions would mitigate this problem.

Finally, different bearings will inherently have different characteristics (dependent upon size, type, application domain, etc.) and as such, this should also be taken into

consideration.

Due to these differences, faults were encapsulated utilising the NASA bearing data. In this case, only 1 instance of each of the inner race and roller element faults were present with 2 outer race faults available for analysis. As such, bearing 1 from the 2nd test set was employed to validate the technique. In order to encapsulate the faults (and normal behaviour), failure and suspension histories from other bearings within the NASA dataset were used, namely:

- **Test 1, Bearing 3**: Inner race fault (failure history).

- **Test 1, Bearing 4:** Roller element fault (failure history).

- **Test 3, Bearing 3:** Outer race fault (failure history).

- **Test 2, Bearing 2:** Normal behaviour (suspension history).

Retrospective analysis of these bearings was performed in the univariate domain based upon the 4 time domain statistical features utilised in the multivariate moments model developed in section 4.2. For inner race, outer race and roller element bearing faults, once failures were deemed to have started to develop (by a distinct change in the underlying statistical metric), the following 155 data points were used to encapsulate failure behaviour. It is due to small quantities of failure data that 155 data points were utilised (as degradation occurs prior to failure). For normal operational behaviour, 10% to 40% of bearing life was taken from the healthy bearing (Test 2, bearing 1) and these samples were utilised to encapsulate behaviour.

In order to assess the quality of these encapsulations, ensembles of robust Mahalanobis distances were utilised to train the metrics as before. The results of this can be seen in figure 6.16. Although strong encapsulation of failure modes is achieved (100% diagnostic accuracy), this does not translate in practice to accurate diagnosis in this case.

As can be seen in figure 6.17, although an improvement is made over the results achieved in figure 6.15, diagnosis is still not possible in this case. Fault detection, however, is possible in this case due to the benefits gained with regards to utilising the NASA bearing failures for the purposes of encapsulation of failure behaviour. It is likely that diagnosis in this case is not possible due to the position of bearings employed in the failure encapsulation. As a radial load is applied, it is possible that this will effect different bearing positions in different ways, and as such, may potentially explain this weak result.

6.3.7 Conclusion

In this work, a novel application employing ensembles of robust multivariate distance metrics for the real-time analysis of the condition of a bearing is presented, along with an expert system to automatically determine the most likely current behaviour of the bearing. Diagnostic signatures for each statistical feature (namely the standard deviation, root mean square, skewness and kurtosis) were analysed and found to be significant in discriminating between normal behavioural conditions and three distinct surface fatigue failure modes applicable to bearings; inner race faults, outer race faults and roller element faults.

A robustly derived Mahalanobis distance metric was used to determine multivariate centres and attribute covariance based upon fault data seeded through electro-discharge machining from a publicly available dataset made available by Case Western Reserve University (2011). The 48 kHz data had the relevant statistical features extracted in order to encapsulate the current behaviour of the bearing. This snapshot enabled the use of the robust Mahalanobis distance (RMD) to explicitly quantify the current operational state of the bearing without the need for a data-driven prognostic tool. By employing this technique, larger models consisting of many variables would not hinder analysis as the problem is reduced to the univariate domain and is computationally tractable even for highly dimensional data. However, this is dependent upon
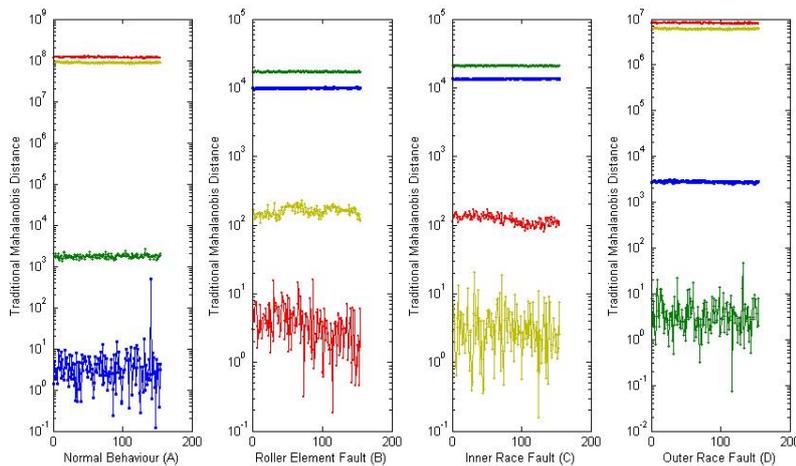
Figure 6.16: An ensemble of traditional Mahalanobis distances trained on normal bearing data using inferior features showing (A), roller element fault data (B), inner race fault data (C) and outer race fault data (D). Normal behaviour (blue) is represented by □, roller element behaviour (red) by ⋆, inner race behaviour (yellow) by ○ and outer race behaviour (green) by ⋄.
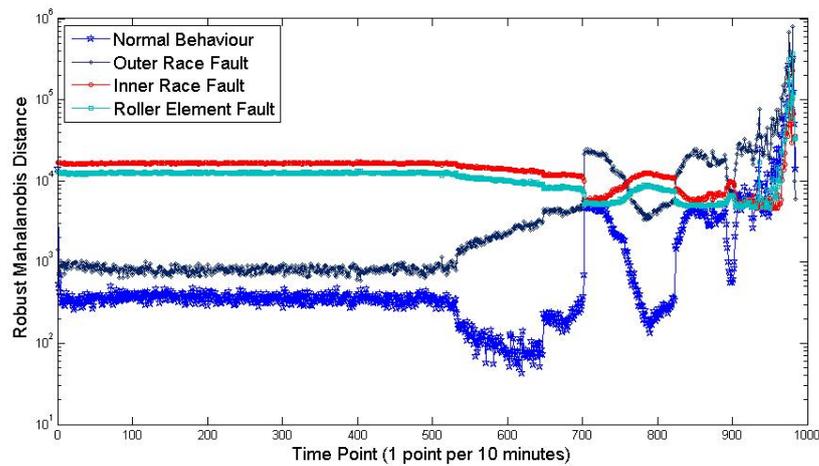
Figure 6.17: Additional validation of ensemble distances for diagnosis on the NASA bearing dataset (Lee et al., 2007) based upon NASA dataset failure encapsulation.

the technique employed for estimating the MCD subset, as explored in section 5.3.

The approach is verified, showing 100% success of both fault detection and fault identification on the Case Western Reserve University (2011) dataset. Following this, the technique was verified on an independent dataset (the NASA bearing data set) made available by Lee et al. (2007). Initially, centres encapsulating failures from the seeded fault dataset (Case Western Reserve University, 2011) are employed, but shown to not work well in practice, potentially due to differences in the data acquisition process and test setup. Failures from the NASA dataset were then employed to encapsulate failure mechanisms, but also shown to have less than desirable results.

Although limitations exist, a proof of concept has been shown for the possibility of utilising ensembles of robust multivariate distance metrics for the purposes of diagnosis. Thus, once refined, this technique would allow for significant cost savings to be created by maintenance teams by increasing overall bearing reliability, minimising spare inventory held, and optimizing both the maintenance planning and maintenance scheduling processes. Optimising is possible as the quantification of degradation on the bearing allows for prioritisation of maintenance activities, and scheduling of these activities at periods of minimal usage.

The main strengths of this approach lie in the transparent nature of the statistical methods employed. No data intensive calculations are required and no 'black-box' systems which are unaccountable and unauditable are used. Domain knowledge is available for incorporation into the analysis through simple FMEA. No normalisation of the data is required, assisting in the implementation of the technique in practice.

No additional sensors were required for the data collection, and as such, no further capital expenditure was required. Storage costs were minimised as the need to store high frequency data was mitigated by utilising time-domain statistical features.

6.4 Bearing remaining useful life (RUL) estimation

6.4.1 Introduction

Having explored the predictive power of artificial neural networks in section 2.5.7, it is possible to extend this analysis in order to gain further business intelligence through the estimation of remaining useful life (RUL).

It should be noted that whilst condition prediction is a form of prognosis, this alone is potentially not enough to make effective business decisions. As by their nature distance metrics are unbounded, each application of this technique will have varying operating limits. Although the parameterised $F-$distribution (Hardin and Rocke, 2005) effectively enables bounds to be placed, this can only be performed when the data is multivariate normal. As such, in cases when this does not hold (which is typically often in practice, as demonstrated in section 6.3), an alternative bounding is required.

This work presents a methodology to estimate the remaining useful life of components based upon failure histories. Remaining useful life estimation (or prediction) enables more effective maintenance decisions to be made than relying on distance metrics alone. Although a distance metric may exceed a specified threshold, this may not necessarily be indicative of an immediate failure. By utilising past failure histories, it is possible to perform a mapping from the condition of the component (the health index derived from the robust distance metric) to a remaining useful life.

The ultimate goal of any prognostics system is to provide an accurate remaining useful life estimation or prediction based upon component specific information. This is referred to as type III prognostics (also known as effects-based prognostics), and moves away from type I (traditional, population based reliability analysis) and type II (stress based, typically moderation of traditional reliability data taking environmental stresses into consideration) in order to provide a more accurate estimation of life for a given component of interest (Coble and Hines, 2008).

By mapping the current operational behaviour of the component to a remaining useful life, more knowledge is provided to an organisation to assist in the scheduling and planning of maintenance. It is easier to identify maintenance opportunities and to plan maintenance activities when insight into the remaining useful life of an asset is available. Thus, accurate estimation of remaining useful life is essential for any proactive or predictive maintenance strategy and also in the optimisation of current maintenance resources.

6.4.2   Remaining useful life estimation

Dataset employed

Due to the supervised nature of artificial neural networks, training is required to take place on failure histories from which the remaining useful life of a component is known.

Within the Reliawind database (Wilkinson et al., 2010), only a single catastrophic gearbox failure is available, this cannot be employed for the RUL estimation. An artificial neural network would simply over fit the data to achieve a high predictive accuracy of remaining useful life as no validation can be performed. As such, the NASA bearing dataset (Lee et al., 2007) is employed due to multiple failure histories

Within the NASA bearing dataset, 4 failures of covering 3 failure modes exist. Within this dataset, three distinct sets of data are employed, namely:

- ***Independent normal operational behaviour:*** In order to determine the multivariate centres and covariance, an independent bearing (which did not fail) is employed. This was Test 3, dataset 1.

- ***Training data:*** In order to train the network, three bearing failure histories were employed so that the neural network could approximate the remaining useful life function of the set of bearings. These were bearings:

    - *Training bearing 1:* Test 2, Bearing dataset 1.
    - *Training bearing 2:* Test 1, Bearing dataset 8.
    - *Training bearing 3:* Test 1, Bearing dataset 6.

- ***Test data:*** In order to ensure that no over-fitting is taking place and to verify the methodology, two further independent bearings are employed. These are:

    - *Validation bearing 1:* Test 2, Bearing dataset 2.
    - *Validation bearing 2:* Test 1, Bearing dataset 5.

The bearing moments model identified in section 4.2 as utilised in chapter 5 was again employed as the principal features of the condition metric based upon the robust Mahalanobis distance. These features are:

- RMS.

- Standard deviation.

- Skewness.

- Kurtosis.

Methodology

In this analysis, we again employ a back propagation feed forward neural network with Levenberg-Marquardt training to predict bearing remaining useful life. The network structure employed for this analysis consisted of three layers; an input layer, a single hidden layer and an output layer. The number of neurons in the hidden layer were determined by Huang (2003):

$$Z_l = 2\sqrt{(r+2)Z_m} \qquad (6.11)$$

The hyperbolic tangent sigmoid transfer function is employed within the hidden layer, with the linear transfer function employed in the output layer. A single neuron was employed in the output layer to represent our predicted value. In order to assess the impact of context, the inputs to the neural network were moderated to allow a sensitivity analysis to take place.

Given a vector of length t representing bearing behaviour $\mathbf{X}$ which is defined as $\mathbf{X} = (RMD_1, RMD_2, \cdots, RMD_t)$, we have the output vector $O^i = \langle RMD_{t+i} \rangle$, and input vector $\mathbf{I^j} = (RMD_t, RMD_{t-1}, \cdots, RMD_{t-j})$.

Due to the use of the hyperbolic tangent sigmoid transfer function employed in the neural network requiring an input between -1 and 1, the robust Mahalanobis distance must be normalised so that all values fall within this range. Both min-max normalisation and z-score normalisation (Kumar and Bhandare, 2011) are unsuitable in these cases.

Min-max normalisation will not work as the robust Mahalanobis function is not bounded within a predefined range; an ill-defined maximum value for scaling could lead to poor condition prediction. Similarly, z-score normalisation will not be effective due to degraded behaviour often falling many deviations above normal behaviour, as such value will be outside the range of those anticipated by the transfer function. Due to the exponential nature of the robust Mahalanobis metric function, a scaled logarithmic transform is employed:

$$\widehat{RMD}_i = \frac{log_{10}(RMD_i)}{log_{10}(c)} \qquad (6.12)$$

Where $\widehat{RMD}_i$ represents the robust Mahalanobis metric for the current observation, and c is an application specific constant which represents the maximal level of degradation permitted within the system before failure has deemed to occur. In this case, $c = 100,000$ and was chosen based upon retrospective analysis of the condition time series. Given the exponential nature of the robust distance function and degradation

phenomena in general, this is the natural transform to employ.

As the remaining useful life of the bearing can be thought of as a function of both time and condition, the input vector can be modified to incorporate this additional feature. This would then be defined as $\mathbf{I^{j,v}} = \left(\hat{T}_t, VAR_v, \hat{RMD}_t, \hat{RMD}_{t-1}, \cdots, \hat{RMD}_{t-j} \right)$. Where VAR_v represents the variance of the previous v observations and $\hat{T}_t$ represents the current elapsed life of this bearing at time t linearly scaled to a percentage of the maximum known life of any similar bearing, for instance:

$$\hat{T}_i = \frac{T_i}{T_{Max}} \tag{6.13}$$

Where T_i is the current life (in elapsed minutes, revolutions, or similar) and T_{Max} is the maximum recorded (or estimated maximum) life of this particular bearing. This will be application specific, however, in this case T_{Max} was determined to be two weeks. The neural network output was $O^t = \langle \overline{RUL_t} \rangle$.

Data from the NASA bearing dataset was employed for this analysis (Lee et al., 2007). The bearing moments model derived in section 4.2 was employed and attribute location and covariance were determined from an independent bearing in the dataset (test 3, dataset 1). These parameters were then used to create both a training set and test set to verify and validate the methodology. Due to the supervised nature of the neural network, only failed bearings could be employed to train the network in this work.

It should be noted that techniques to employ suspension histories in this training process exist. For instance as in Heng (2009), Tian et al. (2010) andWidodo and Yang (2011a). However, these techniques often estimate the RUL of suspension histories to essentially convert them to failure histories. Although this can aid the RUL estimation process by increasing the training sample size, poor estimations of suspension history failure times may introduce noise or bias (both positive and negative) into the estimation. As such, in this preliminary analysis, suspension histories are not employed.

The training set consisted of three failure histories. These were processed as above and validated against two further bearings. A failed bearing of which the remaining useful life was known, and also a bearing which did not fail during the test in order to truly estimate the remaining useful life of the bearing.

6.4.3   Results

Initially, only the scaled RMD estimates were utilised in conjunctions with the autoregressive inputs of this feature (lags) in order to provide non-linear mapping of condition to the remaining useful life. As can be seen in figures 6.18 and 6.19, the derived
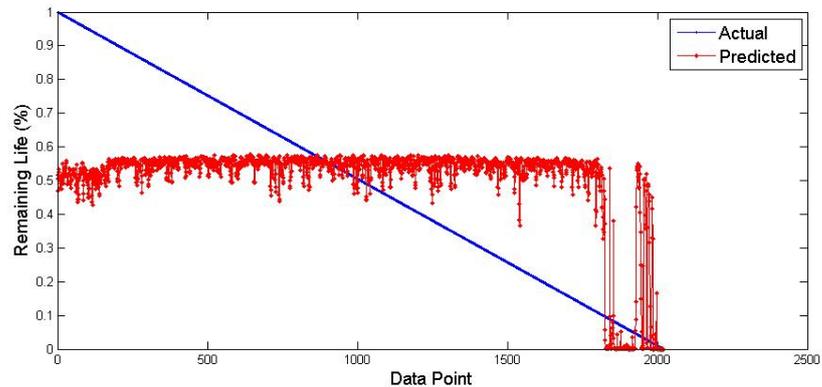
Figure 6.18: RUL estimation of bearing 2 (training dataset).
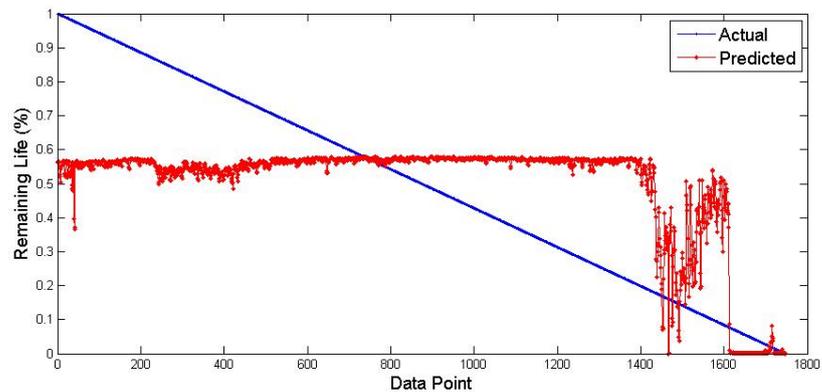
Figure 6.19: RUL estimation of bearing 3 (training dataset).

predictive function is of little quality. In this case, $r = .45$ and $r = .57$ (respectively). The life of the bearing is shown to be constant over time until the condition degrades substantially. As such, it clear that condition alone in this case is not capable of predicting the remaining useful life of the bearing.

Following this, it was noted that within the literature, degradation is often referred to as a function of condition and time. Thus, the feature set (lagged condition of the bearing) was augmented with the age of the bearing. Training of this neural network resulted in higher quality mapping of features to the remaining useful life of the bearings.

As can be seen in the training data (figures 6.21, 6.20 and 6.22), the error is reduced dramatically. As such, the correlation of this predictive function to the underlying remaining useful life is increased. In this case, $r = .99$, $= .96$ and $r = .99$ (respectively). This shows that the neural network is effectively providing a mapping from the input
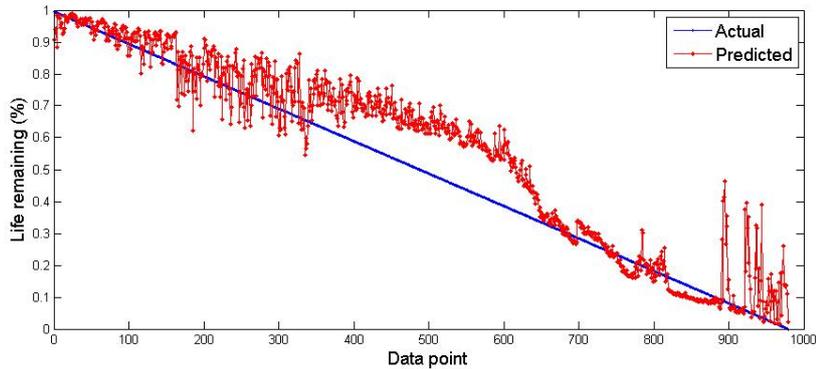
Figure 6.20: Refined RUL estimation of bearing 1 (training dataset).
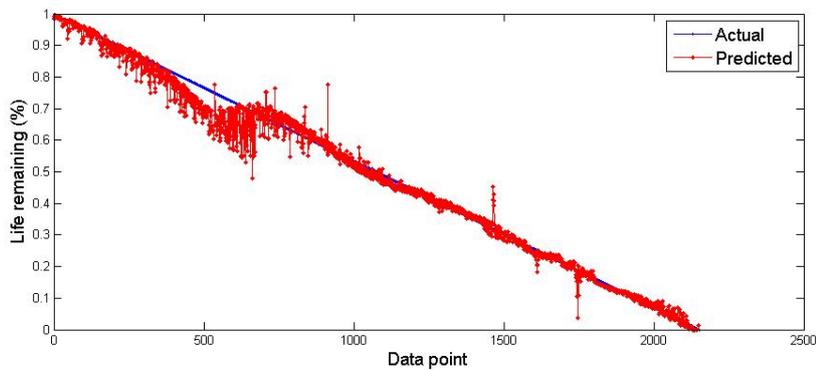
Figure 6.21: Refined RUL estimation of bearing 2 (training dataset).

feature space to the remaining useful life of the bearings.

In order to validate that the neural network truly is providing a strong mapping from the feature space to the remaining useful life, two further bearings were employed to ensure that the network was not over fitting to the training data; the first bearing represents a failure history, with the second representing a bearing suspension history.

Figure 6.23 shows the prediction of the RUL for the independent failed bearing. In this case, a correlation of $r = .96$ is attained with an error $MSE = .007$. This shows that the neural network is capable of providing an adequate mapping from the input space to the remaining useful life of the bearing.

Following this, the second validation bearing employed was a suspension history. As this did not fail, it is of interest to see if the network was capable of determining the remaining useful life of data of this context. Figure 6.24 shows the prediction of RUL over time for this bearing. As the bearing did not fail, it is not possible to present an actual RUL of this bearing. For completeness, the condition of this bearing over time is presented is figure 6.25.
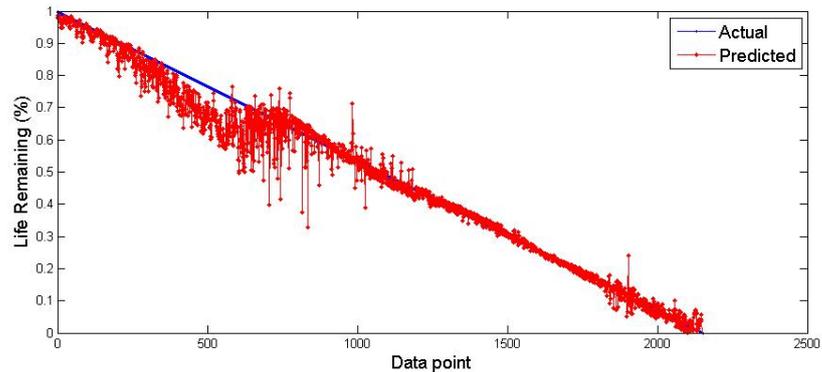
Figure 6.22: Refined RUL estimation of bearing 3 (training dataset).
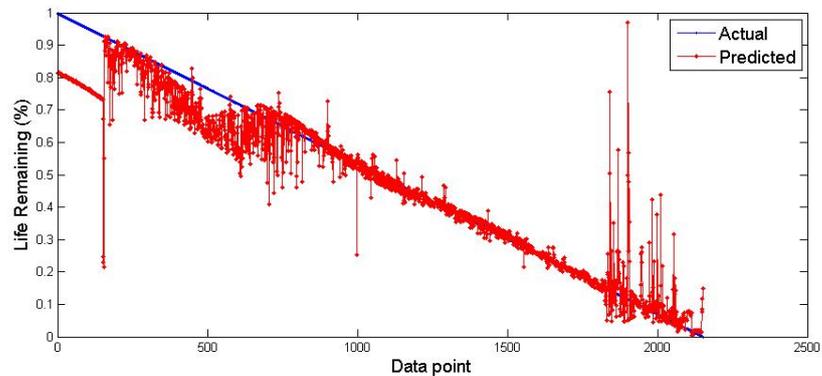
Figure 6.23: Validation of refined RUL estimation on independent failure history with $r = .96$. (bearing 1, validation dataset).

6.4.4 Discussion

Initially, utilising only the auto-regressive condition index values, it is shown that whilst RUL estimation is possible, only a low quality prognostic is attained. This is due to the mapping from condition to remaining useful life not providing the additional context of age to the neural network. Once the age of the bearing is incorporated into the analysis, the error of the RUL estimation reduces dramatically. The inclusion of this knowledge, correlations to the remaining useful life are $r = .45$ and $r = .57$. After inclusion, this increases to $r = .99$, $r = .96$ and $r = .99$. These are near perfect correlations for the estimation of remaining useful life. As such, it is recommended that this meta-data be incorporated when performing this analysis. A similar technique is employed in Tian et al. (2010). However, in their work, no empirical evidence or justification is provided for including this attribute. This work provides some insight as to why this meta-data should be incorporated into the RUL estimation process.
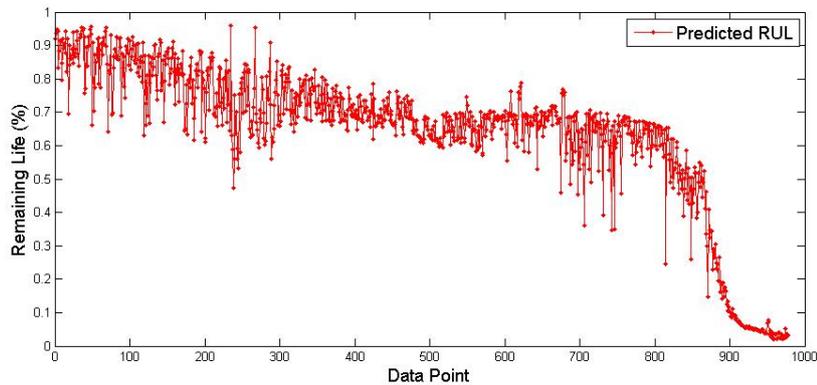
Figure 6.24: Validation of refined RUL estimation on independent suspension history (bearing 2, validation dataset).
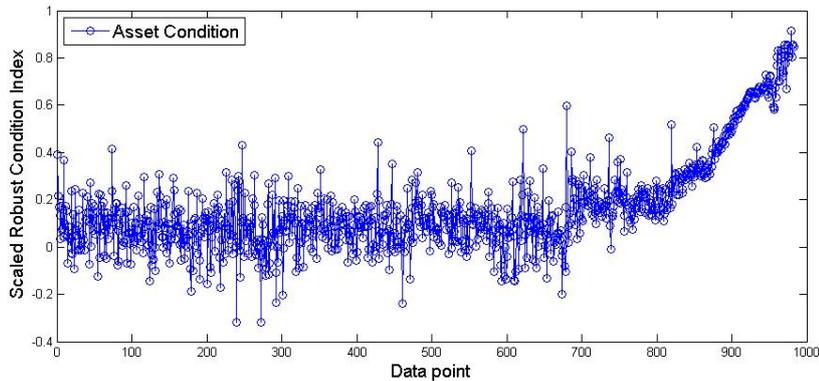
Figure 6.25: Time series of bearing condition of suspended bearing (bearing 2, validation dataset).

Whilst a sufficient training set can be created for the supervised training of the neural network, this is potentially an inaccurate representation. Although each data point from each failure history can be included within the training set (creating a large training dataset), only three failure histories are employed for the training of the network. As such, limited information regarding the variability in operational behaviour (due to operational conditions and inherent manufacturing defects) is present within the data. Ideally, additional failure data would be made available in order to enable more accurate encapsulation of the degradation process so that more accurate estimates of RUL could be attained. Although it is possible to use synthetic data and suspension histories, in this paradigm, it is unlikely that these would be representative of the true failure histories employed in this analysis and may reduce the quality of the RUL estimate.

With regards to the choice of distance metric, in this case the robust Mahalanobis distance was employed. Due to the unbounded nature of this metric, normalisation of the data was performed. It is intuitive to premise that similarity metrics (such as Cosine similarity) may be a more effective choice of metric due to the natural bounds between -1 and 1 which are created. However, as shown in chapter 5, these techniques may not be as sensitive as the robust Mahalanobis distance, and as such, may not provide as accurate an estimate of RUL as provided by the methodology employed in this work.

6.4.5 Conclusion

This work has presented a methodology which enables the estimation of the remaining useful life (RUL) of bearings based upon time-domain statistical features. The bearing moments model developed in section 4.2 was employed in conjunction with the robust Mahalanobis distance as explored in section 5.2.

In this case, only one structure of artificial neural network is employed. Namely, the feed forward back propagation multilayer perceptron network. Other network structures are available, as discussed in chapter 2. It may be of interest, for example, for future work to explore the use of other network topologies (such as cascade forward neural networks) or self organising maps to determine if this initial exploration can be improved upon.

As this analysis only explored the bearing moments model developed in section 4.2, it may be the case that other features may provide more prognostic insight and reduce the RUL estimate error.

Future work will explore the potential to utilise this methodology with survival analysis in order to incorporate failure histories, suspension histories and potentially synthetic (simulated) data in order to increase the accuracy of the RUL estimation.

6.5 Chapter summary

In this chapter, the exploitation of the developed condition index has been exploited for the purposes of tacit knowledge codification (section 6.2), diagnosis (section 6.3), and finally estimation of the asset RUL (section 6.4).

By extracting propositional rules from the dataset in section 6.2, the accuracy and agreement between the condition metric and also developed expert systems can be explored. In this case, the RIPPER algorithm attained an average classification accuracy of 96.11% across three independent validation sets with an average inter-model agreement between the created expert system and the condition index of $\kappa = 94.16\%$.

In total, 12 rules were extracted from the data repository pertaining to the operational characteristics of the wind turbine gearbox. Of these, 8 were found to be intuitive and provide insight into the condition index.

In section 6.3, seeded fault data was employed to explore the possibility of utilising ensembles of distance metrics for diagnosis. In total, four operational modes of interest (normal behaviour, inner race fault, outer race fault and roller element fault) are considered. Utilising the bearing moments model developed in section 4.2, fault detection is shown to be 100% accurate in this case, and is then shown to be 100% with regards to fault identification (diagnosis). Validation of the technique is performed on an independent dataset, and initially shown to have little practical diagnostic capability. Upon refinement, the model gains some diagnostic capability. However, due to substantial differences between data sources, little to no diagnostic capability is found to exist in this case.

Finally, artificial neural networks are shown to be capable of providing a non-linear mapping from the feature space (compromising the bearing moments model developed in section 4.2) to the estimated remaining useful life of the component. Initially, only the condition index (a long with associated autoregressive inputs) are employed, and shown not to correlate effectively with remaining useful life. Due to this, attribute variance and bearing age were added to the model. This enabled the effective estimation of bearing RUL, with validation of the technique on an independent bearing found to correlate highly with the estimated RUL ($r = .96$, $MSE = .007$).

6.6   Research objective 3 (RO3)

This chapter has explored the use of exploitation of the condition metrics for both bearings and gearboxes across high and low frequency data. Tacit knowledge was shown to be possible, along with diagnosis, condition prediction and remaining useful life estimation. As such, this chapter provides insight into research objective 2 (RO2) specified in this thesis in section 3.2.3. RO3 is stated as:

RO3: Exploitation of multivariate distance metrics for early detection, diagnosis and prognosis

> *To exploit any suitably derived metric for the early detection, diagnosis and prognosis of both bearings and gearboxes.*

This is further broken down into three research questions (RQ5, RQ6, RQ7). The results of the analysis performed in this chapter will now be utilised to answer this research objective and the associated research questions.

6.6.1   RQ5: Tacit knowledge codification

Research question 5 (RQ5) explores the use of multivariate distance metrics to extract tacit knowledge from data repositories. As such, this research question corresponds to the research undertaken in section 6.2 of this thesis. RQ5 is:

> *RQ5: Can robust multivariate distance metrics be used as a means for rapidly labelling data for the development of expert systems and extraction of tacit knowledge from historical data repositories? If so, what level of accuracy and agreement can be attained? (RO3)*

In order to answer this question, the four sub-questions proposed in section 3.3.3 must be answered.

Initially, the identification of relevant techniques was undertaken in the systematic literature review undertaken in chapter 2. This identified a substantial number of data mining techniques which could be explored. The work of Hall et al. (2009) provides many potential candidates for this. Based upon this work, a selection of techniques were chosen based upon their computational complexity, relevance and also knowledge output format.

Within the literature, rule extraction within PHM literature is not common. This technique is considered in Brotherton et al. (2000) however, artificial neural networks

are chosen instead due to their superior performance. Similarly, the work of Kusiak and Verma (2011) performed rule extraction, however, the rules are potentially difficult to interpret and potentially have a somewhat lower accuracy (69% to 86%) than the technique presented here.

The trade-off which exists between the techniques is typically of accuracy against both the size of the knowledge-base and also algorithmic complexity. It is possible to achieve a higher classification accuracy, however, this is often at the expense of a substantially larger rule base (as in the decision table and C4.5 algorithms). For practical reasons, it is recommended that the RIPPER algorithm is employed due to the order-independent nature of the rules.

Validation of the codified knowledge can be performed through any of the traditional validation techniques. In this case, Cohens' κ statistic (Cohen, 1960) was employed to ensure the learned model was accurately representing the underlying condition metric. Furthermore, validation against three independent datasets was performed. Cross-fold validation or bootstrapping are alternatives techniques for this, however, these were not required in this case. Qualitative validation of the techniques was performed by an independent domain expert who deemed 8 of the 12 rules intuitive.

As such, in answering RQ5, multivariate distance metrics can be exploited to rapidly label data to develop both expert systems and also to codify tacit knowledge. In this case, an average accuracy of 96.11% was attained with the RIPPER algorithm, with an agreement (as measured by Cohens' κ) to the underlying condition metric of 94.16%, as explored in section 6.2.4

6.6.2 RQ6: Fault diagnosis

Research question 6 (RQ6) explores the possibility of utilising ensembles of multivariate distance metrics to encapsulate various operational behaviours for the purposes of fault diagnosis. Research question 6 is stated in section 3.3.6 as:

RQ6: Can the application of robust multivariate distance or similarity metrics be employed for the diagnosis of developing bearing faults? If so, how can this be performed, and with what accuracy can fault diagnosis take place? (RO3)

In order to answer this, the two sub-questions proposed in section 3.3.6 must be answered.

Firstly, the bearing moments model identified in section 4.2 (compromising the RMS, standard deviation, skewness and kurtosis) were found to differ substantially and significantly due to degradation through. This was determined through traditional

analysis of variance (ANOVA). This enabled the detection of faults through these features. Post-hoc analysis showed each feature differed significantly between each failure model. As such, when utilising these features with the robust Mahalanobis distance, diagnosis between these failure modes is possible in this case. Verification of this was found to be 100% on the seeded fault dataset.

Although not explored in detail in the preliminary analysis performed in section 6.3, the underlying categories of features does matter. As has been demonstrated, poor feature and distance metric selection substantially and significantly reduces the quality of the developed diagnostic tool. Due to the nature of degradation, it is recommended that features from multiple categories are utilised so that multiple aspects of the underlying data are measured.

Due to the availability of failure data with faults (seeded or otherwise) it is possible to use this a priori knowledge to label failures for training. Whilst clustering would provide a means to perform a labelling of the data, the labels for these would not be available prior to the analysis. As such, it would be possible to identify faults, however, knowing which fault had developed would require further analysis.

Secondly, multivariate distance metrics have been shown to be capable of being utilised for diagnosis. As the robust Mahalanobis distance has been shown to be the most sensitive, this was employed for the analysis. The increased sensitivity of this metric enables more accurate diagnosis as smaller differences between samples can be identified. It is likely that less sensitive techniques (such as the Minkowski distances) would not achieve the level of diagnostic capability which has been achieved by the robust Mahalanobis distance.

In this case, full coverage of failure modes (with regards to bearing pitting and spalling) was achieved. Although substantial differences in the validation dataset reduced the capabilities of the technique in practice, verification of the technique on the validation dataset was also found to be 100% accurate across all failure modes.

As such, in answering RQ6, multivariate distance metrics can be utilised for both early detection and also diagnosis, by employing ensembles of metrics trained on seeded fault data. In this case, 100% verification accuracy was achieved, however, due to practical limitations, this performance could not be validated on the independent dataset.

6.6.3   RQ7: Prognosis

Research question 7 (RQ7) explores the possibility of utilising artificial neural networks to perform non linear state estimation for both asset condition and remaining useful life. Research question 7 is stated in section 3.3.7 as:

*RQ7: Can the application of robust multivariate distance or similarity metrics be exploited by non-linear techniques for either the prognosis of asset condition by future condition prediction or remaining useful life estimation? If so, how accurately can these be performed? (**RO3**)*

In order to answer this, the four sub-questions proposed in section 3.3.7 must be answered.

Initially, the systematic review undertaken in chapter 2 enabled techniques explored within the literature to be identified. Artificial neural networks were found to be the most commonly utilised technique to perform prognosis, with support vector machines (SVM) also commonly being utilised. These techniques are preferred due to their non-linear nature, as linear techniques are often not capable of modelling the non-linear degradation process (as shown in this thesis in section 6.4). In this work, the artificial neural network was chosen due to their superiority over the SVM.

With regards to remaining useful life estimation, a correlation coefficient of $r = .96$ is achieved for an independent bearing in section 6.4. This represents a strong correlation between the trained neural network and the underlying function of bearing degradation. In order to achieve this, data from three failure histories bearings was employed. Whilst the artificial neural network does not account for all the variance in the degradation function, the prediction is capable of being utilised within a condition based maintenance (CBM) framework in order to realise reductions in maintenance budgets. It is anticipated that more data would lead to more accurate models being created, however, given the law of diminishing returns, it is premised only two failure histories would be necessary to utilise this technique effectively in practice.

As such, in answering RQ7, multivariate distance metrics can be utilised for prognosis by employing artificial neural networks to estimate remaining useful life or estimate future asset condition. In this case, 5 step ahead prediction is shown to be achievable on low frequency wind turbine gearbox SCADA data as in section **??** and remaining useful life estimation is shown to be achievable with a correlation of $r = .96$ to the underlying degradation function on high frequency bearing accelerometer data in section 6.4.

6.6.4 RO3: Exploitation of multivariate distance metrics for early detection, diagnosis and prognosis

Having explored RQ5, RQ6 and RQ7, it is possible to answer research objective 3 (RO3). This states:

RO3: Exploitation of multivariate distance metrics for early detection, diagnosis and prognosis

> *To exploit any suitably derived metric for the early detection, diagnosis and prognosis of both bearings and gearboxes.*

Building upon the research undertaken in chapter 5, the developed models and derived distance metric were exploited to enable PHM. Tacit knowledge codification was performed found to be beneficial in understanding the causes of failure in section 6.2. Average classification accuracy of 96.11% was achieved in determining gearbox normal operation, inspection conditions and also damage conditions. Furthermore, diagnosis was found to be possible in section 6.3. Verification accuracy was high (100%) however, in practice this was shown not to perform as desired due to differences between datasets.

Finally asset prognosis were undertaken.Remaining useful life estimation was performed with the assistance of artificial neural networks.These were used to derive a non-lineap mapping between the input space (asset condition) and output space (remaining useful life). This was found to be possible within this paradigm, and the developed neural network was found to perform this task well ($r = .96$) when utilising the developed bearing moments model (section 4.2).

As such, RO3 has been answered throughout this chapter. It is possible to develop a PHM system performing early detection, diagnosis and prognosis based upon the two models developed in this thesis. Furthermore, favourable results are achieved when utilising multivariate distance metrics in order to perform this analysis. However, it is essential to employ both a sensitive distance metric (such as the robust Mahalanobis distance), strong encapsulation of behaviours (as in the bearing and gearbox models) and also powerful non-linear techniques (such as artificial neural networks).

Chapter 7

Conclusions

This thesis has detailed seven research questions across three research objectives. This chapter summarises the results of the research undertaken in this thesis and provides avenues for future exploration which may be of interest to readers of this thesis.

Initially, this thesis presented multivariate condition assessment models for bearings (the moments model) and gearboxes (the extended physics of failure model) in chapter 4. These were transformed into health metrics through the assistance of multivariate distance and similarity metrics in chapter 5, with encapsulation of asset behaviour being performed by artificially intelligent search techniques in section 5.3. These metrics were then exploited in chapter 6 by enabling tacit knowledge extraction in section 6.2, fault diagnosis in section 6.3 and remaining useful life estimation in section 6.4.
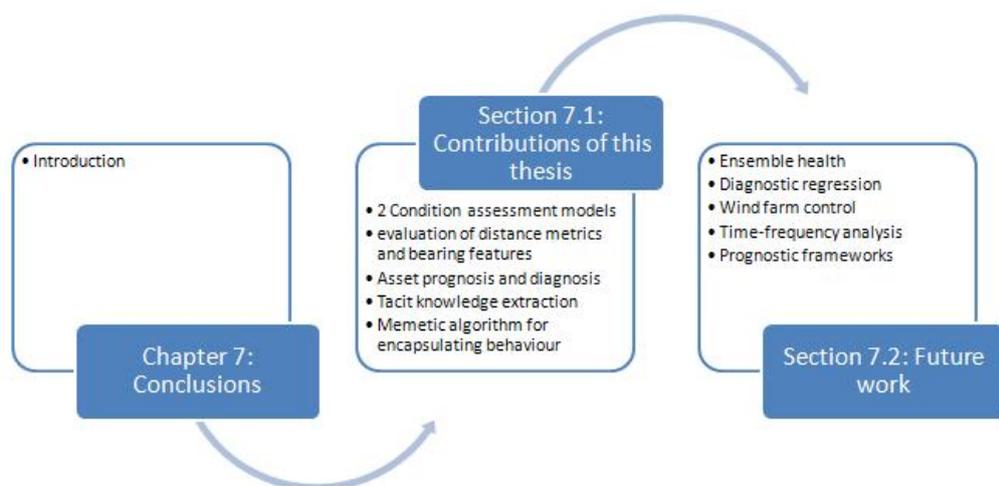
Figure 7.1: Overview of the structure of chapter 7.

7.1 Conclusions of this thesis

It is clear that there is a distinct need to move away from traditional 'fail and fix' maintenance practices towards 'predict and prevent' maintenance strategies within wind energy. The potential exists to substantially reduce the operations and maintenance (O&M) budget of operators by pro-actively monitoring critical components and pre-emptively maintaining these components prior to failure. However, due to the infantile nature of many predictive and proactive techniques, the uptake of these techniques has been limited in practice.

This thesis has attempted to mitigate the risks associated with the development – and subsequent use – of a PHM system by presenting methodologies which are transparent, based upon sound statistical underpinnings and and exploit powerful predictive techniques. The contributions of this thesis will now be summarised, bringing together the research undertaken throughout this thesis.

7.1.1 Analysis of time-domain statistical features

Analysis of time-domain statistical features for the purpose of condition assessment based upon high frequency was undertaken in section 4.2. In this, 28 statistical features in the time-domain were comparatively analysed across five distinct categories of feature. These were:

- *Measures of central tendency:* Mean, median, mode, tri-mean, root mean square (RMS) and the winsorized mean.

- *Measures of variability:* Standard deviation, variance, interquartile range (IQR), range, minimum, maximum and median absolute deviation (MAD).

- *Measures of shape:* Skewness, kurtosis, hyperskewness, hyperflatness, crest factor (CF) and the peak to average power ratio (PAPR).

- *Measures of position:* 10th percentile, 25th percentile, 75th percentile and the 90th percentile.

- *Measures of impurity:* Shannon entropy, Chao-Shen entropy, Miller-Madow entropy, James-Stein style shrinkage estimator of entropy and a Bayesian estimate of entropy (Krichevsky-Trofimov).

Of these, measures of variability were found to perform the most consistently, followed by measures of shape, position, impurity and finally entropy.

Whilst this was an extensive analysis of these features, it was not by any means exhaustive. Due to the nature of these metrics, a comparative analysis of all of these features was not possible and beyond the scope of this thesis. However, the research undertaken provided a means to utilise statistical process control as a means to quantify feature quality when the standard assumptions of statistical process control are not met (for instance, normally distributed data). Furthermore, metrics were defined in order to quantify the sensitivity of the metric (by determining how much data is deemed to be in control over the asset history) and also to quantify the potential as a leading indicator of degradation (by determining the point at which the features leave predetermined bounds of control).

7.1.2 Encapsulation of bearing behaviour based upon statistical time-domain features and high frequency data (moments model)

Encapsulation of bearing behaviour was performed in section 4.2.11. In this, a comparative analysis covering the verification and validation of 28 time-domain statistical features was undertaken. Five categories of features were explored, namely; measures of central tendency, measures of variability, measures of shape, measures of position and measures of impurity. A robust statistical process control chart utilising exponentially weighted moving averages (EWMA) was employed due to practical limitations of the underlying data. A multivariate model compromising 4 of the 28 features was created as a result of this analysis, incorporating: (1) RMS, (2) standard deviation, (3) skewness and (4) kurtosis.

This has been referred to as the moments model throughout this thesis due to the incorporation of statistical moments within the model.

Although skewness and kurtosis were found to be mediocre features in this instance, given the prevalence of these features in the literature, it is likely that the poor results attained for these features are due to the metrics employed for the analysis and stochastic artefacts in the underlying dataset. As such, these features were chosen due to their prevalence within the literature, and also as a means to further encapsulate the bearing behaviour. As these features (in essence) represent the first four moments, encapsulation of the underlying distribution is more accurate as a result when analysis is performed in the multivariate domain.

7.1.3 Wind turbine gearbox extended physics of failure model based upon low frequency SCADA data

The extension of a physics of failure model for condition assessment of the wind turbine gearbox was performed in section 4.3. In this, the model described in Feng et al. (2011) was extended to incorporate further improvements, namely:

- Normalisation of transient states.

- Normalisation against turbine de-rating and curtailment (though performance charts).

This has been referred to as the extended physics of failure model throughout this thesis due to the further refinements over the previous model of Feng et al. (2011).

This model is shown to inherently take wind turbine wake into consideration, and overcomes the limitations of other models – such as that of Kusiak and Verma (2013) – in that the practical application of the model is not limited by turbine rating or curtailment. This not only enables a more holistic model to be utilised in the condition assessment of the wind turbine gearbox, but also enables the specification of new performance charts from which additional metrics can be created to quantify the current condition of the gearbox. The transient nature of wind is normalised for, further increasing the quality of the model.

7.1.4 Analysis of multivariate distance and similarity metrics

Analysis of multivariate distance and similarity metrics was undertaken in section 5.2. In this, eight multivariate distance metrics and one multivariate similarity metric were explored for their capabilities in encapsulating asset condition in a univariate manner. The multivariate distance metrics explored were three Minkowski distances (Euclidean, Manhattan and Chebyshev), the Canberra distance, the Lance-Williams distance, the Penrose distance, the traditional Mahalanobis and the robust Mahalanobis distance. The similarity metric explored was that of cosine similarity.

Exploration across two datasets covering three experiments designed to identify the strengths and weaknesses of the technique were performed. The Minkowski metrics were identified as adequate for the purposes of condition assessment in the multivariate domain, however, should be avoided as the interactions between attributes is not taken into consideration. The same conclusions were reached for the Lance-Williams distance and also the Canberra distance metrics. The cosine similarity metric was found to perform well, but inconsistently in this case due to the nature of the technique. The Penrose distance was found to be more sensitive than the previous metrics, however

suffers from potential problems when attributes are highly correlated. As these interactions are accounted for in the Mahalanobis distance, this is both more sensitive and also more practical to employ. Finally, the robust Mahalanobis distance is shown to provide more insight than the traditionally defined metric due to its increased sensitivity.

7.1.5 Elitist memetic algorithm for encapsulation of asset behaviour

Encapsulation of asset behaviour was performed with the assistance of artificially intelligent search techniques in section 5.3. In this case, an elitist memetic algorithm was employed to identify a robust subset of points which could be utilised in conjunction with the Mahalanobis distance in order to increase the sensitivity of the metric and assist in the identification of anomalous behaviour (such as degradation). A comparative analysis of four algorithms was undertaken on real world wind turbine gearbox SCADA data in order to evaluate the effectiveness of the identified techniques. The elitist memetic algorithm was compared and contrasted against a random search of the search space, a traditional genetic algorithm and also the Fast-MCD heuristic.

In this case, the elitist memetic algorithm was found to outperform both the random search and genetic algorithm in all instances which were analysed. Furthermore, for larger datasets ($N = 1,000$), the elitist memetic algorithm was found to be superior to that of the widely utilised Fast-MCD heuristic. For the largest dataset explored ($N = 100,000$), within 15 seconds the elitist memetic algorithm had identified a superior solution than the Fast-MCD heuristic was able to find in an hour of search. Similarly, for the case $N = 10,000$, all of the 40 subsets identified by the elitist memetic algorithm were superior to be best subset identified over 40 iterations of the Fast-MCD heuristic ($\det(\Sigma_{EMA}^{-1}) = 20.50$, $\det(\Sigma_{Fast-MCD}^{-1}) = 20.51$). All of the explored techniques consistently outperformed the traditional estimation of the covariance matrix.

7.1.6 Tacit knowledge extraction from historical data repositories

Tacit knowledge extraction was performed in section 6.2. In this, low frequency wind turbine gearbox SCADA data processed as per the extended physics of failure model from section 4.3 was utilised for a comparative analysis of three algorithms (decision table, C4.5 and RIPPER) to explore their capabilities in extracting tacit knowledge from the underlying health metric. Statistically sound thresholds based upon analysis in the χ^2 domain were utilised due to the non-normality of the underlying multivariate data.

Cross-validation was performed utilising data from three wind turbines, with data from one turbine being held out of each analysis. The RIPPER algorithm attained

the highest average accuracy (96.11%) and highest agreement to the underlying model on average ($\kappa = 94.16\%$). This was followed by the C4.5 algorithm which attained 92.68% accuracy on average along with an average agreement to the underlying model of $\kappa = 89.03\%$. Finally, the decision table algorithm attained an average accuracy of 85.87% and an average agreement to the underlying model of $\kappa = 78.80\%$.

Although all three techniques worked well, the RIPPER algorithm was found to have the highest accuracy and agreement whilst simultaneously having the smallest knowledge base (12 rules, in comparison to the 33 nodes of the C4.5 algorithm and the 412 rules of the decision table algorithm). As such, the RIPPER algorithm is recommended over both the C4.5 and decision table algorithms in this case.

7.1.7 Fault diagnosis through multivariate distance ensembles

Fault diagnosis was performed with the assistance of ensembles of multivariate distance metrics in section 6.3. In this, bearing data encapsulated by the moments model (developed in section 4.2) with faults seeded through electro-static discharge machining (0.007 inches in size) was employed covering the three failure modes (inner race, outer race and roller elements) pertaining to bearing surface fatigue (spalling and pitting). In the analysis, 10 samples across each failure mode were employed in addition to 10 samples representing normal operational behaviour for a total of 40 samples.

Statistical analysis employing analysis of variance revealed that each feature differed significantly due to the seeded faults. Further insight was gained through employing traditional and robust post-hoc analysis. In this case, the bearing model derived in section 4.2 was found to be capable of not only fault detection but also identification (fault diagnosis). Verification of the diagnostic accuracy was found to be 100%, with no false negatives or positives reported. This was reduced substantially when utilising inferior features (as identified in section 4.2) and an inferior distance metric (as identified in section 5.2). Furthermore, validation on an independent test set identified practical limitations of this technique due to underlying differences in the datasets. Re-training of the diagnostic ensembles on the validation set enabled some diagnostic insight to be achieved, but this was limited.

7.1.8 Prognosis of bearing remaining useful life

Prognosis of bearing remaining useful life was performed with the assistance of non-linear mapping performed by artificial neural networks in section 6.4. In this, bearing data processed with the moments model (developed in section 4.2) from three failure histories was utilised as supervised training data in order to provide an estimate of the

remaining useful life of bearings in real-time. Data was pre-processed for the artificial neural network with the assistance of a logarithmic-transform.

Initially, utilising only the auto-regressive inputs of asset condition, it was not possible to accurately estimate the remaining useful life of the bearings as the trained function did not correlate well to the underlying function of remaining useful life on two of the three verification bearings ($r = .45$ and $r = .57$). However, modelling the function of degradation as a function of both condition and time, and incorporating the variance of the condition substantially increased the quality of the prognostic technique developed. The developed technique was then able to correlate highly to the three bearings in the verification set; $r = .99$, $r = .96$ and $r = .99$ (respectively). Independent validation on an unseen dataset also found a strong correlation of $r = .96$. Whilst this is not perfect, it does provide a means to perform prognostics in practice for the purposes of condition based maintenance.

7.2 Outcomes of this thesis

This thesis has detailed the development of two models for the condition assessment of components. Specifically, a multivariate model for the encapsulation of high frequency bearing accelerometer data (the moments model) and an extended physics of failure model for the quantification of wind turbine gearbox condition based upon SCADA data.

Throughout this thesis, emphasis has been given on practical techniques which can readily be adopted in industry. To assist this, analysis has been performed on publicly available datasets utilising transparent techniques where possible, empirically deriving models from the data itself.

This thesis has presented a holistic set of techniques which can be utilised within a PHM solution in order to enable the early detection, diagnosis and prognosis of faults. As such, this thesis provides a means to enable organisations to derive more value from their assets through state-of-the-art techniques. Commonly available sources of condition monitoring data were utilised, mitigating the necessity for additional investment. Furthermore, low frequency techniques were presented in order to reduce data storage and bandwidth requirements, mitigating concerns with regards to the implementation of these techniques, whilst assisting with the industrial uptake of these methodologies and techniques.

As specified in section 3.1 of this thesis, the overall aim of this thesis was to explore:

To what extent can the use of robust multivariate statistical methodologies

and data-driven techniques be employed and exploited as a metric for con-
dition assessment to assist in the early detection, diagnosis and prognosis
of both bearings and gearboxes with respect to wind turbines?

Given the research undertaken through this thesis – specifically in chapters 4, 5 and 6 – this question has been answered. Early detection was shown to be possible by leading features for bearings in section 4.2 and also gearboxes in 5.2. Diagnosis was performed based upon high frequency bearing data in section 6.3, with prognosis of asset condition found to be possible (section **??**) and ultimately RUL estimation being performed (section 6.4).

As such, this thesis has presented a complete set of general methodologies which enable PHM to be undertaken on a variety of assets and components at varying levels of data resolution. As such, the methodologies presented in this thesis are valuable in a wide range of domains, and applicable far beyond wind turbine gearbox assessment.

7.3   Recommendations for future work

In the course of this research various new techniques and methodologies have been presented. As such, there exists many natural extensions to this work which not only complement the research undertaken in this thesis, but also provide a means to verify and validate the research presented here. The following areas have been identified as being of value in the continuation of the research presented in this thesis.

7.3.1   Ensemble turbine health

The models developed in chapter 4 provide the means to develop a condition index for a single bearing (potentially within a gearbox), or the gearbox itself within a wind turbine. However, various other critical components exist within the wind turbine. By exploiting the notions of ensembles which have been presented and discussed in this thesis in section 6.3 for the purposes of diagnosis, further insights into the current operational behaviour of the wind turbine can be achieved.

For instance, it would be possible to develop models for each subsystem of the wind turbine (for instance, those specified in section 1.5). These models can then be hierarchically incorporated into a further multivariate distance metric which takes the condition of each subsystem as an input, and presents the condition of the turbine as a whole as the output. This would enable the identification of wind turbines which were

potentially at risk or failure, whilst providing a means to identifying the subsystem which was likely to fail (similar to the research undertaken in section 6.3 of this thesis).

7.3.2 Diagnostic regression

As the diagnostic ensemble model developed in section 6.3 utilises fixed size seeded failures (0.007 inch), the developed technique can detect and identify the initial development of spalling and pitting on the bearings. In practice however, failures will not be of a fixed size and this may reduce the effectiveness of this technique in practice.

As such, it is of interest to utilise empirical data to quantify not only the type of fault, but to also estimate the size (severity of the fault). Multivariate distance metrics provide a means to perform this (as larger distances potentially indicate a more severe fault), however, this mapping is not anticipated to be linear. As such, utilising non-linear state estimation techniques (such as artificial neural networks) in conjunction with the models and concepts developed throughout this thesis, it would be possible to estimate the size of developing faults. This would provide maintenance operators and decision makers the necessary information required to make effective decisions regarding the operation of the component.

7.3.3 Wind farm control strategies

Whilst beyond the scope of this thesis, the implications of this research can be used to assist in the development of control strategies for wind farms as a whole. As the research undertaken throughout this work has provided a means to quantify degradation on the wind turbine gearbox, this then presents the opportunity to exploit this as both a metric and constraint with regards to the control of the entire wind farm.

Typically, wind turbines within a wind farm operate at full capacity in order to generate the maximum amount of energy from the wind. This increases the loading of the components. Due to wind turbine wake effects, down-wind turbines generate substantially less energy as less energy is available within the wind itself. State-of-the-art research on wind turbine control strategies have shown that this is in some ways inefficient, as it may be possible to extract more energy from the wind by careful curtailment and de-rating of up-wind turbines. If it is possible to curtail or de-rate specific wind turbines to increase power production, this also reduces the load (and thus, fatigue) on the wind turbine.

As the techniques presented in this thesis provide a means to indirectly measure the fatigue of the gearbox, this can be used as a means to minimise the total fatigue presented to the wind farm as a whole. This would increase the life of components,

whilst providing equal (or potentially greater) power production. As such, the use of holistic wind farm control strategies is essential for ensuring maximum return on investment is achieved.

7.3.4 Time-frequency domain analysis of features

In this thesis, analysis of features has typically been performed in the time domain. This enabled statistical features to be combined into multivariate models for the purposes of condition assessment (as per the moments model as presented in section 4.2 of this thesis). However, due to the a priori knowledge which is often available for inclusion within analysis as meta-data, analysis in the frequency or time-frequency domains should not be ignored. It may be the case the more accurate diagnosis of failure modes can be attained when utilising the this meta-data. Furthermore, these techniques may also provide a means to quantify the severity of degradation in a similar manner to those explored in throughout this thesis.

7.3.5 Frameworks for comparative prognostic analysis

Due to the infantile nature of prognostics and health management, there is a lack of publicly available datasets, methodologies and frameworks which enable the comparative analysis of different techniques. As prognostics is often employed on mission critical assets, often data is not made publicly available due to its inherent commercial sensitivity. As such, it can be difficult to determine which techniques should be employed for a given dataset, why this is the case and also the typical power of these techniques. It would be of great value to derive a framework to enable the comparative evaluation of various different prognostic techniques with a common metric. As different domains will typically have different data sources and also different requirements in a PHM system, being able to compare and contrast these is of great interest.

Recently, preliminary methodologies to perform this have been presented at various conferences (such as the 2014 IEEE PHM conference and the 2014 COMADEM conference). Whilst these provide a potential means to perform this analysis, further research is necessary.

Bibliography

Herv Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. ISSN 1939-0068.

Paul Addison. *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance.* CRC Press, 2002.

Charu C. Aggarwal. *Outlier Analysis.* Springer: New York, 2013.

Rosmaini Ahmad and Shahrul Kamaruddin. A review of condition-based maintenance decision-making. *European journal of industrial engineering*, 6 (5):519–541, 2012.

W. Albring. A. betz, introduction to the theory of flow machines. xvi + 281 s. m. 168 fig. pergamon pres. preis geb. 75 s. net. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift fr Angewandte Mathematik und Mechanik*, 47(2):140–141, 1967. ISSN 1521-4001.

Kjell Aleklett, Mikael Höök, Kristofer Jakobsson, Michael Lardelli, Simon Snowden, and Bengt Söderbergh. The peak of the oil age - analyzing the world oil production reference scenario in world energy outlook 2008. *Energy Policy*, 38(3):1398–1414, 2010. ISSN 0301-4215. Security, Prosperity and Community Towards a Common European Energy Policy? Special Section with Regular Papers.

William R. L. Anderegg, James W. Prall, Jacob Harold, and Stephen H. Schneider. Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107 (27):12107–12109, 2010.

Q. Guo andW. Wu, D. L. Massart, C. Boucon, and S. De Jong. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(12):123–132, 2002. ISSN 0169-7439.

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, June 1999. ISSN 0163-5808.

European Wind Energy Association. Wind in our sails. Technical report, European Wind Energy Association, November 2011.

L. F. Awosika, G. T. French, R. J. Nicholls, and C. E Ibe. The impact of sea level rise on the coastline of nigeria. In *the CC Symposium on the Rising Challenges of the Sea*, Magaritta, Venezuela, March 1992.

Bart Baesens, Rudy Setiono, Christophe Mues, and Jan Vanthienen. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, 49(3):312–329, 2003.

Gustavo Batista, Ronaldo Prati, and Maria Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

David. S. Battisti and Rosamond L. Naylor. Historical warnings of future food insecurity with unprecedented seasonal heat. *Science*, 323(5911):240–244, 2009.

N. Baydar, Q. Chen, A. Ball, and U. Kruger. Detection of incipient tooth defect in helicar gears using multivariate statistics. *Mechanical Systems and Signal Processing*, 15(2):303–321, 2001. ISSN 0888-3270.

Eric Bechhoefer and Praneet Menon. Bearing envelope analysis window selection. In *Proceedings of the annual conference of the prognostics and health management society (PHMSociety)*, 2009.

E. Becker and P. Poste. Creative condition monitoring. Online, July 2009.

R. J. Beckman and R. D. Cook. [outlier..........s]: Response. *Technometrics*, 25 (2): 161–163, May, 1983.

Thorsten Bernholt and Paul Fischer. The complexity of computing the MCD-estimator. *Theoretical Computer Science*, 326(1):383–398, 2004.

Bruce P. Bogert, Michael JR Healy, and John W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the symposium on time series analysis. Vol. 15.*, 1963.

Ørnulf Borgan. *Encyclopaedia of biostatistics*, chapter Kaplan-Meier Estimator, pages 2154–2160. John Wiley & Sons, Ltd, 2014. ISBN 9781118445112.

Nick Bostrom, Yudkowsky Eliezer, Keith Frankish (Ed.), and William Ramsey (Ed.). *Cambridge Handbook of Artificial Intelligence*, chapter The Ethics of Artificial Intelligence. New York: Cambridge University Press., 2013.

Xavier Chiementin Bovic Kilundu and Pierre Dehombreux. Singular spectrum analysis for bearing defect detection. *Journal of vibration and acoustics*, 133 (5):1–7, 2011.

George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26 (2):211–252, 1964.

George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control, 4th Edition*. Wiley Publishing, 2008.

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

Peter Brockwell and Richard Davis. *Introduction to Time Series and Forecasting*. Taylor & Francis, 2002.

E. Brodsky and Boris S. Darkhovsky. *Nonparametric methods in change point problems*. Number volume 243 in Monograph series. Springer, 1993.

David Broomhead, D. S.; Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321355, 1988.

Tom Brotherton, Gary Jahns, Jerry Jacobs, and Dariusz Wroblewski. Prognosis of faults in gas turbine engines. In *Aerospace Conference Proceedings, 2000 IEEE*, volume 6, pages 163–171. IEEE, 2000.

Ralph T. Buscarello. *Practical Solutions to Machinery and Maintenance Vibration Problems*. Update International, 1994.

Shane Butler, John Ringwood, and Frank OConnor. Exploiting SCADA system data for wind turbine performance monitoring. In *Proceedings of the Conference on Control and Fault-Tolerant Systems (SysTol),October 9-11, 2013. Nice, France*, 2013.

Wahyu Caesarendra, Achmad Widodo, and Bo-Suk Yang. Application of relevance vector machine and logistic regression for machine degradation assessment. *Mechanical Systems and Signal Processing*, 24(4):1161–1171, 2010. ISSN 0888-3270.

Case Western Reserve University. Seeded fault test data. Online, 2011. URL http://csegroups.case.edu/bearingdatacenter/home.

J. Cendrowska. PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27 (4):349–370, 1987.

Cai-Wan Chang-Jian and Shiuh-Ming Chang. Bifurcation and chaos analysis of spur gear pair with and without nonlinear suspension. *Nonlinear Analysis: Real World Applications*, 12(2):979–989, 2011. ISSN 1468-1218.

Anne Chao and Tsung-Jen Shen. Nonparametric estimation of Shannons index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, 10(4):429–443, 2003.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, 16:321–357, 2011.

Samira Chekkal, Narimen Aouzellag Lahaani, Djamal Aouzellag, and Kaci Ghedamsi. Fuzzy logic control strategy of wind generator based on the dual-stator induction generator. *International Journal of Electrical Power & Energy Systems*, 59(0):166–175, 2014. ISSN 0142-0615.

B. Chen, Y. N. Qiu, Y. Feng, P. J. Tavner, and W. W. Song. Wind turbine SCADA alarm pattern recognition. In *Renewable Power Generation (RPG 2011), IET Conference on*, pages 1–6, September 2011a.

Bindi Chen. *Automated On-line Fault Prognosis for Wind Turbine Monitoring using SCADA data*. PhD thesis, Univesity of Durham, 2014.

Bindi Chen, Peter Tavner, Yanhui Feng, William Song, and Yingning Qiu. Bayesian network for wind turbine fault diagnosis. In *Proceedings of the annual conference of the european wind energy association, Copenhagen, Denmark, 16-19 April, 2012*, 2012.

Bindi Chen, Christoper Crabtree, Donatella Zappalá, and Peter Tavner. Survey of commercially available SCADA data analysis tools for wind turbine health monitoring. Technical report, SUPERGEN WIND UK, May 2014.

Hongbo Xuand Guohua Chen, Xinhua Wang, and Jun Liang. The application of EMD and ARMA bi-cepstrum fault diagnosis method in gearbox of overhead traveling crane. *Advanced materials research*, 308 - 310:88–91, 2011b.

Jigang Chen and Guowen Hao. Research on the fault diagnosis of wind turbine gearbox based on Bayesian networks. In Yinglin Wang and Tianrui Li, editors, *Practical Applications of Intelligent Systems*, volume 124 of *Advances in Intelligent and Soft Computing*, pages 217–223. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-25657-8.

Kai-Ying Chen and Yi-Cheng Shaw. Applying back propagation network to cold chain temperature monitoring. *Advanced Engineering Informatics*, 25(1):11–22, 2011. ISSN 1474-0346. RFID and sustainable value chains.

Zhongsheng Chen and Yongmin Yang. Fault diagnostics of helicopter gearboxes based on multi-sensor mixtured hidden Markov models. *Journal of Vibration and Acoustics*, 134 (3):1–8, 2012.

Chia Chen Ciang, Jung-Ryul Lee, and Hyung-Joon Bang. Structural health monitoring for a wind turbine system: a review of damage detection methods. *Measurement Science and Technology*, 19(12):122001, 2008.

J. B. Coble and J. W. Hines. Prognostic algorithm categorization with PHM challenge application. In *Prognostics and Health Management, 2008. PHM 2008. International Conference on*, pages 1–11, October 2008.

Jamie Coble and Wesley Hines. Applying the general path model to estimation of remaining useful life. In *Proceedings of the annual conference of the prognostics and health management society*, 2009.

Marco Cocconcelli, Radoslaw Zimroz, Riccardo Rubini, and Walter Bartelmus. STFT based approach for ball bearing fault detection in a varying speed motor. In Tahar Fakhfakh, Walter Bartelmus, Fakher Chaari, Radoslaw Zimroz, and Mohamed Haddar, editors, *Condition Monitoring of Machinery in Non-Stationary Operations*, pages 41–50. Springer Berlin Heidelberg, 2012a. ISBN 978-3-642-28767-1.

Marco Cocconcelli, Radoslaw Zimroz, Riccardo Rubini, and Walter Bartelmus. Kurtosis over energy distribution approach for STFT enhancement in ball bearing diagnostics. In Tahar Fakhfakh, Walter Bartelmus, Fakher Chaari, Radoslaw Zimroz, and Mohamed Haddar, editors, *Condition Monitoring of Machinery in Non-Stationary Operations*, pages 51–59. Springer Berlin Heidelberg, 2012b. ISBN 978-3-642-28767-1.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

L. Cohen. Time-frequency distributions - a review. *Proceedings of the IEEE*, 77(7): 941–981, July 1989. ISSN 0018-9219.

Leon Cohen. *Time-frequency analysis.* Prentice Hall: New Jersey, 1995a.

W. W. Cohen and Y. Singer. A simple, fast, and effective rule learner. In *Proceedings of the National Conference on Artificial Intelligence*, pages 335–342, July 1999.

William W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995b.

Harry M. Collins. Tacit knowledge, trust and the Q of sapphire. *Social studies of science*, 31(1):71–85, 2001.

F. Combet and L. Gelman. An automated methodology for performing time synchronous averaging of a gearbox signal without speed sensor. *Mechanical Systems and Signal Processing*, 21(6):2590–2606, 2007. ISSN 0888-3270.

Feiyun Cong, Jin Chen, and Yuna Pan. Kolmogorov-smirnov test for rolling bearing performance degradation assessment and prognosis. *Journal of Vibration and Control*, 17 (9):1337–1347, 2010.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp*, 19:297–301, 1965.

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

C. J. Crabtree, Y. Feng, and P. J. Tavner. Detecting incipient wind turbine gearbox failure: a signal analysis method for on-line condition monitoring. In *Proceedings of the Scientific Track of the European Wind Energy Conference*, pages 154–156, 2010.

Christopher Crabtree. *Condition Monitoring Techniques for Wind Turbines*. PhD thesis, University of Durham, 2011.

Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

G. Dalpiaz, A. Rivola, and R. Rubini. Effectiveness and sensitivity of vibration processing techniques for local fault detection in gears. *Mechanical Systems and Signal Processing*, 14(3):387–412, 2000. ISSN 0888-3270.

Christine P. Dancey and John Reidy. *Statistics without maths for psychology*. Pearson Education, 2007.

D. J. Davis. An analysis of some failure data. *Journal of the American Statistical Association*, 47(258):113–150, 1952.

A.C. Davison. *Statistical Models*. Cambridge University Press, 2003.

R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

G. D'Elia, E. Mucchi, and G. Dalpiaz. On the time synchronous average in planetary gearboxes. In *International Conference Surveillance 7*, October 2013.

Department of Energy and Climate Change. UK renewable energy roadmap. Technical report, Department of Energy and Climate Change, July 2011.

Department of Energy and Climate Change. DECC electricity generation costs 2013. Technical report, Department of Energy and Climate Change, July 2013.

Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, 2009.

I. Dinwoodie, F. Quail, and D. McMillan. Analysis of offshore wind turbine operation and maintenance using a novel time domain meteo-ocean modeling approach. In *ASME Turbo Expo 2012: Turbine Technical Conference and Exposition*. ASME, June 2012.

Ming Dong and Ying Peng. Equipment PHM using non-stationary segmental hidden semi-markov model. *Robotics and Computer-Integrated Manufacturing*, 27(3):581–590, 2011. ISSN 0736-5845.

Yuhua Dong and Yanqiu Cui. Analysis of a new joint time-frequency distribution of suppressing cross-term. *Research Journal of Applied Sciences, Engineering and Technology*, 4 (11):580–1584, 2012.

E. Elbeltagi, T. Hegazy, and D. Grierson. Comparison among five evolutionary-based optimization algorithms. *Advanced Engineering Informatics*, 19 (1):43–53, 2005.

James B. Elsner and Anastasios A. Tsonis. *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Springer, 1996.

Michael Kingsley Eric Bechhoefer. A review of time syncronous average algorithms. In *Annual conference of the prognostics and health management society (PHMSociety)*, 2009.

The Crown Estate. Offshore wind cost reduction: pathways study. Technical report, The Crown Estate, May 2013.

Martin Ester, Hans peter Kriegel, Jörg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge discovery and data mining (KDD)*, pages 226–231. AAAI Press, 1996.

European Wind Energy Association. A summary of opinion surveys on wind power. Technical report, European wind energy association, September 2003.

European Wind Energy Association. Wind energy - the facts. Volume 2. Costs and prices. Technical report, European wind energy association, 2005.

Ryan D. Evans. Classic bearing damage modes. In *Wind Turbine Tribology Seminar*, Broomfield, CO, USA, November 2011. NREL.

Shu Fang, Deliang Li, and Litao Liuand Zhenwei Zhang. Pitting corrosion diagnosis of bearing based on power cepstrum and histogram. *Advanced materials research*, 512 - 515:1672–1676, 2012.

Yanhui Feng, Yingning Qiu, Christopher J Crabtree, Hui Long, and Peter J Tavner. Use of SCADA and CMS signals for failure detection and diagnosis of a wind turbine gearbox. In *Proceedings of the annual conference of the european wind energy association*, 2011.

Zhipeng Feng, Ming Liang, and Fulei Chu. Recent advances in time-frequency analysis methods for machinery fault diagnosis: A review with application examples. *Mechanical Systems and Signal Processing*, 38(1):165–205, 2013. ISSN 0888-3270. Condition monitoring of machines in non-stationary operations.

David Ferguson and Victoria M. Catterson. Big data techniques for wind turbine condition monitoring. In *Annual conference of the European wind energy association*, 2014.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Carlos M. Fonseca and Peter J. Fleming. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary computation*, 3(1):1–16, 1995.

Michael L. Fugate, Hoon Sohn, and Charles R. Farrar. Vibration-based damage detection using statistical process control. *Mechanical Systems and Signal Processing*, 15 (4):707–721, 2001. ISSN 0888-3270.

Haodong Gao and Yidu Zhang. Nonlinear behavior analysis of geared rotor bearing system featuring confluence transmission. *Nonlinear Dynamics*, 76(4):2025–2039, 2014. ISSN 0924-090X.

M. C. Garcia, M. A. Sanz-Bobi, and J. del Pico. SIMAP: Intelligent system for predictive maintenance: Application to the health condition monitoring of a windturbine gearbox. *Computers in Industry*, 57(6):552–568, 2006.

Nagi Z. Gebraeel, Mark A. Lawley, Rong Li, and Jennifer K. Ryan. Residual-life distributions from component degradation signals: A Bayesian approach. *IIE Transactions*, 37(6):543–557, 2005.

Thomas Gellermann. Extension of the scope of condition monitoring systems for multi-MW and offshore wind turbines. *VGB powertech*, 9:81–85, 2013.

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian data analysis: third edition.* CRC Press, 2014.

S. H. Ghafari, F. Golnaraghi, and F. Ismail. Effect of localized faults on chaotic vibration of rolling element bearings. *Nonlinear Dynamics*, 53(4):287–301, 2008. ISSN 0924-090X.

V. N. Ghate and S. V. Dudul. Cascade neural-network-based fault classifier for three-phase induction motor. *Industrial Electronics, IEEE Transactions on*, 58(5):1555–1563, May 2011. ISSN 0278-0046.

Ramanathan Gnanadesikan and John R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28 (1):81–124, 1972.

Jamie Godwin and Peter Matthews. Classification and detection of wind turbine pitch faults through SCADA data analysis. *Internation Journal of the Prognostis and Health management society (PHMSociety)*, 4 (11):1–11, 2013.

Jamie Godwin and Peter Matthews. A low frequency uni-variate model for the effective diagnosis and prognosis of bearing signals based upon high frequency data. In *Proceedings of the annual conference of the prognostics and health management society (PHMSociety)*, 2014a.

Jamie L. Godwin and Peter Matthews. Rapid labelling of SCADA data to extract transparent rules using RIPPER. In *Reliability and Maintainability Symposium (RAMS), 2014 Proceedings - Annual*, pages 1–6, 2014b.

Jamie L. Godwin and Peter Matthews. *Data Mining and Analysis in the Engineering Field*, chapter Robust Statistical Methods for Rapid Data Labelling., pages 107–141. IGI Global, 2014c.

I. Gordon, L. Carnel, D. Van Gestel, G. Beaucarne, and J. Poortmans. 8% efficient thin-film polycrystalline-silicon solar cells based on aluminum- induced crystallization and thermal CVD. *Progress in Photovoltaics: Research and Applications*, 15(7):575–586, 2007. ISSN 1099-159X.

Emily Gosden. Offshore wind farm developer DONG energy claims it can beat government cost target, March 2013. [Online; posted 1-March-2013].

S. A. Grady, M. Y. Hussaini, and M. M. Abdullah. Placement of wind turbines using genetic algorithms. *Renewable Energy*, 30 (2):259–270, 2005.

Huageng Luo Hai Qiu and Neil Eklund. On-board aircraft engine bearing prognostics: Enveloping or FFT analysis? In *The ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2009.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

Z. Hameed, Y. S. Hong, Y. M. Cho, S. H. Ahn, , and C. K. Song. Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renewable and Sustainable energy reviews*, 13(1):1–39, 2009.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, volume 3644 of *Lecture Notes in Computer Science*, pages 878–887. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-28226-6.

Anca D. Hansen and Lars H. Hansen. Wind turbine concept market penetration over 10 years (1995 - 2004). *Wind Energy*, 10(1):81–97, 2007. ISSN 1099-1824.

Johanna Hardin and David M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4)(4):928–946, 2005.

S. P. Harsha, K. Sandeep, and R. Prakash. The effect of speed of balanced rotor on non-linear vibrations associated with ball bearings. *International Journal of Mechanical Sciences*, 45(4):725–740, 2003. ISSN 0020-7403.

P. Hart. The condensed nearest neighbor rule (corresp.). *Information Theory, IEEE Transactions on*, 14(3):515–516, May 1968. ISSN 0018-9448.

Hossein Hassani. Singular spectrum analysis: Methodology and comparison. *Journal of data science*, 5 (2):239–257, 2007.

Jean Hausser and Korbinian Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10:1469–1484, 2009.

D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

Qingbo He, Fanrang Kong, and Ruqiang Yan. Subspace-based gearbox condition monitoring by kernel principal component analysis. *Mechanical Systems and Signal Processing*, 21(4):1755–1772, 2007. ISSN 0888-3270.

Aiwina Soong Yin Heng. *Intelligent prognostics of machinery health utilising suspended condition monitoring data*. PhD thesis, Queensland University of Technology, 2009.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617, 1990.

D. Hochmann and E. Bechhoefer. Envelope bearing analysis: theory and practice. In *Aerospace Conference, 2005 IEEE*, pages 3658–3666, March 2005.

D. Hochmann and R. B. Randall. Optimising of bearing diagnostic techniques using simulated and actual bearing fault signal. *Mechanical Systems and Signal Processing*, 14(5):763–788, 2000. ISSN 0888-3270.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

David W. Hosmer Jr., Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time to event data*. Wiley, 2011.

David W. Hosmer Jr., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley, 2013.

Shengguo Hu, Yanfeng Liang, Lintao Ma, and Ying He. MSMOTE: Improving classification performance when training data is imbalanced. In *Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on*, volume 2, pages 13–17, October 2009.

Guang-Bin Huang. Learning capability and storage capacity of two-hidden-layer feedforward networks. *Neural Networks, IEEE Transactions on*, 14(2):274–281, March 2003. ISSN 1045-9227.

Marion King Hubbert. Nuclear energy and the fossil fuels. Technical Report Vol. 95, Shell Development Company, Exploration and Production Research Division, Houston, TX., June 1956.

Peter J Huber. *Robust statistics*. Springer, 2011.

Mia Hubert, Peter J. Rousseeuw, and Tim Verdonck. A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21 (3):618–637, 2012.

Intergovernmental Panel on Climate Change. Organisation, 1988.

H. Janicke, M. Bottinger, U. Mikolajewicz, and G. Scheuermann. Visual exploration of climate variability changes using wavelet analysis. *IEEE Transactions on Visualization and Computer Graphics*, 15 (6):1375–1382, 2009.

A. K. S. Jardine, D. Lin, and D. Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

D. N. Joanes and C. A. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189, 1998.

Preston Johnson. TSA methods for bearing and gearbox signature enhancements in wind turbines. Technical report, National Instruments, 2011.

I. T. Jolliffe. *Principal Component Analysis, Series: Springer Series in Statistics, (2nd ed)*. Springer: New York, 2002.

Cheng Junsheng, Yu Dejie, and Yang Yu. A fault diagnosis approach for roller bearings based on EMD method and AR model. *Mechanical Systems and Signal Processing*, 20(2):350–362, 2006. ISSN 0888-3270.

Abhinav Saxena Kai Goebel, Bhaskar Saha. A comparison of three data driven techniques for prognostics. In *62nd meeting of the machinery failure and prevention technology society (MFPT)*, 2008.

Haiying Kang, Yanjie Qi, Guangsheng Liu, Jinhua Liu, and Chunquan Li. Application of the ARMA model in non-stationary vibration signals. In *Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2012 International Conference on*, pages 751–754, June 2012.

Chinmaya Kar and A. R. Mohanty. Technical note: Gearbox health monitoring through multiresolution Fourier transform of vibration and current signals. *Structural Health Monitoring*, 5(2):195–200, 2006.

L. Kaufman and P.J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis (1 ed.)*. New York: John Wiley. ISBN 0-471-87876-6., 1990.

Sanaa Kerroumi, Xavier Chiementin, and Lanto Rasolofondraibe. Dynamic classification method of fault indicators for bearings monitoring. *Mechanics & Industry*, 14:115–120, January 2013. ISSN 2257-7750.

K. Kim, G. Parthasarathy, O. Uluyol, W. Foslien, S. Shuangwen, and P. Fleming. Use of SCADA data for failure detection in wind turbines. In *Energy Sustainability Conference and Fuel Cell Conference,Washington DC, August 7 - 10*, 2011.

Ron Kohavi. The power of decision tables. In *Machine Learning: ECML-95*, pages 174–189. Springer, 1995.

S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, and W. Gronwald. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8 (Suppl 1):146–160, 2012.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982. ISSN 0340-1200.

B. Kosasih, W. Caesarendra, K. Tieu, A. Widodo, C. A. Moodie, and A. K. Tieu. Degradation trend estimation and prognosis of large low speed slewing bearing lifetime. *Applied Mechanics and Materials*, 493:343–348, 2014.

S. B. Kotsiantis and P. E. Pintelas. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1 (1):46–55, 2003.

M. A. Kramer. Autoassociative neural networks. *Computers & Chemical Engineering*, 16(4):313–328, 1992. ISSN 0098-1354. Neutral network applications in chemical engineering.

R. Krichevsky and V. Trofimov. The performance of universal encoding. *Information Theory, IEEE Transactions on*, 27(2):199–207, 1981.

Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

Jain Y. Kumar and Santosh Kumar Bhandare. Min max normalization based data perturbation method for privacy protection. *International Journal of Computer and Communication Technology*, 2 (8):45–50, 2011.

Cheng-Chien Kuo, Po-Hung Chen, An Liu, and Li-Ming Chen. Defect type recognition system for wind turbine by subtractive clustering. In *Intelligent System Design and Engineering Application (ISDEA), 2012 Second International Conference on*, pages 1404–1408, January 2012.

A. Kusiak and A. Verma. A data-mining approach to monitoring wind turbines. *Sustainable Energy, IEEE Transactions on*, 3(1):150–157, January 2012a. ISSN 1949-3029.

A. Kusiak and A. Verma. Monitoring wind farms with performance curves. *Sustainable Energy, IEEE Transactions on*, 4(1):192–199, 2013. ISSN 1949-3029.

Andrew Kusiak and Anoop Verma. A data-driven approach for monitoring blade pitch faults in wind turbines. *Sustainable Energy, IEEE Transactions on*, 2(1):87–96, 2011.

Andrew Kusiak and Anoop Verma. Analyzing bearing faults in wind turbines: A data-mining approach. *Renewable Energy*, 48(0):110–116, 2012b. ISSN 0960-1481.

Dae-Ho Kwak, Dong-Han Lee, Jong-Hyo Ahn, and Bong-Hwan Koh. Fault detection of roller-bearings using signal processing and optimization algorithms. *Sensors*, 14 (1):283 – 298, 2014.

G. N. Lance and W. T. Williams. Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, 9(1):60–64, 1966.

G. N. Lance and W. T Williams. Mixed-data classificatory programs I.) Agglomerative systems. *Australian Computer Journal*, 1(1):15–20, 1967.

Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In Silvana Quaglini, Pedro Barahona, and Steen Andreassen, editors, *Artificial Intelligence in Medicine*, volume 2101 of *Lecture Notes in Computer Science*, pages 63–66. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-42294-5.

Marc LeBlanc and Anne Marie Graves. Condition monitoring systems: Trends and cost benefits. National Renewable Energy Laboratory (NREL) Wind Turbine Condition Monitoring Workshop, September 2011.

Hyungdae Lee, C. Byington, and M. Watson. PHM system enhancement through noise reduction and feature normalization. In *Aerospace Conference, 2010 IEEE*, pages 1–10, 2010.

J Lee, H Qiu, G Yu, and J Lin. Rexnord technical services, bearing data set, IMS, University of Cincinnati. NASA Ames prognostics data repository, 2007.

Jay Lee, Fangji Wu, Wenyu Zhao, Masoud Ghaffari, Linxia Liao, and David Siegel. Prognostics and health management design for rotary machinery systems: Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 42(1 - 2): 314–334, 2014. ISSN 0888-3270.

Lawrence M Leemis. *Reliability: probabilistic models and statistical methods*. Prentice-Hall, Inc., 1995.

E. Levrat, B. Iung, and A. Crespo Marquez. E-maintenance: review and conceptual framework. *Production Planning & Control*, 19(4):408–429, 2008.

Chuan Li, Ming Liang, Yi Zhang, and Shumin Hou. Multi-scale autocorrelation via morphological wavelet slices for rolling element bearing fault diagnosis. *Mechanical Systems and Signal Processing*, 31(0):428–446, 2012. ISSN 0888-3270.

Fucai Li, Lin Ye, Guicai Zhang, and Guang Meng. Bearing fault detection using higher-order statistics based ARMA model. *Key engineering materials*, 347:271–276, 2007.

G. Liao, S. Liu, T. Shi, and G. Zhang. Gearbox condition monitoring using self-organizing feature maps. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 218(1):119–129, 2004.

Haitao Liao, Wenbiao Zhao, and Huairui Guo. Predicting remaining useful life of an individual unit using proportional hazards model and logistic regression model. In *Reliability and Maintainability Symposium, 2006. RAMS '06. Annual*, pages 127–132, January 2006.

J. Lin and M. J. Zuo. Gearbox fault diagnosis using adaptive wavelet filter. *Mechanical systems and signal processing*, 17(6):1259–1269, 2003.

Po-Hung Chen Lin, Deng-Fa and Mike Williams. Measurement and analysis of current signals for gearbox fault recognition of wind turbine. *Measurement Science Review*, 13 (2):89–93, 2013.

B. Liu, S. Riemenschneider, and Y. Xu. Gearbox fault diagnosis using empirical mode decomposition and Hilbert spectrum. *Mechanical Systems and Signal Processing*, 20 (3):718–734, 2006. ISSN 0888-3270.

David B. Lobell, Marshall B. Burke, Claudia Tebaldi, Michael D. Mastrandrea, Walter P. Falcon, and Rosamond L. Naylor. Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863):607–610, 2008.

H. Long, J. Wu, and P. J. Tavner. Analysis of statistical loading conditions of wind turbine gearboxes based on SCADA data. In *Proceedings of the annual conference of the european wind energy association*, 2012.

C. Joseph Lu and William Q. Meeker. Using degradation measures to estimate a time-to-failure distribution. *Technometrics*, 35 (2):161–174, May, 1993.

Chen Lu, Qian Sun, Laifa Tao, Hongmei Liu, and Chuan Lu. Bearing health assessment based on chaotic characteristics. *Shock and Vibration*, 20(3):519–530, 2013.

Huageng Luo, Hai Qiu, George Ghanime, Melinda Hirz, and Geo van der Merwe. Synthesized synchronous sampling technique for differential bearing damage detection. *Journal of engineering for gas turbines and power*, 132 (7):1–8, 2010.

David J. C. MacKay. *Sustainable Energy - Without the Hot Air*. UIT Cambridge LTD., 2008.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

A Malhi and R .X. Gao. PCA-based feature selection scheme for machine defect classification. *Instrumentation and Measurement, IEEE Transactions on*, 53(6):1517–1525, December 2004. ISSN 0018-9456.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. London: Academic Press., 1979.

Kanti V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

Ignacio Santa María, Carlos Pantaleón, and Jesús Ibañez. A comparative study of high-accuracy frequency estimation methods. *Mechanical Systems and Signal Processing*, 14(5):819–834, 2000. ISSN 0888-3270.

Fabio Cecconi Massimo Cencini and Angelo Vulpiani. *Chaos From Simple models to complex systems*. World Scientific, 2009.

Allan May, David McMillan, and Sebastian Thöns. Economic analysis of condition monitoring systems for offshore wind turbine sub-systems. In *Proceedings of the annual conference of the European Wind Energy Association (EWEA)*, 2014.

P. D. McFadden. Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration. *Journal of Vibration and Acoustics*, 108 (2):165 – 170, 1986.

S. A. McInerny and Y. Dai. Basic vibration signal processing for bearing fault detection. *Education, IEEE Transactions on*, 46(1):149–156, February 2003. ISSN 0018-9359.

Geoffrey McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 2004.

Becki Meadows. Offshore wind O&M challenges. In *Wind Turbine Condition Monitoring Workshop*. National Renewable Energy Laboratory, September 2011.

A. Mesemanolis, C. Mademlis, and I. Kioskeridis. A fuzzy-logic based control strategy for maximum efficiency of a wind energy conversion system. In *Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 2012 International Symposium on*, pages 7–12, June 2012.

Richard G. Miller and Steven R. Sorrell. The future of oil supply. *Philosophical Transactions of the Royal Society A: Mathematical,Physical and Engineering Sciences*, 372 (2006):20130179, 2014.

Douglas C Montgomery. *Introduction to statistical quality control*. John Wiley & Sons, 2007.

A. Moosavian, H. Ahmadi, and A. Tabatabaeefar. Fault diagnosis of main engine journal bearing based on vibration analysis using fisher linear discriminant, K-nearest neighbor and support vector machine. *Journal of Vibroengineering*, 14(2):894–906, 2012.

B. Muruganatham, M. A. Sanjith, B. K. Kumar, S. A. V. S. Murty, and P. Swami-nathan. Inner race bearing fault detection using singular spectrum analysis. In *Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on*, pages 573–579, October 2010.

Bubathi Muruganatham, M. A. Sanjith, B. Krishnakumar, and S. A. V. Satya Murty. Roller element bearing fault diagnosis using singular spectrum analysis. *Mechanical Systems and Signal Processing*, 35(1 - 2):150–166, 2013. ISSN 0888-3270.

Anand Natarajan. An overview of the state of the art technologies for multi-MW scale offshore wind turbines and beyond. *Wiley Interdisciplinary Reviews: Energy and Environment*, 3(2):111–121, 2014. ISSN 2041-840X.

H. Joseph Newton and Emanuel Parzen. Forecasting and time series model types of 111 economic time series. Technical report, Texas A & M University College Station Institute of statistics, 1983.

Hieu V. Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Computer Vision–ACCV 2010*, pages 709–720. Springer, 2011.

Thomas Dyhre Nielsen and Finn Verner Jensen. *Bayesian networks and decision graphs*. Springer, 2009.

N. G. Nikolaou and I. A. Antoniadis. Demodulation of vibration signals generated by defects in rolling element bearings using complex shifted morlet wavelets. *Mechanical Systems and Signal Processing*, 16(4):677–694, 2002. ISSN 0888-3270.

V. Niola, G. Quaremba, and V. Avagliano. Further developments on gear transmission monitoring. In *the 9th WSEAS International Conference on simulation, modelling and optimization*, 2009.

Sebastian Nusser. *Robust Learning in Safety-Related Domains: Machine Learning Methods for Solving Safety-Related Application Problems*. PhD thesis, Otto-von-Guericke-Universitat, Magdeburg, Germany, 2009.

Frank Nyame-Asiamah. Exploiting tacit knowledge through knowledge management technologies. In *Proceedings of the Learning Forum London Conference*, June, 2009.

John S. Oakland. *Statistical Process Control*. Butterworth-Heinemann, 2003.

Farag K. Omar and A. M. Gaouda. Dynamic wavelet-based tool for gearbox diagnosis. *Mechanical Systems and Signal Processing*, 26(0):190–204, 2012. ISSN 0888-3270.

T. V .I. (Ed.) Onofeghara, F. A. Akpata and D. U. U. (Ed.) Okali. *Nigerian Wetlands: An Overview.* (Port Harcourt, Nigeria: UNESCO/MAB, 1990), 1990.

A. V. Oppenheim. *Superposition in a class of nonlinear systems.* PhD thesis, Res. Lab. Electronics, Massachusetts Institute of Technology, Cambridge, MA., 1965.

R. K. Pachauri and A. Reisinger, editors. *Fourth Assessment Report: Climate Change 2007: The AR4 Synthesis Report.* Geneva: IPCC, 2007.

Sudhakar Madhavrao Pandit and Shien-Ming Wu. *Time series and system analysis with applications.* New York: Wiley, 1983.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

P. Paris and F. Erdogan. A critical analysis of crack propagation laws. *Journal of fluid engineering*, 85 (4):528–533, 1963.

Choon-Su Park, Young-Chul Choi, and Yang-Hann Kim. Early fault detection in automotive ball bearings using the minimum variance cepstrum. *Mechanical Systems and Signal Processing*, 38(2):534–548, 2013. ISSN 0888-3270.

Mukund R. Patel. *Wind and solar power systems: design, analysis, and operation.* CRC Press, 2012.

Lionel Sharples Penrose. Distance, size and shape. *Annals of Eugenics*, 17(1):337–343, 1952.

C. Pesch. *Eigenschaften des gegenüber Ausreißern robusten MCD-Schätzers und Algorithmen zu seiner Berechnung.* PhD thesis, Universität Passau, 2000.

Paul F. Ploutz. *Global Warming: Handbook of Ecological Issues.* Xlibris Corporation, 2011.

Lech Polkowski. *Rough Sets: Mathematical Foundations.* Springer-Verlag, 2013.

Kevin L. Priddy and Paul E. Keller. *Artificial Neural Networks: An Introduction.* The international society for optical engineering, 2005.

Hai Qiu, Jay Lee, Jing Lin, and Gang Yu. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of Sound and Vibration*, 289(4):1066–1090, 2006.

Yingning Qiu, Yanhui Feng, Peter Tavner, Paul Richardson, Gabor Erdos, and Bindi Chen. Wind turbine SCADA alarm analysis for improving reliability. *Wind Energy*, 15(8):951–966, 2012. ISSN 1099-1824.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

J. Rafiee, M. A. Rafiee, and P. W. Tse. Application of mother wavelet functions for automatic gear and bearing fault diagnosis. *Expert Systems with Applications*, 37(6): 4568–4579, 2010.

Ahmad Rafsanjani, Saeed Abbasion, Anoushiravan Farshidianfar, and Hamid Moeenfard. Nonlinear dynamic modeling of surface defects in rolling element bearing systems. *Journal of Sound and Vibration*, 319(35):1150–1174, 2009. ISSN 0022-460X.

Matiur Rahman. *Applications of Fourier Transforms to Generalized Functions*. WIT Press, 2011.

V. K. Rai and A. R. Mohanty. Bearing fault diagnosis using FFT of intrinsic mode functions in hilbert-huang transform. *Mechanical Systems and Signal Processing*, 21 (6):2607–2615, 2007. ISSN 0888-3270.

Frank Rehmet. *Wind Energy - Technology and Planning*. World Wind Energy Association (WWEA), 2006.

Alvin C. Rencher. *Wiley Series in Probability and Statistics*, pages 709–715. John Wiley & Sons, Inc., 2003. ISBN 9780471271352.

RenewableUK. Wind energy in the UK: state of the industry report. Technical report, Renewable UK, October 2013.

RenewableUK. UK wind energy database (UKWED), 2014.

Lim Chi Keong Reuben and David Mba. Bearing time to failure estimation using spectral analysis features. *Structural health monitoring*, 13 (2):219–230, 2014.

Geoffrey Robson. *Multiple Outlier Detection and Cluster Analysis of Multivariate Normal Data*. PhD thesis, University of Stellenbosch, 2003.

Timothy J. Ross. *Fuzzy Logic with Engineering Applications*. Wiley, 2009.

Peter J. Rousseeuw. *Robust regression and outlier detection*. Wiley, 2005.

Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

J. P. Royston. Some techniques for assessing multivarate normality based on the Shapiro-Wilk W. *Applied Statistics*, 32(2):121–133, 1983.

Eugene P. Sabini, Jerome A. Lorenc, Oakley Henyan, and Kenneth L. Hauenstein. Bearing defect detection using time synchronous averaging (TSA) of an enveloped accelerometer signal, 2004.

E. Sainz, A. Llombart, and J. J. Guerrero. Robust filtering for the characterization of wind turbines: Improving its operation and maintenance. *Energy Conversion and Management*, 50(9):2136–2147, 2009.

N. R. Sakthivel, V. Sugumaran, and Binoy. B. Nair. Comparison of decision tree-fuzzy and rough set-fuzzy methods for fault categorization of mono-block centrifugal pump. *Mechanical Systems and Signal Processing*, 24(6):1887–1906, 2010. ISSN 0888-3270.

D. R. Salgado and F. J. Alonso. Tool wear detection in turning operations using singular spectrum analysis. *Journal of Materials Processing Technology*, 171(3):451–458, 2006. ISSN 0924-0136.

Luis Alberto Torres Salomao, Hugo Gámez Cuatzin, Juan Anzurez Marín, and Isidro I. Lázaro Castillo. Fuzzy-pi control, pi control and fuzzy logic control comparison applied to a fixed speed horizontal axis 1.5 MW wind turbine. In *Proceedings of the World Congress on Engineering and Computer Science 2012 Vol II, October 24-26, 2012, San Francisco, USA*, 2012.

N. Saravanan, V. N. S. Kumar Siddabattuni, and K. I. Ramachandran. A comparative study on classification of features by svm and psvm extracted using morlet wavelet for fault diagnosis of spur bevel gear box. *Expert Systems with Applications*, 35(3): 1351–1366, 2008. ISSN 0957-4174.

Nader Sawalhi, Robert B. Randall, and David Forrester. Separation and enhancement of gear and bearing signals for the diagnosis of wind turbine transmission systems. *Wind Energy*, 17(5):729–743, 2014. ISSN 1099-1824.

J. Schöpfel. Grey literature on bilingualism in belgium. In *proceedings of the ninth conference on grey literature (GL9)*, pages 1–9, 2008.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMO-BILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Renping Shao, Wentao Hu, Yayun Wang, and Xiankun Qi. The fault feature extraction and classification of gear using principal component analysis and kernel principal component analysis based on the wavelet packet transform. *Measurement*, 54(0): 118–132, 2014. ISSN 0263-2241.

Lixiang Shen, Francis E.H. Tay, Liangsheng Qu, and Yudi Shen. Fault diagnosis using rough sets theory. *Computers in Industry*, 43(1):61–72, 2000. ISSN 0166-3615.

Shawn Sheng. Gearbox typical failure modes, detection and mitigation methods. In *AWEA Operations & Maintenance and Safety Seminar*. NREL, January 2014.

David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*. CRC Press, 2003.

Louise Smith. Planning for onshore wind farms, May 2014.

Willoughby Smith. Effect of light on selenium during the passage of an electric current. *Nature*, 7:303, 1873.

Jan Snyman. *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms*, volume 97. Springer, 2005.

F. Spinato, P. J. Tavner, G. J. W. van Bussel, and E. Koutoulakos. Reliability of wind turbine subassemblies. *IET Renewable Power Generation*, 3:387–401(14), December 2009. ISSN 1752-1416.

David Stevens. Equipment condition monitoring. Technical report, AV-Technology Ltd., January 2011.

Thomas F. Stocker, Q. Dahe, and Gian-Kasper Plattner. Climate change 2013: The physical science basis. *Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers (IPCC, 2013)*, 5:1–1458, 2013.

Zachary G. Stoumbos and Jr. Reynolds, Marion R. The robustness and performance of CUSUM control charts based on the double-exponential and normal distributions. In Hans-Joachim Lenz and Peter-Theodor Wilrich, editors, *Frontiers in Statistical Quality Control 7*, volume 7 of *Frontiers in Statistical Quality Control*, pages 79–100. Physica-Verlag HD, 2004. ISBN 978-3-7908-0145-3.

Wensheng Su, Fengtao Wang, Hong Zhu, Zhixin Zhang, and Zhenggang Guo. Rolling element bearing faults diagnosis based on optimal morlet wavelet filter and autocorrelation enhancement. *Mechanical Systems and Signal Processing*, 24(5):1458–1472, 2010. ISSN 0888-3270. Special Issue: Operational Modal Analysis.

Nobuo Takeda. Characterization of microscopic damage in composite laminates and real-time monitoring by embedded optical fiber sensors. *International Journal of Fatigue*, 24(2 - 4):281–289, 2002. ISSN 0142-1123.

P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson, 2005.

P. J. Tavner. *Offshore Wind Turbines: Reliability, availability & maintenance*. IET Renewable Energy, 2012. ISBN 978-1-84919-229-3.

Murat C. Testik, George C. Runger, and Connie M. Borror. Robustness properties of multivariate EWMA control charts. *Quality and Reliability Engineering International*, 19(1):31–38, 2003.

Margaret Thatcher. Speech at 2nd world climate conference, November 1990.

Zhigang Tian, Lorna Wong, and Nima Safaei. A neural network approach for remaining useful life prediction utilizing both failure and suspension histories. *Mechanical Systems and Signal Processing*, 24(5):1542–1555, 2010.

Zhigang Tian, Tongdan Jin, Bairong Wu, and Fangfang Ding. Condition based maintenance optimization for wind power generation systems under continuous monitoring. *Renewable Energy*, 36(5):1502–1509, 2011.

Ivan Tomek. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(6):448–452, 1976.

Onder Uluyol, Girija Parthasarathy, Wendy Foslien, and Kyusung Kim. Power curve analytic for wind turbine performance monitoring and prognostics. In *Proceedings of the annual conference of the prognostics and health management society*, 2011.

U.S. Energy Information Administration. History of energy consumption in the united states, 1775 to 2009, February 2011.

Willem Van Gool. Innovation and technology – strategies and policies. In Olivério D.D. Soares, A. Martins da Cruz, G. Costa Pereira, Isabel M. R. T. Soares, and Albino J.P .S. Reis, editors, *Energy Policy: Fairy Tales and Factualities*, pages 93–105. Springer Netherlands, 1997. ISBN 978-0-7923-4435-3.

Anoop Verma and Andrew Kusiak. Predictive analysis of wind turbine faults: A data mining approach. In *Proceedings of the 2011 Industrial Engineering Research Conference*, May 2011.

Luisa F. Villa, Aníbal Reñones, Jose R. Perán, and Luis J. de Miguel. Angular resampling for vibration analysis in wind turbines under non-linear speed fluctuation. *Mechanical Systems and Signal Processing*, 25(6):2157–2168, 2011. ISSN 0888-3270. Interdisciplinary Aspects of Vehicle Dynamics.

W. J. Wang and P. D. McFadden. Early detection of gear failure by vibration analysis I. calculation of the time-frequency distribution. *Mechanical Systems and Signal Processing*, 7(3):193–203, 1993. ISSN 0888-3270.

W. J. Wang, J. Chen, X. K. Wu, and Z. T. Wu. The application of some non-linear methods in rorating machinery fault diagnosis. *Mechanical Systems and Signal Processing*, 15(4):697–705, 2001. ISSN 0888-3270.

Xiyang Wang, Viliam Makis, and Ming Yang. A wavelet approach to fault diagnosis of a gearbox under varying load conditions. *Journal of Sound and Vibration*, 329(9): 1570–1585, 2010. ISSN 0022-460X.

Yue Wang and David Infield. Scada data based nonlinear state estimation technique for wind turbine gearbox condition monitoring. In *Proceedings of the annual conference of the european wind energy association*, 2012.

Thomas M. Welte and Kesheng Wang. Models for lifetime estimation: an overview with focus on applications to wind turbines. *Advances in Manufacturing*, 2(1):79–87, 2014. ISSN 2095-3127.

David Vernon Widder. *The Laplace Transform, Princeton Mathematical Series*. Princeton University Press, MR 0005923, 1941.

Achmad Widodo and Bo-Suk Yang. Machine health prognostics using survival probability and support vector machine. *Expert Systems with Applications*, 38(7):8430–8437, 2011a.

Achmad Widodo and Bo-Suk Yang. Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Systems with Applications*, 38(3):2592–2599, 2011b. ISSN 0957-4174.

M. Wilkinson, B. Hendriks, F. Spinato, E. Gomez, H. Bulacio, J. Roca, and H. Long. Methodology and results of the reliawind reliability field study. In *European Wind Energy Conference (April)*, 2010.

Michael Wilkinson, Brian Darnell, Thomas van Delt, and Kier Harman. comparison of methods for wind turbine condition monitoring with SCADA data. In *Proceedings of the annual conference of the european wind energy association*, 2013.

Richard Williams. Becquerel photovoltaic effect in binary compounds. *The Journal of Chemical Physics*, 32(5):1505–1514, 1960.

Michelle Winny. SKF windcon saves £1 million for global wind farm operator. In Energy Management, December 2013.

M. L. D. Wong, L. B. Jack, and A. K. Nandi. Modified self-organising map for automated novelty detection applied to vibration signal monitoring. *Mechanical Systems and Signal Processing*, 20(3):593–610, 2006. ISSN 0888-3270.

S. Wu and D. Clements-Croome. Preventive maintenance models with random maintenance quality. *Reliability Engineering & System Safety*, 90(1):99–105, 2005.

Yuan-Kang Wu and Jing-Shan Hong. A literature review of wind forecasting technology in the world. In *Power Tech, 2007 IEEE Lausanne*, pages 504–509, July 2007.

Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, and Jorg Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *proceedings of the 14th international conference on data engineering (ICDE)*, 1998.

Yanjun Yan and James Zhang. Using edge-detector to model wake effects on wind turbines. In *IEEE International Conference on Prognostics and Health Management (PHM 2014). Spokane, Washington State, USA. 23 - 25 June*, 2014.

Wenxian Yang and Jiesheng Jiang. Wind turbine condition monitoring and reliability analysis by SCADA information. In *Mechanic Automation and Control Engineering (MACE), 2011 Second International Conference on*, pages 1872–1875, July 2011.

Wenxian Yang, Richard Court, and Jiesheng Jiang. Wind turbine condition monitoring by the approach of SCADA data analysis. *Renewable Energy*, 53(0):365–376, 2013. ISSN 0960-1481.

Yang Yang, Yinxia Liao, Guang Meng, and Jay Lee. A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. *Expert Systems with Applications*, 38(9):11311–11320, 2011. ISSN 0957-4174.

L. A. Zadeh. Fuzzy sets. *Information and control*, 8 (3):338–353, 1965.

A. Zaher, S. D. J. McArthur, D. G. Infield, and Y. Patel. Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy*, 12(6):574–593, 2009. ISSN 1099-1824.

D. Zappalá, C. Crabtree, and P. Tavner. Gear fault detection automation using Wind-Con frequency tracking. In *Proceedings of the annual conference of the european wind energy association*, 2012.

D. Zappalá, P.J. Tavner, C.J. Crabtree, and S. Sheng. Side-band algorithm for automatic wind turbine gearbox fault detection and diagnosis. *Renewable Power Generation, IET*, 8(4):380–389, May 2014. ISSN 1752-1416.

D. Zhang, W. Liu, X. Gong, and H. Jin. A novel improved SMOTE resampling algorithm based on fractal. *Journal of Computational Information Systems*, 7(6):2204–2211, 2011.

Yuping Zhang and Shufeng Ai. EMD based envelope analysis for bearing faults detection. In *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, pages 4257–4260, June 2008.

Zijun Zhang, A. Verma, and A. Kusiak. Fault analysis and condition monitoring of the wind turbine gearbox. *Energy Conversion, IEEE Transactions on*, 27(2):526–535, June 2012. ISSN 0885-8969.

Stanislav Ziaran and Radoslav Darula. Determination of the state of wear of high contact ratio gear sets by means of spectrum and cepstrum analysis. *Journal of vibration and acoustics*, 135 (2):1–10, 2013.

Walter Zucchini and Iain Macdonald. *Hidden markov models for time series: an introduction using R*. CRC Press, 2009.

Matej Zvokelj, Samo Zupan, and Ivan Prebil. Non-linear multivariate and multiscale monitoring and signal denoising strategy using kernel principal component analysis combined with ensemble empirical mode decomposition method. *Mechanical Systems and Signal Processing*, 25(7):2631–2653, 2011. ISSN 0888-3270.

Appendices

Appendix A

Bearing Model Feature Definitions

A.1 Measures of central tendency

Measures of central tendency are summary statistics which are employed to describe a set of data by identifying the most "central" point within the dataset. As it is expected that degradation will cause high frequency accelerometer data to change, it may be the case that degradation causes a bias within the data. It is possible that this bias may be observed through the use of summary statistics describing the central location of a distribution of values. Due to this, five measures of central tendency are explored in this thesis, namely:

- Mean (arithmetic).

- Median.

- Mode.

- Tri-Mean.

- Root mean square (RMS).

- Winsorized Mean.

Although other measures of central tendency exist (such as the geometric mean or harmonic mean), these require strictly positive values and as such, high frequency bearing data is not well suited to these techniques. The five measures of central tendency explored within this thesis are defined below.

Mean (arithmetic)

The arithmetic mean (or the average) is the sum of the univariate data, divided by the number of points within the univariate data set. Formally, this can be defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{A.1}$$

Where $\bar{x}$ represents the estimation for the sample mean, n is the sample size (number of observations) and x_i is the ith observation of x.

Median

The median is more robust measure of central tendency than the arithmetic mean. As such, it is less sensitive to outliers and noise. As it is possible that degradation will

introduce noise or outlying points into the data, ensuring features are robust may aid the detection of the degradation. The median can formally be defined as:

$$\tilde{x} = \begin{cases} X_{(N+1)/2} & \text{if } N \text{ is odd} \\ \frac{1}{2}X_{N/2} + X_{1+(N/2)} & \text{if } N \text{ is even} \end{cases} \tag{A.2}$$

Where N represents the sample size, and X represents an ordered sample with $\min_x = X_1$ and $\max_x = X_N$.

Mode

The mode represents the most commonly occurring observation from a sample. Informally, this is the value at the "peak" of the probability density function of the empirical distribution. As such, it differs in notion from both the mean and the median. The mode is formally defined as:

$$x_{mode} = L_1 + \frac{f_2}{f_1 + f_2}c \tag{A.3}$$

Where L_1 is the class boundary of the lower modal class, f_1 is the frequency of the previous class to the modal class, f_2 is the frequency of the next class to the model class and c is the modal interval size.

Tri-Mean

The tri-mean is a more robust than the traditional sample mean, and as such, is more similar to the median than the sample mean. It is a measurement of central tendency (location) which is defined as a weighted average of the empirical distributions median and two quartiles. Formally, this is defined as:

$$x_{TM} = \frac{Q_1 + 2Q_2 + Q_3}{4} \tag{A.4}$$

Where Q_1 represents the first quartile (25th percentile), Q_2 the second quartile (median or 50th percentile) and Q_3 the third quartile (75th percentile). This combines the medians emphasis on central values, whilst acknowledging the extremes of the distribution.

Root Mean Square (quadratic mean)

The root mean square (RMS) or quadratic mean of a distribution is an alternative measure of central tendency that is commonly explored within electrical engineering,

and is useful when performing analysis of high frequency data containing sinusoidal functions (such as bearing data). Formally, the root mean square is defined as:

$$\bar{x}_{rms} = \sqrt{\frac{1}{n}x_1^2 + x_2^2 + \cdots + x_n^2} \tag{A.5}$$

Where n is the sample size.

The RMS is a measure of the magnitude of the attribute, and as such, it is hoped to be beneficial in the assistance of the identification of degradation.

Winsorized Mean

The final measure of central tendency explored within the context of this thesis is the Winsorized mean. This differs from the previous measures of central tendency in that this is more akin to a data cleaning and preprocessing technique. In this, k observations from both the lower and upper extreme ends of the probability distribution are replaced with the next remaining extreme value. Typically, k is defined in terms of n (the sample size), i.e, the 10% winsorized mean would replace all values from the 0 to 10th percentile and the 90th to 100th percentile with the values at the 10th and 90th percentile.

$$x_w = \frac{1}{n}(k+1)x_{(k+1)} \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)} \tag{A.6}$$

Where x_w represents the winsorized mean, n is the sample size, k is the number of observations to winsorize (typically expressed in terms of n) and x_i represents the ith observations of x.

A.2 Measures of variability

Measures of variability are summary statistics can be employed to describe a set of data by quantifying the variability of attributes within the dataset. As it is expected that degradation will cause the vibration data to change, it may be the case that degradation causes a change (either increased, or reduced) in the variability of an attribute. It is possible that this change may be observed through the use of summary statistics describing the variability of a distribution of values. Due to this, seven measures of variability are explored in this thesis for the purposes of bearing feature selection, namely:

- Standard deviation.

- Variance.

- Interquartile range (IQR).

- Range.

- Minimum.

- Maximum.

- Mean absolute deviation.

Although other measures of variability exist, a comparative analysis compromising all summary statistics of variability is beyond the scope of this thesis. As such, the seven identified measures of variability are explored as a basis of commonly occurring measures of variability, which it is anticipated that the reader will be familiar with. The seven measures of variability explored within this thesis are defined below.

Standard deviation

The first measure of variability employed as a feature is that of the standard deviation. This is a non-negative statistical measure of dispersion which quantifies the dispersion (variation) about the sample mean. The corrected sample standard deviation is defined as:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{x})^2} \tag{A.7}$$

Where $\hat{x}$ represents the sample mean, and x_i represents the ith data point of x. This estimator is a biased estimator of standard deviation for a sample. It is anticipated that changes in the data due to degradation will increase the variability within the data – such as that which is measured by the standard deviation – and as such, this feature may be beneficial to the identification of degradation.

Variance

The second measure of variability employed as a feature is that of variance. This is a non-negative statistical measure of dispersion which is closely linked to the standard deviation. This is also referred to as the second central moment. The unbiased sample variance is defined as:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{x})^2 \tag{A.8}$$

Where $\hat{x}$ represents the sample mean, and x_i represents the ith data point of x. This estimator of variance is an unbiased estimator of variance for a sample. As such, it is anticipated that the variance may be able to reveal more information regarding bearing degradation than the standard deviation.

Interquartile range

The interquartile range (IQR) is a measure of variability which quantifies this between the first and third quartiles. As such, it is defined as:

$$x_{IQR} = Q_{75} - Q_{25} \tag{A.9}$$

Where Q_3 represents the data at the third quantile (the 75th percentile) and Q_1 represents the data at the first quantile (the 25th percentile). For the purposes of this thesis, calculation of quantiles is performed as recommended by Hyndman and Fan (1996). That is, the index h of the a given percentile p can be defined as:

$$h = (N + 1/3)p + 1/3 \tag{A.10}$$

Where h represents the index of the ordered statistic, N represents the sample size and p the percentile in question (namely, 25 or 75). This index can then be utilised to identify the value of the percentile Q_p:

$$Q_p = x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor}) \tag{A.11}$$

Where h is the index identified in A.10, x_i is the ith data point of the ordered attribute x and $\lfloor x \rfloor$ represents the largest integer not greater than x.

As with both standard deviation and variance, it it anticipated that degradation on the bearing will change the underlying distribution. Thus, the interquartile range provides a further measure to quantify and explore this.

Range

The range represents a further means to quantify and explore bearing degradation. This is similar to the interquartile range, however, the range compromises the entire distribution. As such, the range is more sensitive to outlying and extreme data points than the interquartile range. If it is the case that degradation cases these outlying and extreme points, the range may provide a means to quantify the degradation. The range can be defined as:

$$x_{range} = (Q_{100} - Q_1) = \max(X) - \min(X) \tag{A.12}$$

Where x_{range} represents the range, and Q represents the quantile function defined in eq. A.11.

Minimum

The next measure of variability explored is the minimum value of the distribution. As the minimum will inherently be sensitive to outliers and extreme values, if these values correlate to degradation on the bearing, this feature would assist in the encapsulation of current degradation on the bearing. The minimum is defined as:

$$x_{min} := \{x_{min} | \forall x_k : x_{min} \leq x_k\} \tag{A.13}$$

Where x_k represents the kth element of x.

Maximum

Similar to the minimum, The next measure of variability explored is the maximum value of the distribution. As with the minimum, the maximum will also inherently be sensitive to outliers and extreme values. If the degradation is not observed by the minimum, the maximum provides a counter point to the minimum to assist in the identification of degradation. The maximum is defined as:

$$x_{max} := \{x_{max} | \forall x_k : x_{max} \geq x_k\} \tag{A.14}$$

Where x_k represents the kth element of x.

Median absolute deviation (MAD)

The final measure of variability explored within this context is that of the median absolute deviation. This is a robust measure of variability and builds upon the robust nature of the median. In brief, this summary statistic is the median of the absolute values of the observed deviations from the median. Formally, this is defined as:

$$x_{MAD} = median_i(|x_i - median_j(x_j)|) \tag{A.15}$$

Where x_i is the ith observation of x and the median is defined as in eq. A.2. This is employed as there may be cases where the variance of a distribution is undefined (as in the case of the Cauchy distribution).

A.3   Measures of shape

Measures of shape are summary statistics can be employed to describe the shape of a set of data by quantifying various parameters of the underlying distribution. As before, it is expected that degradation will cause the vibration data to change, it may be the case that degradation causes a change (either increased, or reduced) in the tails, peak or other aspect of the distribution. It is possible that this change may be observed through the use of summary statistics describing the shape of a distribution of values. Due to this, six measures of shape are explored in this thesis for the purposes of bearing feature selection, namely:

- Skewness.

- Kurtosis.

- Hyperskewness.

- Hyperflatness.

- Crest factor.

- Peak to average power ratio (PAPR).

Other measures of shape do exist, however, as before, a comparative analysis compromising all known summary statistics is beyond the scope of this thesis. The six measures of shape explored within this thesis for the purposes of bearing condition assessment provide a baseline from which other research can build upon. The six measures of shape explored within this thesis are defined below.

Skewness

Skewness is the first measure of shape explored in this thesis and represents the asymmetry of a distribution. A perfect normal distribution will have zero skewness, with negative skewness if the distribution "leans" to the right and positive skewness if it "leans" to the left.

Although there are many definitions of skewness, as explored in Joanes and Gill (1998), for the purposes of this thesis, sample skewness can be formally defined as:

$$G_1 = \frac{K_3}{K_2^{3/2}} = \frac{\sqrt{\{n(n-1)\}}}{n-2} g_1 \tag{A.16}$$

Where the coefficient g_1 is defined as (Joanes and Gill, 1998):

$$g_1 = \frac{m_3}{m_2^{3/2}} \tag{A.17}$$

Furthermore, m_k can be defined as (Joanes and Gill, 1998):

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k \tag{A.18}$$

If degradation on the bearing causes the distribution to "lean" to either side, this measure of shape will quantify this change and provide a basis for the assessment of condition. It should be noted that differences in the calculation of skewness will typically be insignificant for larger samples (i.e. $n > 100$).

Kurtosis

Similar to skewness (the third central moment) kurtosis is the fourth central moment. This quantifies the "peakedness" of the underlying distribution. Typically, a normal distribution will have a kurtosis value of zero (mesokurtic), if the peak of the distribution is excessive, this positive kurtosis is referred to as lepokurtic, and similarly, if the peak is lower than anticipated as compared to a normal distribution this negative kurtosis is referred to as platykurtic. Kurtosis is typically positive if the tails of the distribution are heavier than expected from a normal distribution, and similarly negative if the tails are lighter than expected from a normal distribution.

As skewness has many definitions, so does kurtosis. For the purposes of this thesis, the sample kurtosis is defined as (Cramér, 1999):

$$G_2 = \frac{K_4}{K_2^2} = \frac{(n-1)}{(n-2)(n-3)}\{(n+1)g_2 + 6\} \tag{A.19}$$

Where the coefficient g_2 is defined as (Cramér, 1999):

$$g_2 = \frac{m_4}{m_2^2} - 3 \tag{A.20}$$

Coefficients m_4 and m_2 are defined as per eq. A.18.

If degradation on the bearing causes the distribution to tails to either increase or decrease with respect to the normal distribution, this measure of shape will quantify this change and provide a basis for the assessment of condition. It should be noted that like skewness, differences in the calculation of kurtosis will typically be insignificant for larger samples (i.e. $n > 100$).

Hyperskewness

The third measure of shape explored within this thesis for bearing condition assessment is that of hyperskewness. This is another higher order moment (the 5th moment), however, is often under utilised as it can be difficult to interpret. The 5th central moment can be defined as:

$$\mu_5 = \frac{\sum_{x=1}^n (x - \bar{x})^5 f(x)}{\sum_{x=1}^n f(x)} \tag{A.21}$$

Where $f(x)$ represents the cumulative probability distribution function. Although this statistic may be difficult to interpret, it is anticipated that if degradation is not observed by the four other moments utilised, hyperskewness will provide an alternative means to identify and quantify degradation.

Hyperflatness

The fourth measure of shape explored within this thesis for bearing condition assessment is that of hyperflatness. This is the next higher order moment (the 6th moment). As with hyperskewness, is often under utilised as it can be difficult to interpret. The 6th central moment can be defined as:

$$\mu_6 = \frac{\sum_{x=1}^n (x - \bar{x})^6 f(x)}{\sum_{x=1}^n f(x)} \tag{A.22}$$

Where $f(x)$ represents the cumulative probability distribution function. Although this statistic may be difficult to interpret, it is anticipated that if degradation is not observed by the four other moments utilised, hyperskewness will provide an alternative means to identify and quantify degradation.

Crest factor

Moving away from higher order moments, the next statistical feature explored for bearing degradation assessment is that of the crest factor. The crest factor is a measure of how extreme peaks within the data are, and is defined as:

$$x_{CF} = \frac{|x|_{peak}}{x_{rms}} \tag{A.23}$$

Where $|x|_{peak}$ represents the absolute peak (absolute maximal value) within the data, and x_{rms} represents the quadratic mean, as defined in eq. A.5. As degradation may be caused by shocks to the system, by exploring the relationship between the

maximal point and the quadratic mean, it is hoped that this can be quantified and correlated to bearing degradation.

Peak to average power ratio (PAPR)

The peak to average power ratio (PAPR) is closely related to the crest factor and can be defined as:

$$x_{PAPR} = \frac{|x|^2_{peak}}{x^2_{rms}} = CF^2 \tag{A.24}$$

This expresses the peak amplitude as power (maximal amplitude squared), and expresses this power in terms of the average power (average amplitude – i.e. RMS – squared).

Similarly to the crest factor, it is anticipated that shocks within the data will be observed within this features and as such, the peak to average power ratio will provide a means to quantify these shocks. This would enable quantification of the current state of the bearing.

A.4   Measures of position

Measures of position are statistics which can be employed to quantify the value given a specific position within the underlying distribution. As has been throughout this chapter, it is expected that degradation will cause the vibration data to change, it may be the case that degradation causes a change (either increased, or reduced) in the sides of the distribution. It is possible that these changes may be observed through the use of summary statistics describing the value at a position within a distribution. Due to this, four measures of shape are explored in this thesis for the purposes of bearing feature selection, namely:

- 10th percentile.

- 25th percentile.

- 75th percentile.

- 90th percentile.

The four chosen percentiles were chosen for various reasons. The 10th and 90th percentiles were chosen due to their position close to the lower and upper ends of the distribution. If degradation causes extreme values and outliers to occur, it is anticipated

that these positions will reflect these changes, whilst being more sensitive to the changes than the other positions explored. Similarly, the 25th and 75th percentiles provides a similar quantification, but for the main body of the distribution. If degradation causes a contaminated distribution, it is anticipated that these may change to reflect this. Measures of position are explored in conjunction with measures of central tendency, measures of variability and measures of shape for various reasons. Mainly, it is possible for measures of central tendency to change (such as the arithmetic mean or median) without changes to various positions within the distribution. This is also the case for measures of variability such as the minimum and maximum, and for measures of shape such as skewness and kurtosis. By sampling a range of percentiles across the distribution, it is hoped that any changes to the underlying structure of the data are captured to be exploited for the purposes of condition assessment.

10th percentile

The first measure of position explored within the context of this chapter is the 10th percentile. The technique used within this thesis to identify the required percentile has already been defined in section A.2. For completeness, it is given here again:

$$h = (N + 1/3)p + 1/3 \tag{A.25}$$

Where h represents the index of the ordered statistic, N represents the sample size and p the percentile in question (in this case, 10). This index can then be utilised to identify the value of the percentile Q_p:

$$Q_p = x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor}) \tag{A.26}$$

Where h is the identified index , x_i is the ith data point of the ordered attribute x and $\lfloor x \rfloor$ represents the largest integer not greater than x.

25th percentile

The 25th percentile is the second measure of position explored within this chapter for the purposes of bearing condition assessment. This can be defined as above, with:

$$p = 25 \tag{A.27}$$

As this is a more inlying position within the distribution than the 10th percentile, it should be more robust to extreme or outlying values, whilst still accurately representing changes to the position of the distribution to the left of the centre.

75th percentile

The 75th percentile is the penultimate measure of position explored within this chapter for the purposes of bearing condition assessment. This can be defined as above, with:

$$p = 75 \tag{A.28}$$

This is the natural counterpoint to the 25th percentile, and as such may provide additional information in a multivariate context if changes between these attributes are asymmetric. Due to its position, this percentile should be more robust to the extreme or outlying values encountered by the 90th percentile.

90th percentile

The 90th percentile is the final measure of position explored within this chapter for the purposes of bearing condition assessment. This can be defined as above, with:

$$p = 90 \tag{A.29}$$

This is the natural counterpoint to the 10th percentile, and as such may provide additional information in a multivariate context if changes between these attributes are asymmetric. Due to its position, this percentile should be more sensitive to the extreme or outlying values encountered by the 75th percentile, allowing for the exploration of changes between these attributes to be taken into consideration.

A.5 Measures of impurity

Measures of impurity are differ from the previous attributes in that these measures estimate the contamination (on impurity) of an underlying distribution. To motive the exploration of these measures, take the following example. Degradation is an unpredictable process, and given that measures of impurity such as entropy quantify this unpredictability, this provides a possible means to quantify degradation. Entropy traditionally refers to the information content within a signal. If normal operation behaviour of the bearing contains a certain quantity of information, it is possible that the addition of degradation will change (either increase, or potentially decrease) the information content within the signal. As such, measures of impurity should be explored for their ability to assist with condition assessment. Due to this, five measures of impurity are explored in this thesis for the purposes of bearing feature selection, namely:

- Shannon entropy.

- Chao and Shen entropy.

- Krichevsky-Trofimov entropy.

- Miller Madow entropy.

- James-Stein Shrinkage entropy.

Although five entropy functions are explored within the context of bearing condition assessment, other measures of impurity exist (such as the Gini index). A comparative evaluation of all measures of impurity is beyond the scope of this thesis, and the subset of features selected in this case is performed to motivate the potential of measures of impurity for the purposes of bearing degradation assessment. For the purposes of this thesis, all measures of entropy are given in nats (natural logarithm).

Shannon entropy

The first measure of impurity explored within the context of this chapter is that of Shannon entropy. Shannon entropy was introduced in 1948 by Claude Shannon (Shannon, 2001). Shannon entropy is defined as:

$$H_s = -\sum_{i=1}^{n} p(x_i) \ln p(x_i) \tag{A.30}$$

Where px_i represents the probability of x_i occurring within the discrete variable x. As vibration data containing normal operational behaviour will contain a given quantity of information, it is anticipated that this quantity will change should another source of information contaminate the signal (such as degraded behaviour).

Chao-Shen entropy

The entropy measure defined by (Chao and Shen, 2003) builds upon the notion of Shannon entropy. In their measure, the Horvitz-Thompson (Horvitz and Thompson, 1952) adjustment is employed in order to enhance the estimation of entropy. This adjustment ensures sufficient coverage of the underlying distribution whilst taking into consideration missing information due to potential sampling flaws. The Chao-Shen estimator of entropy is defined as:

$$H_{CS} = -\sum_{i=1}^{s} \frac{\hat{C}\tilde{\pi}_i \log(\hat{C}\tilde{\pi}_i)}{1 - (1 - \hat{C}\tilde{\pi}_i)^n} I(A_i) \tag{A.31}$$

The full details of this and their derivation are beyond the scope of this thesis, but are presented in detail in Chao and Shen (2003).

Krichevsky Trofimov entropy

Krichevsky-Trofimov (KT) entropy (Krichevsky and Trofimov, 1981) – denoted as $_{KT}$ – is a modification of Shannon entropy in that Bayesian techniques are employed. In this case, the bin probabilities are superseded by the Bayesian estimates of frequencies. Within this family of entropy estimators, a parameter is required to specify the (pseudo) count per frequency bin. In this case, Jeffreys prior is employed. This is defined as the square root of the determinant of the Fisher information criterion (Krichevsky and Trofimov, 1981).

The estimator explored in this case is the Krichevsky-Trofimov (KT) estimator. The full details of this estimator are beyond the scope of this thesis, but are presented in Krichevsky and Trofimov (1981) and in Hausser and Strimmer (2009).

Miller-Madow entropy

Miller Madow entropy is again a modification to the traditional Shannon entropy. In this case, the Miller Madow correction is applied to the observed counts for the computation of Shannon entropy, to provide a bias-corrected empirical entropy estimate. In this case, given the Shannon entropy H, the Miller-Madow correction is defined as:

$$H_{MM} = H + \frac{\hat{m} - 1}{2N} \tag{A.32}$$

Where N represents the sample size, and $\hat{m}$ represents the estimate of the number of bins with a non-zero probability. This is discussed in detail in Paninski (2003).

James-Stein Shrinkage entropy

The final measure of impurity and final feature explored in this chapter for the purposes of bearing condition assessment is that of a James-Stein shrinkage estimator for entropy. This is defined as per Hausser and Strimmer (2009):

$$H_{shrink} = -\sum_{i=1}^{n} \hat{\theta}_i^{shrink} \ln(\hat{\theta}_i^{shrink}) \tag{A.33}$$

Where $\hat{\theta}_i^{shrink}$ is defined as per Hausser and Strimmer (2009):

$$\hat{\theta}_i^{shrink} = \lambda t_k + (1 - \lambda)\hat{\theta}_i^{ML} \tag{A.34}$$

Where λ represents the shrinkage intensity (between 0 and 1), t_k represents the shrinkage target and $\hat{\theta}_i^{ML}$ represents the maximum likelihood estimation of entropy

(i.e. empirically derived Shannon entropy). The full details of this estimator and its derivation can be found in Hausser and Strimmer (2009).

Appendix B

Bearing Model Results

B.1   Section 4.2 results - Measures of central tendency
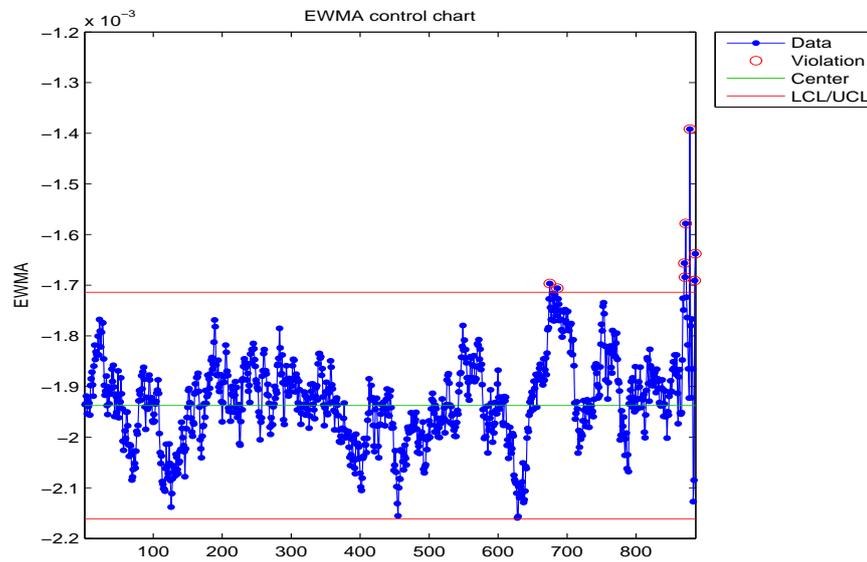

Figure B.1: EWMA chart for the mean of the failed bearing.
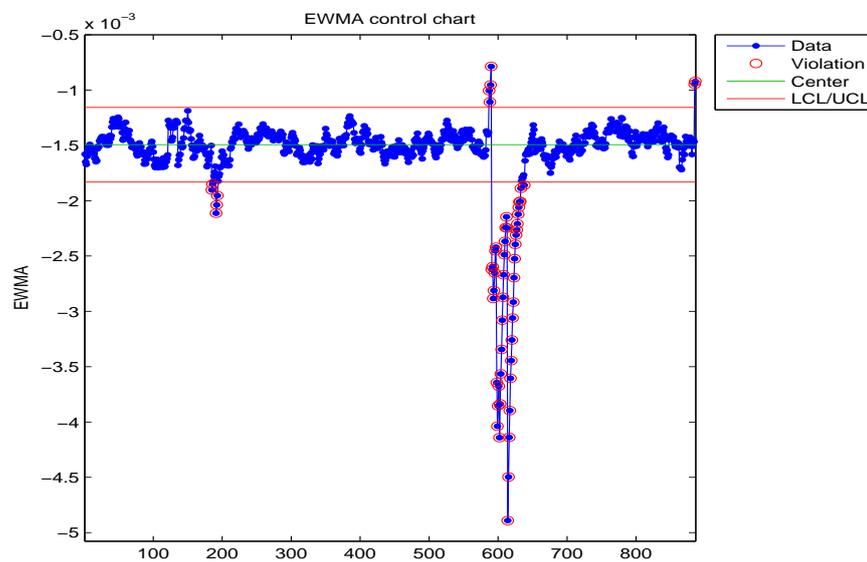
Figure B.2: EWMA chart for the mean of the healthy bearing.
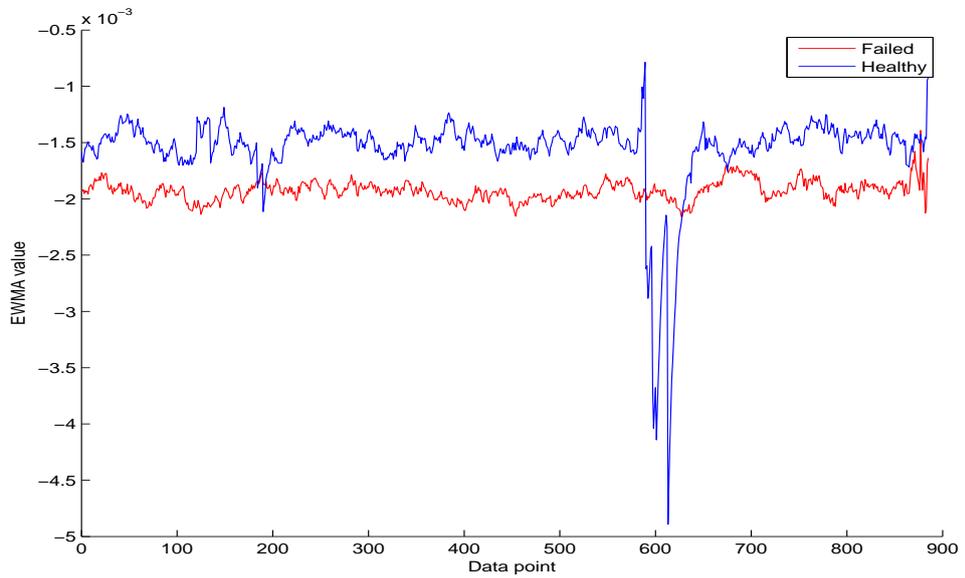
Figure B.3: Mean over time of the failed and healthy bearing.
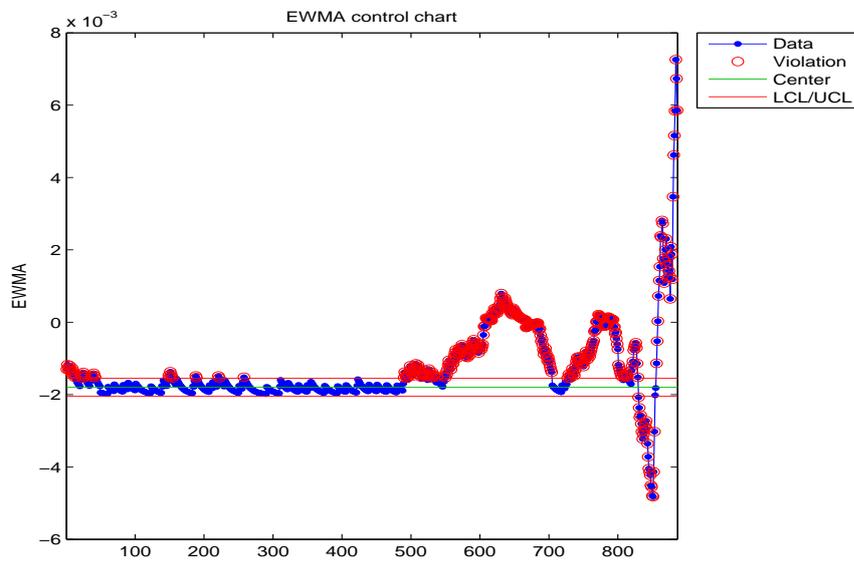
Figure B.4: EWMA chart for the median of the failed bearing.
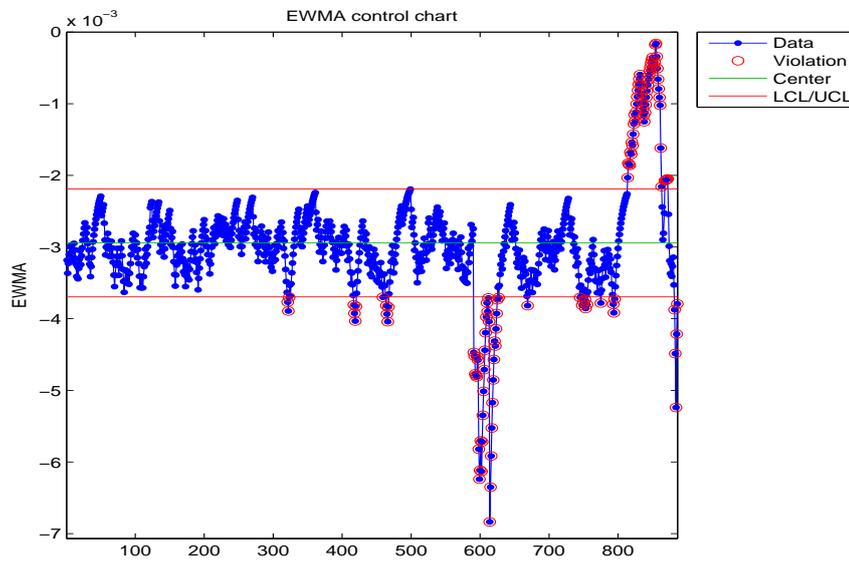
Figure B.5: EWMA chart for the median of the healthy bearing.
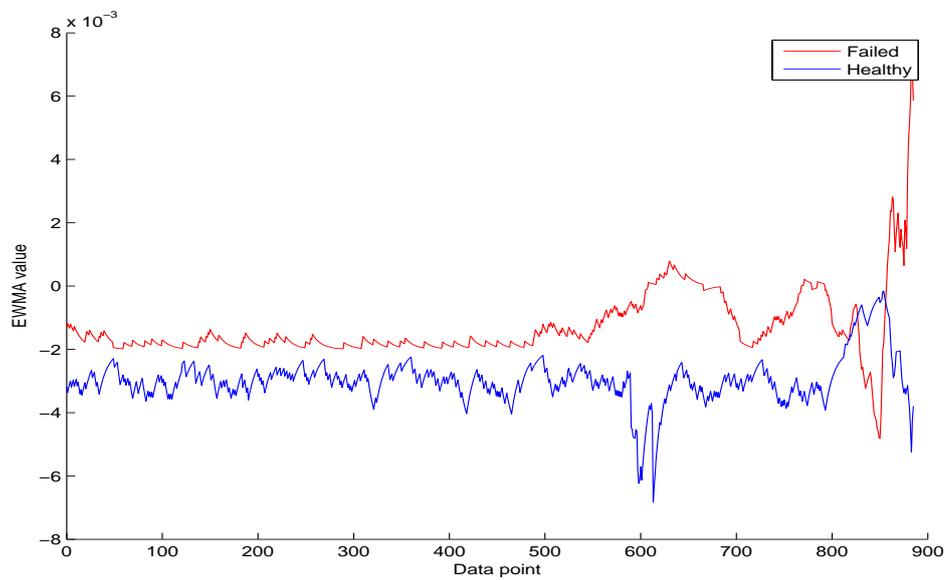
Figure B.6: Median over time of the failed and healthy bearing.
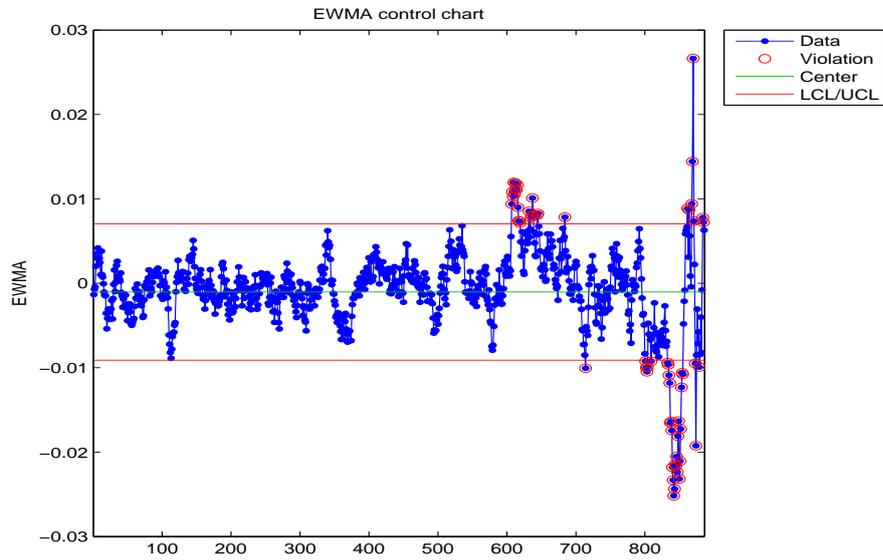
Figure B.7: EWMA chart for the mode of the failed bearing.
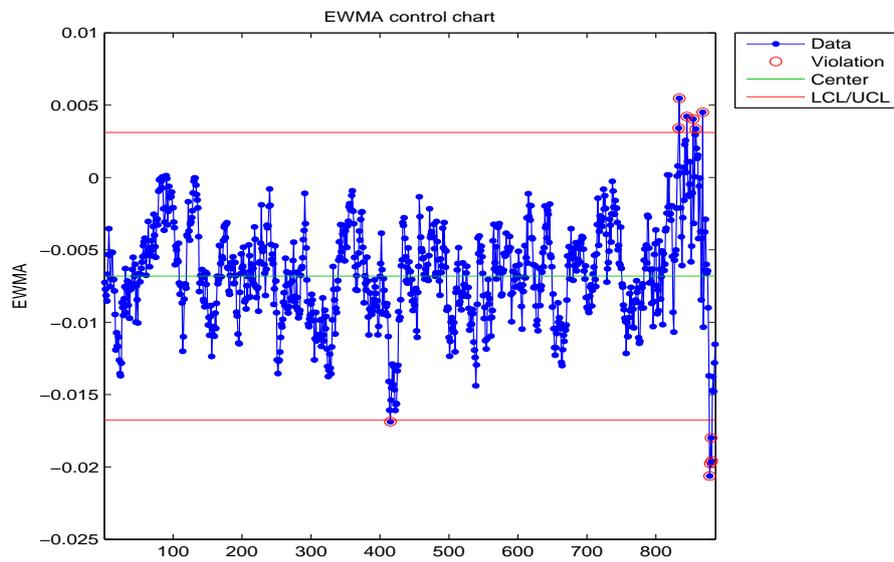

Figure B.8: EWMA chart for the mode of the healthy bearing.
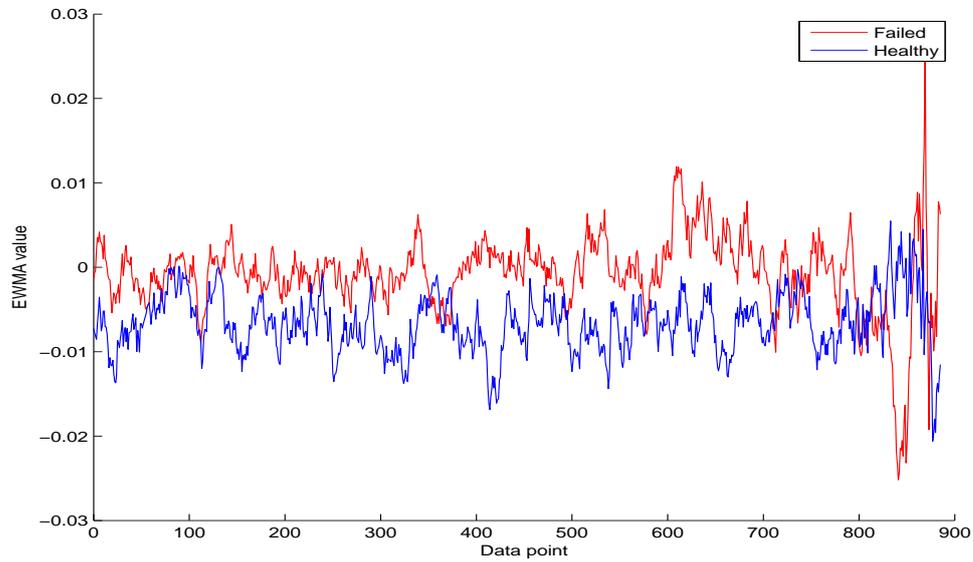
Figure B.9: Mode over time of the failed and healthy bearing.
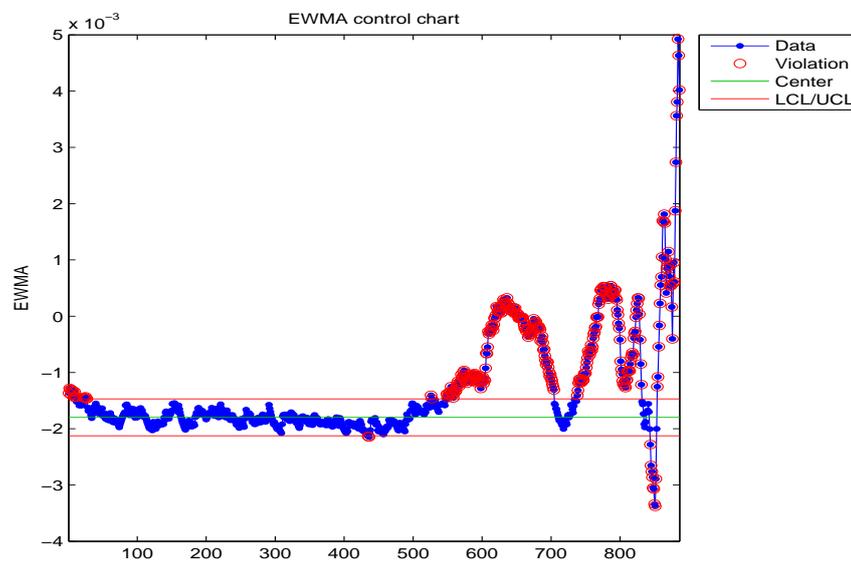
Figure B.10: EWMA chart for the tri-mean of the failed bearing.
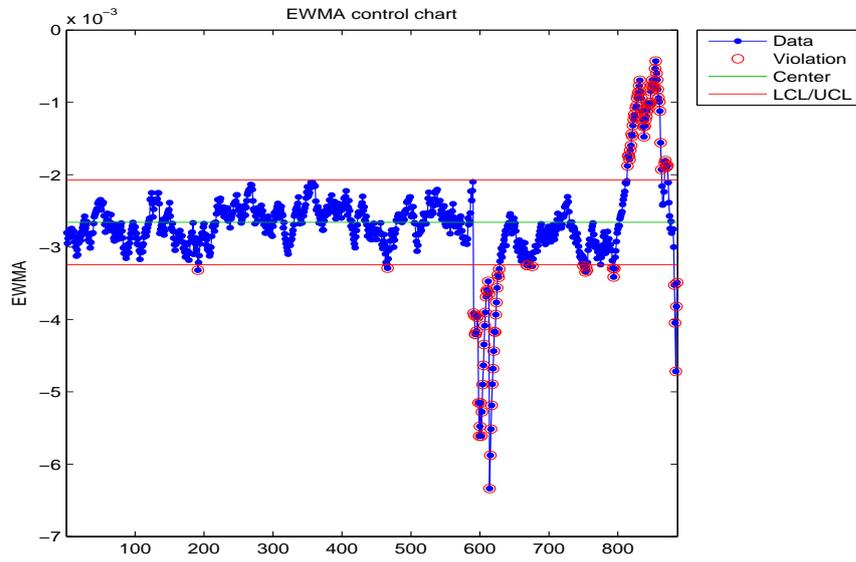
Figure B.11: EWMA chart for the tri-mean of the healthy bearing.
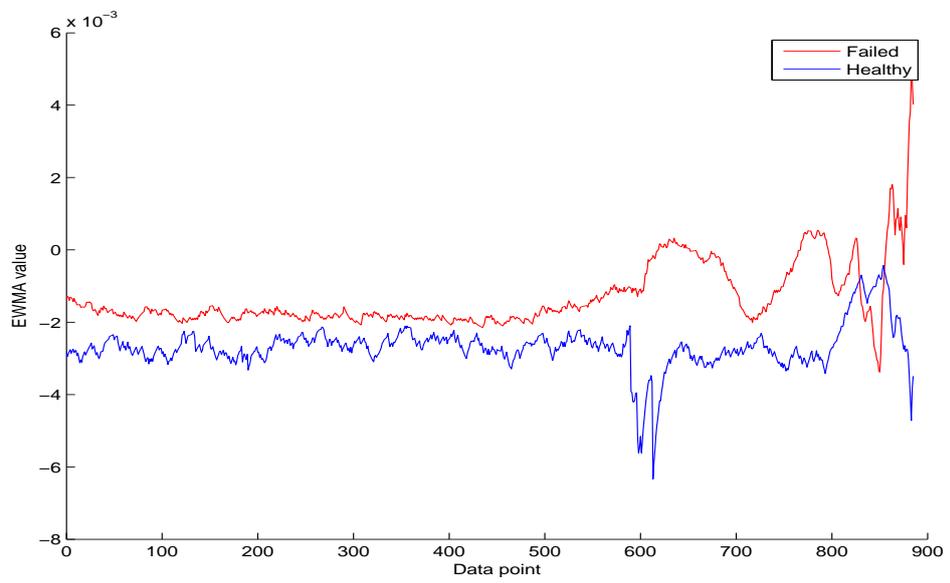
Figure B.12: Tri-mean over time of the failed and healthy bearing.
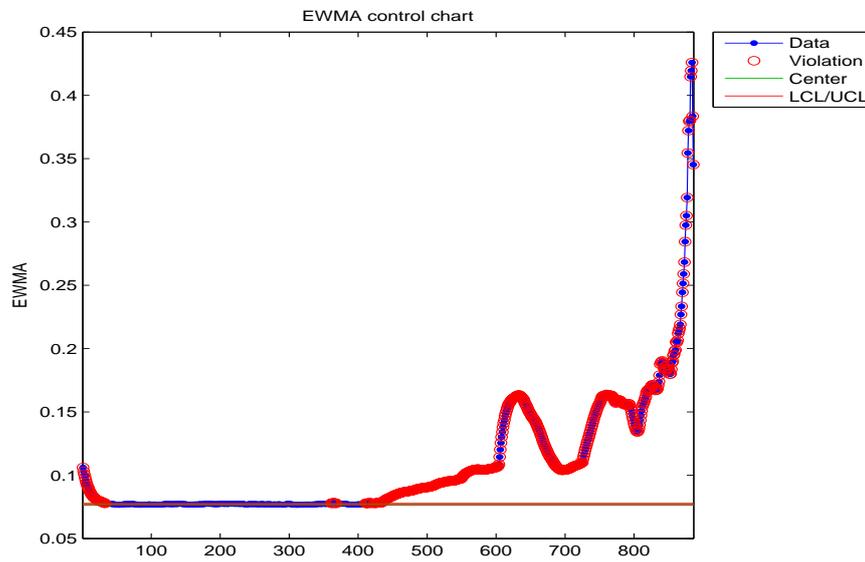
Figure B.13: EWMA chart for the RMS of the failed bearing.
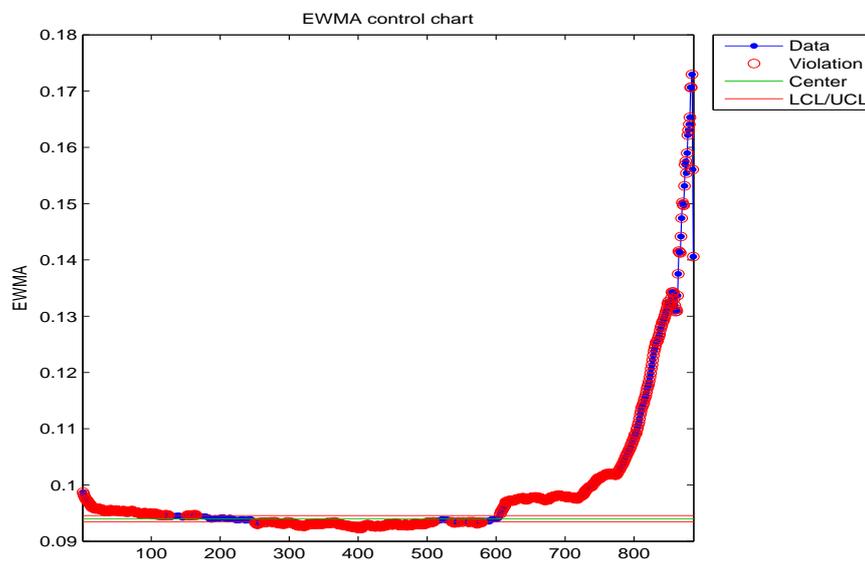
Figure B.14: EWMA chart for the RMS of the healthy bearing.
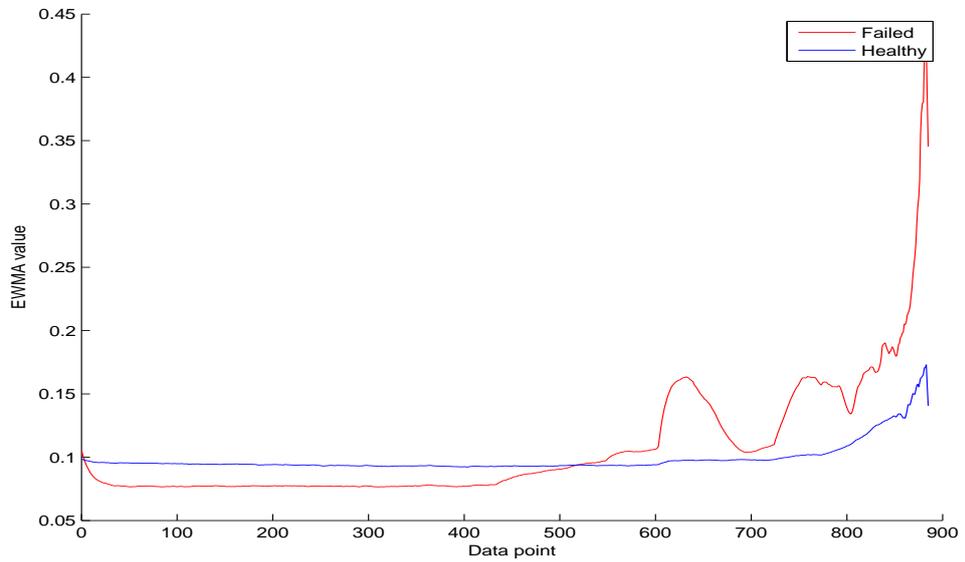
Figure B.15: RMS over time of the failed and healthy bearing.
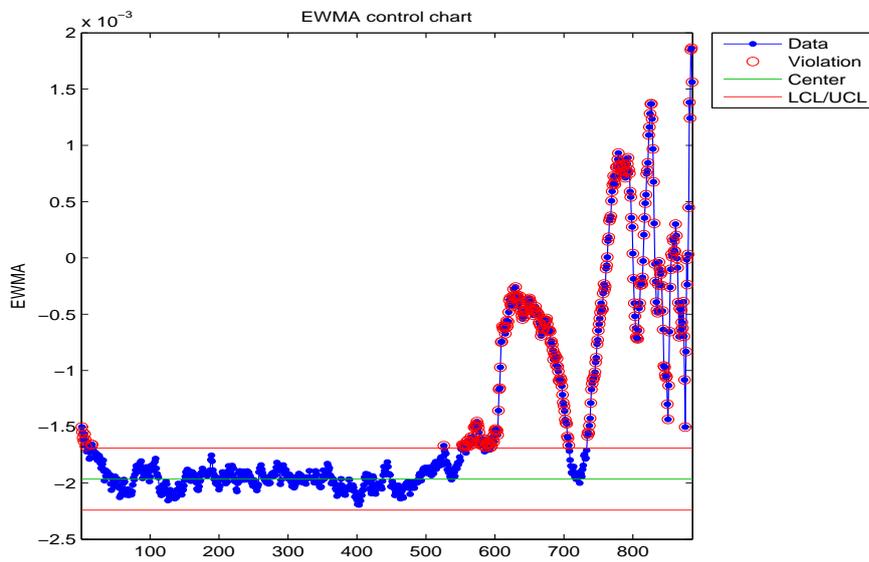
Figure B.16: EWMA chart for the winsorized wean of the failed bearing.

Figure B.17: EWMA chart for the winsorized mean of the healthy bearing.
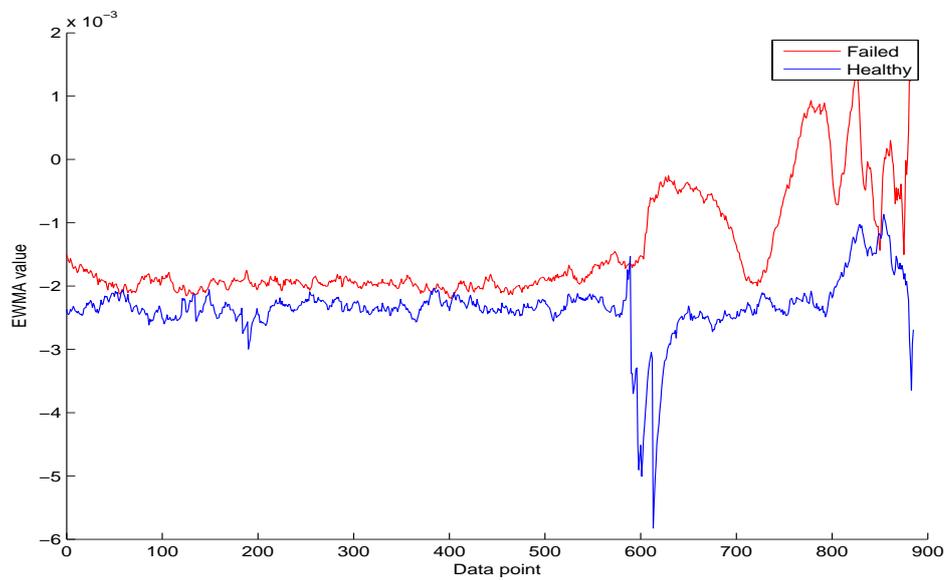
Figure B.18: Winsorized mean over time of the failed and healthy bearing.

B.2 Section 4.2 results - Measures of variability
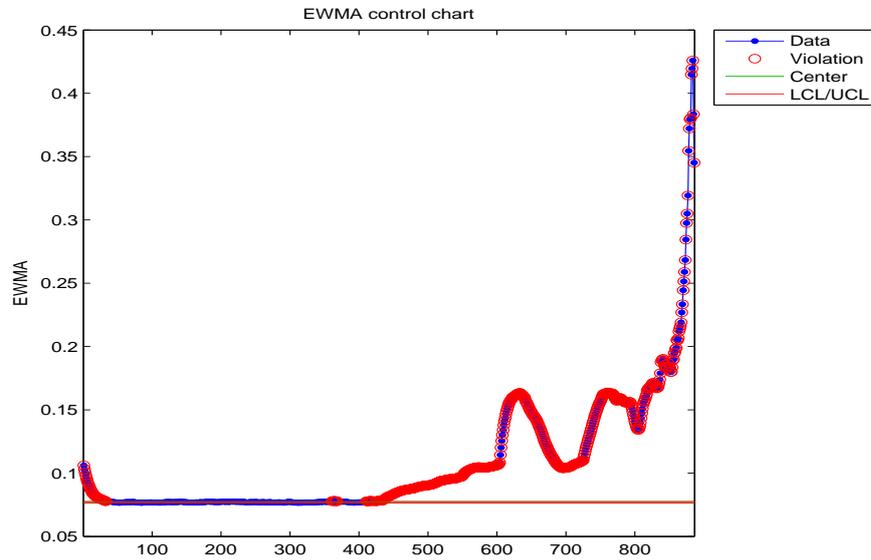
Figure B.19: EWMA chart for the standard deviation of the failed bearing.
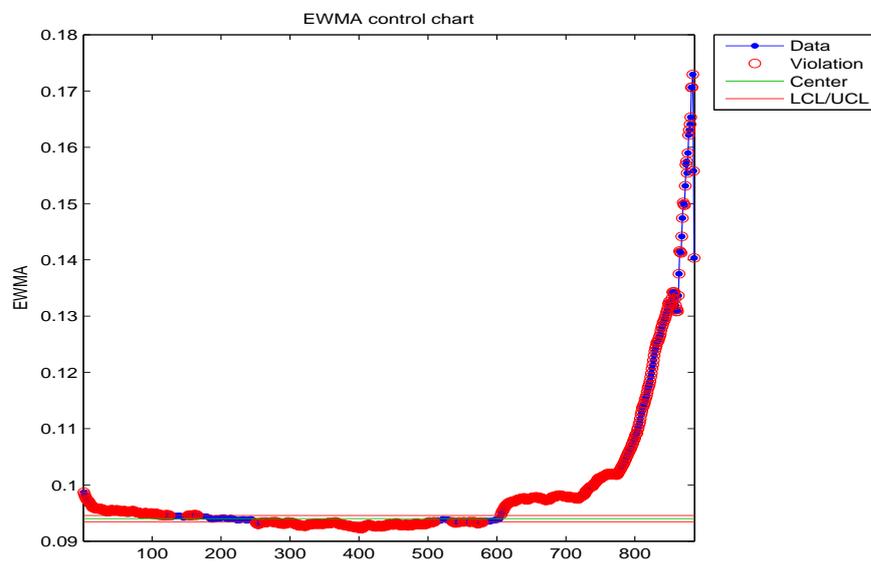
Figure B.20: EWMA chart for the standard deviation of the healthy bearing.
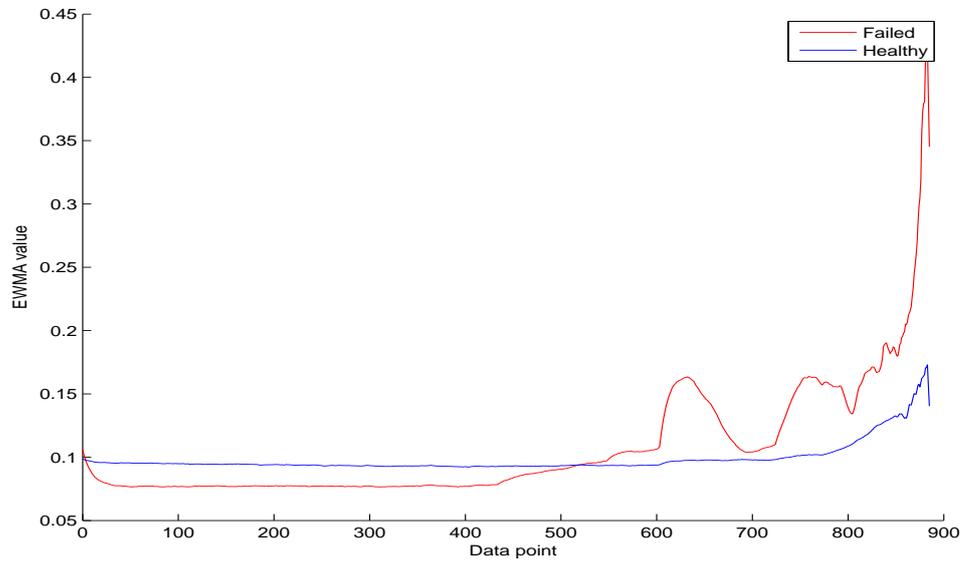
Figure B.21: Standard deviation over time of the failed and healthy bearing.
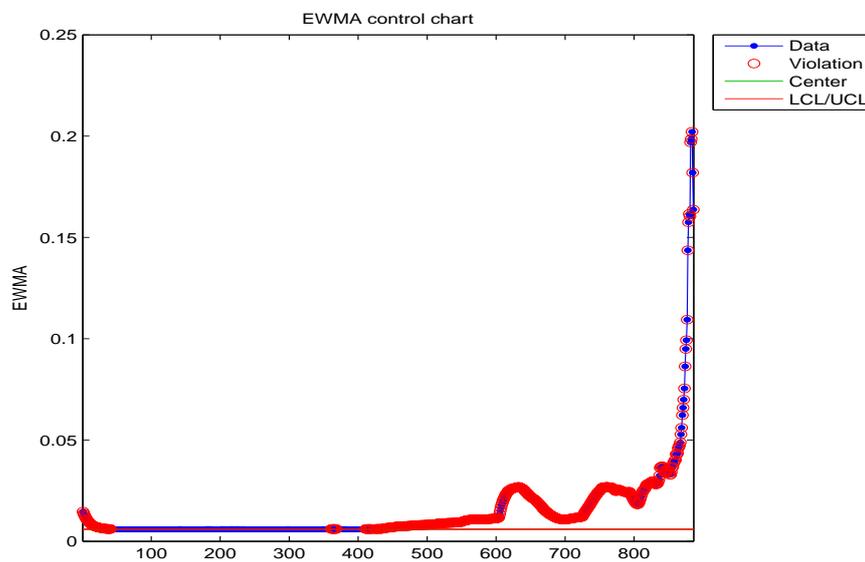
Figure B.22: EWMA chart for the variance of the failed bearing.
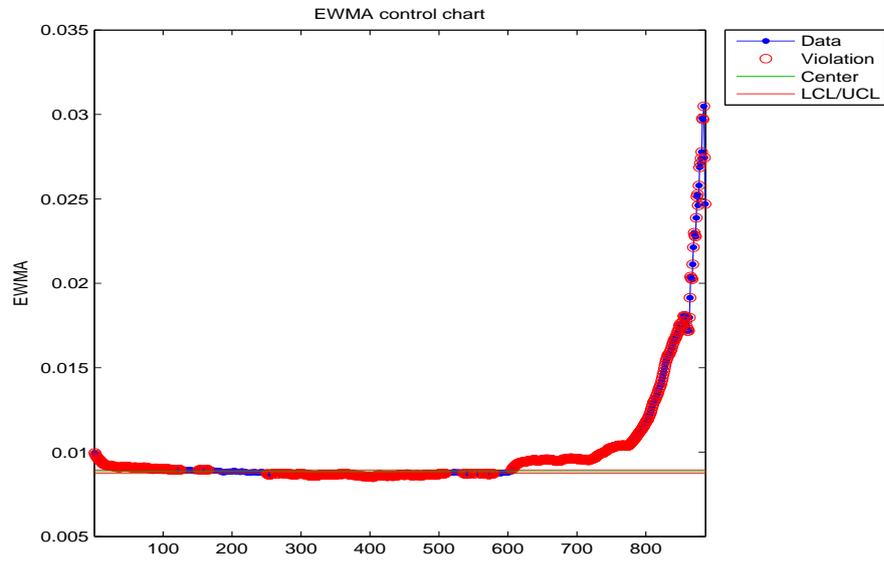
Figure B.23: EWMA chart for the variance of the healthy bearing.
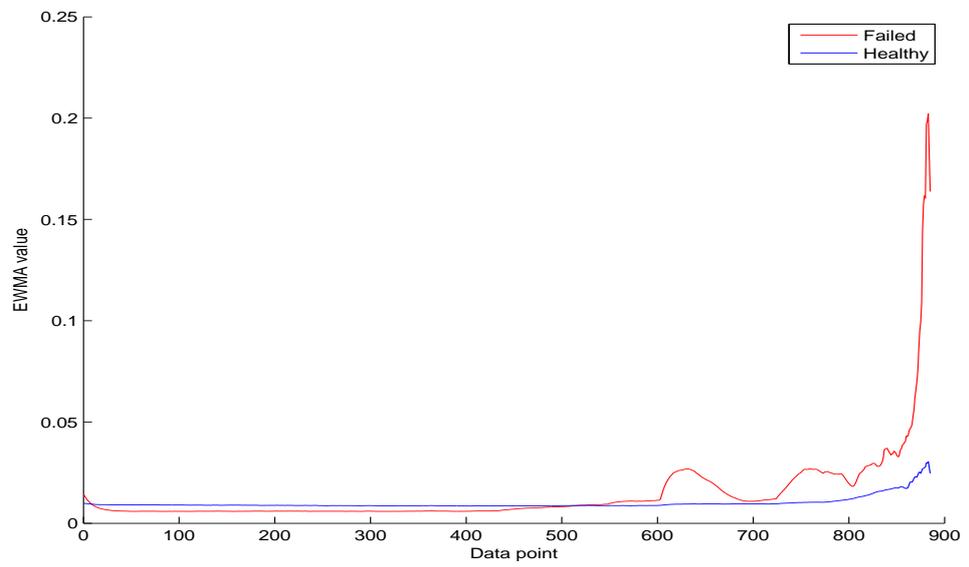

Figure B.24: Variance over time of the failed and healthy bearing.
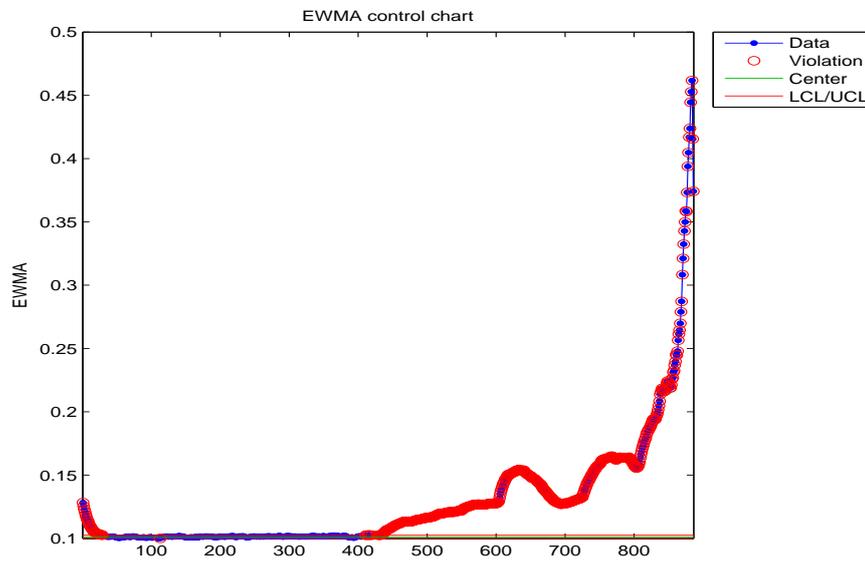
Figure B.25: EWMA chart for the IQR of the failed bearing.
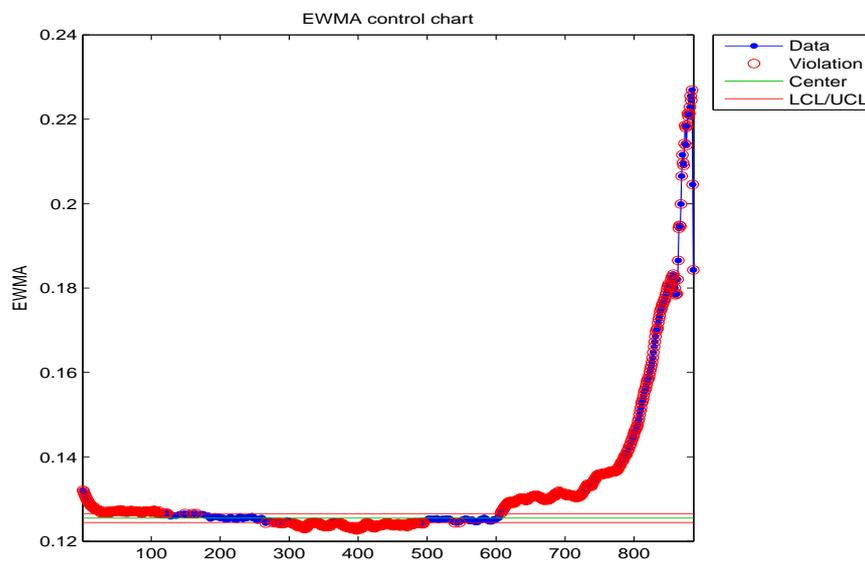
Figure B.26: EWMA chart for the IQR of the healthy bearing.
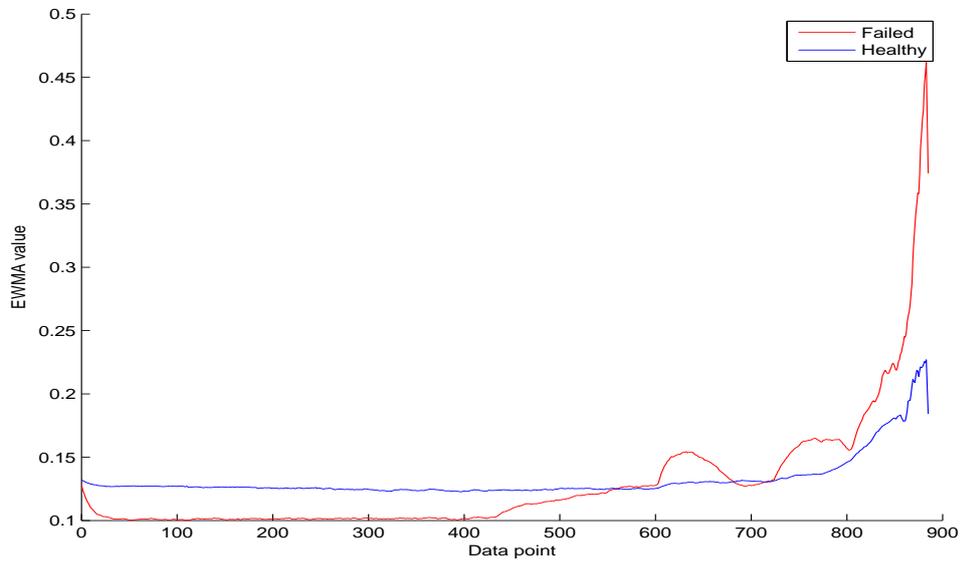
Figure B.27: IQR over time of the failed and healthy bearing.
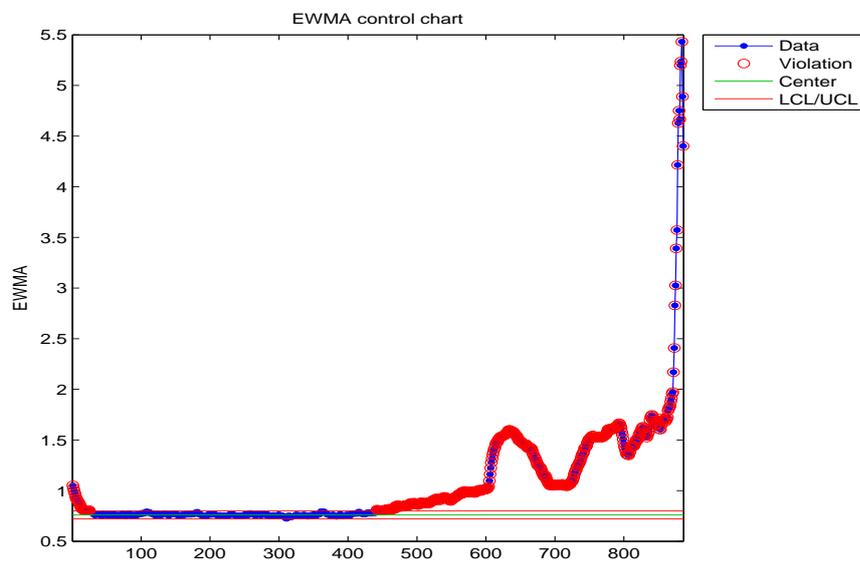
Figure B.28: EWMA chart for the range of the failed bearing.
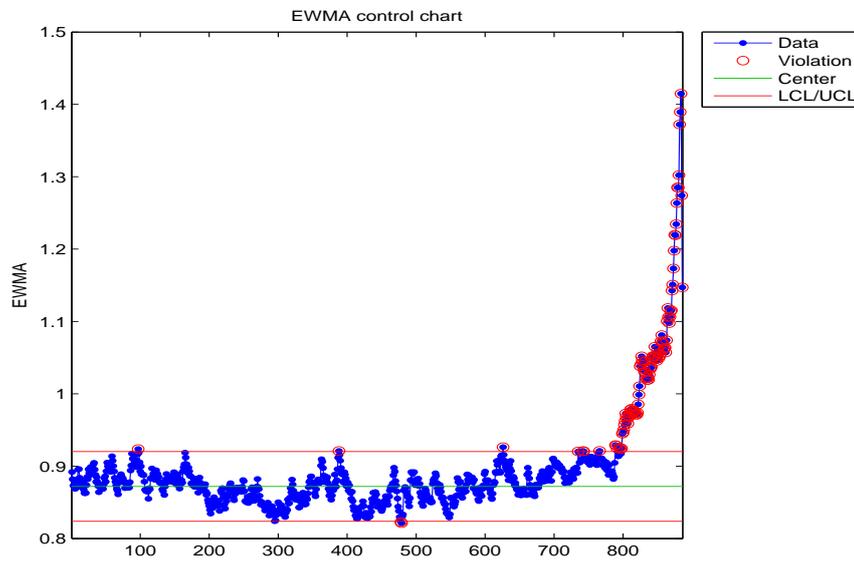
Figure B.29: EWMA chart for the range of the healthy bearing.
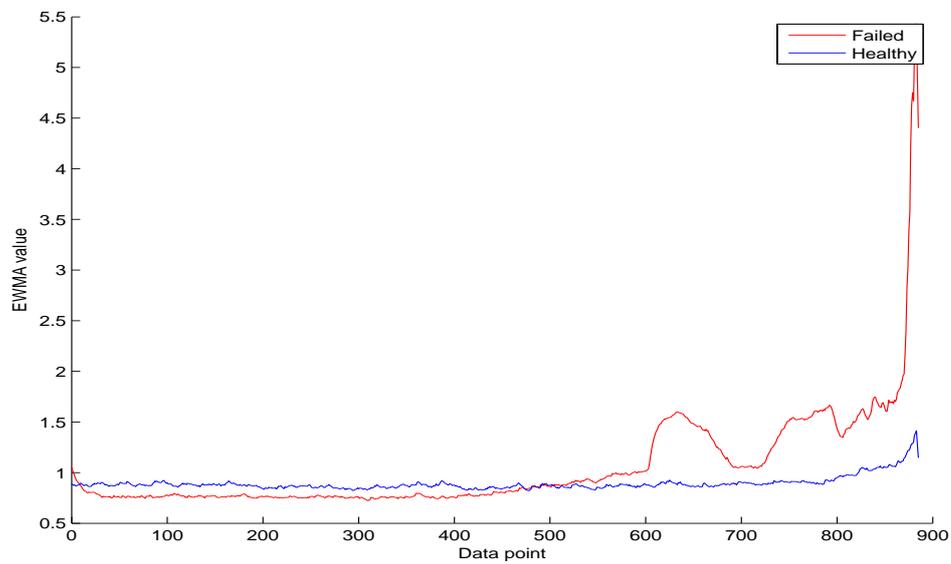
Figure B.30: Range over time of the failed and healthy bearing.
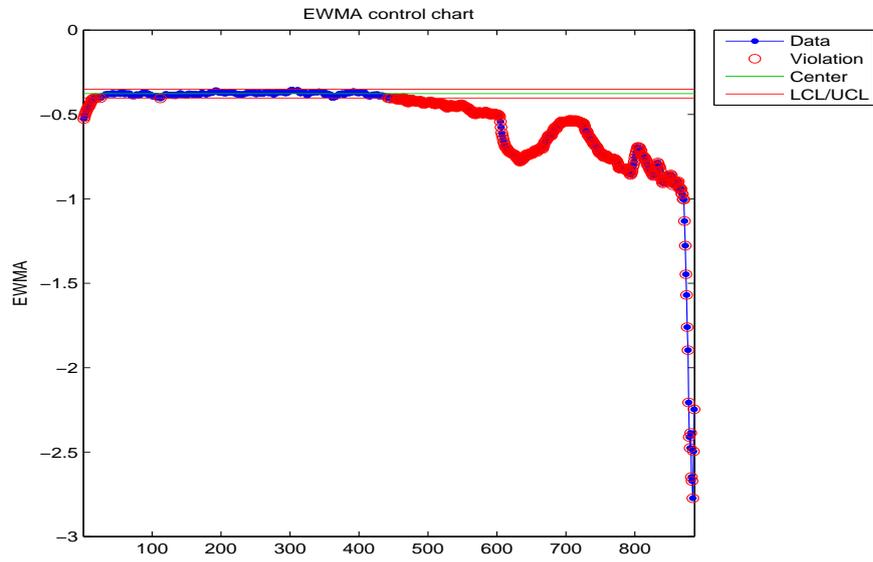
Figure B.31: EWMA chart for the minimum of the failed bearing.
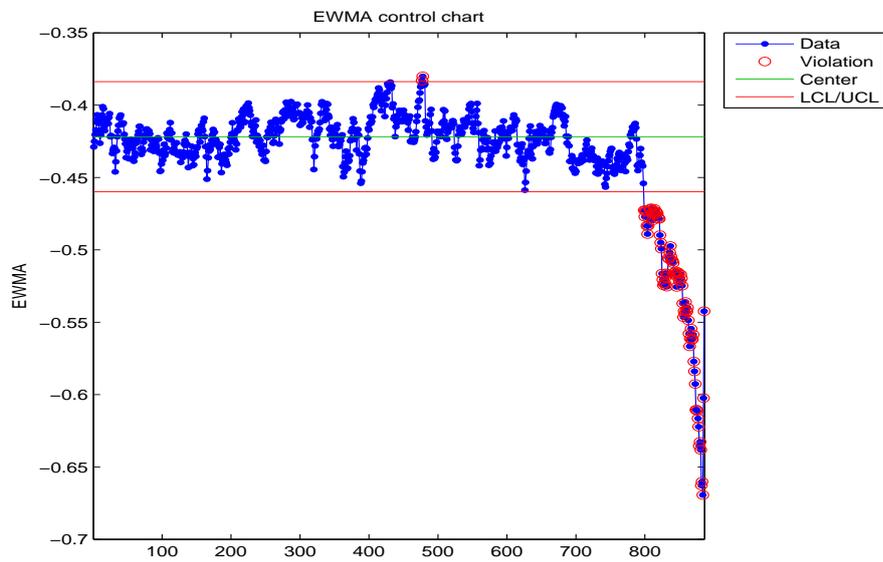
Figure B.32: EWMA chart for the minimum of the healthy bearing.
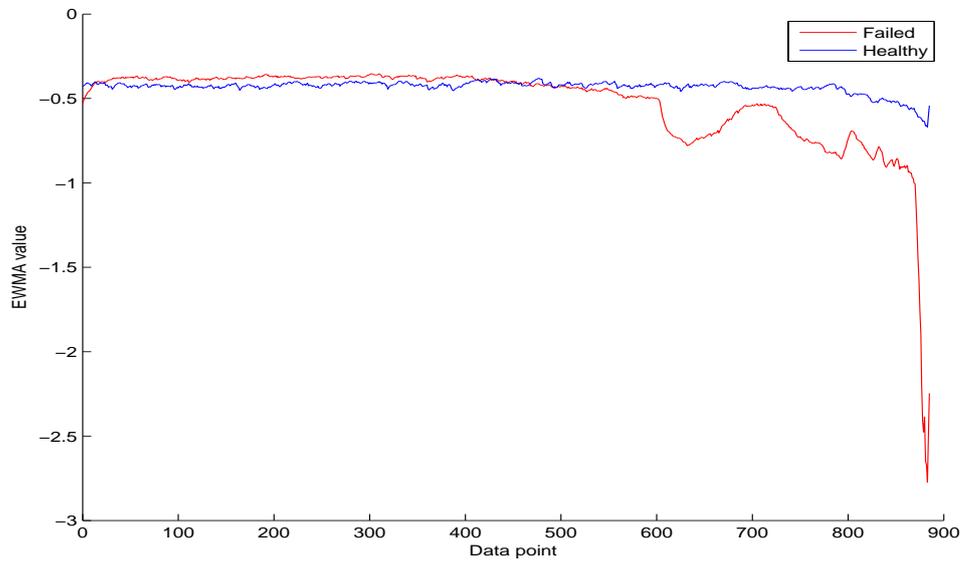
Figure B.33: Minimum over time of the failed and healthy bearing.
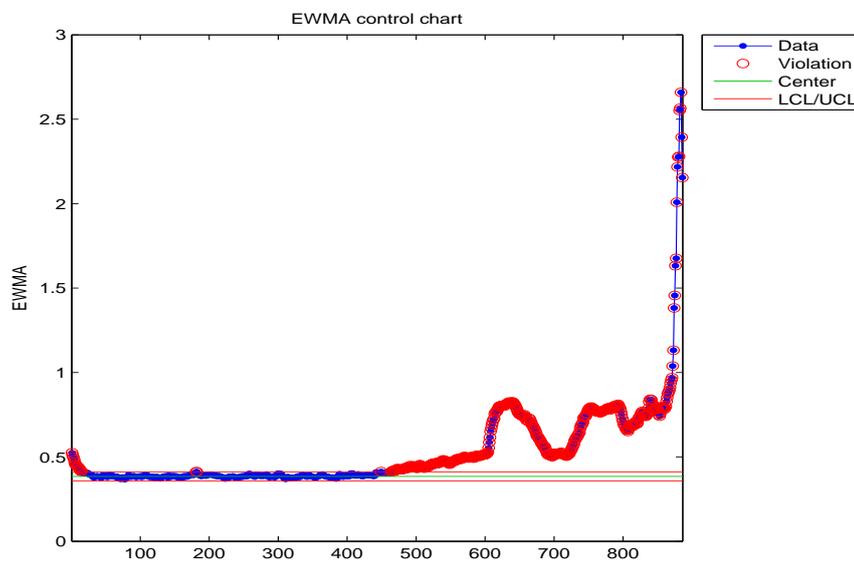
Figure B.34: EWMA chart for the maximum of the failed bearing.
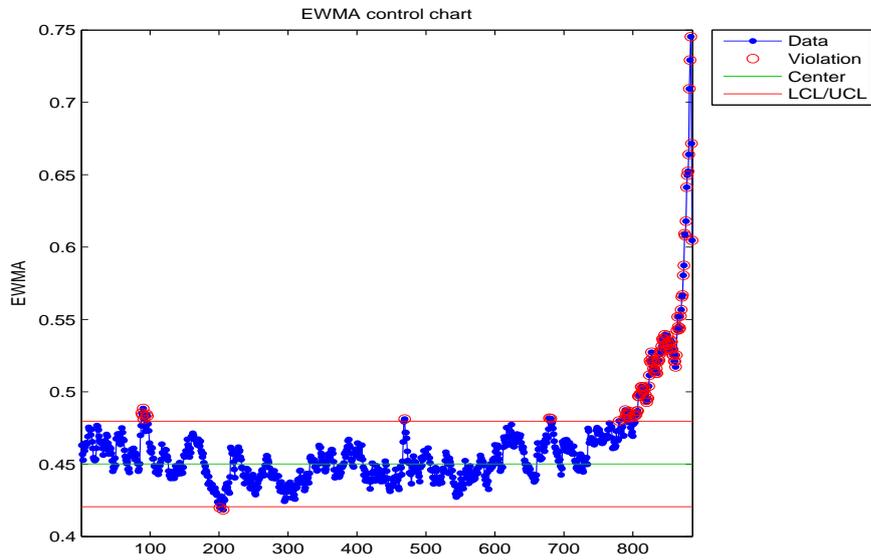
Figure B.35: EWMA chart for the maximum of the healthy bearing.
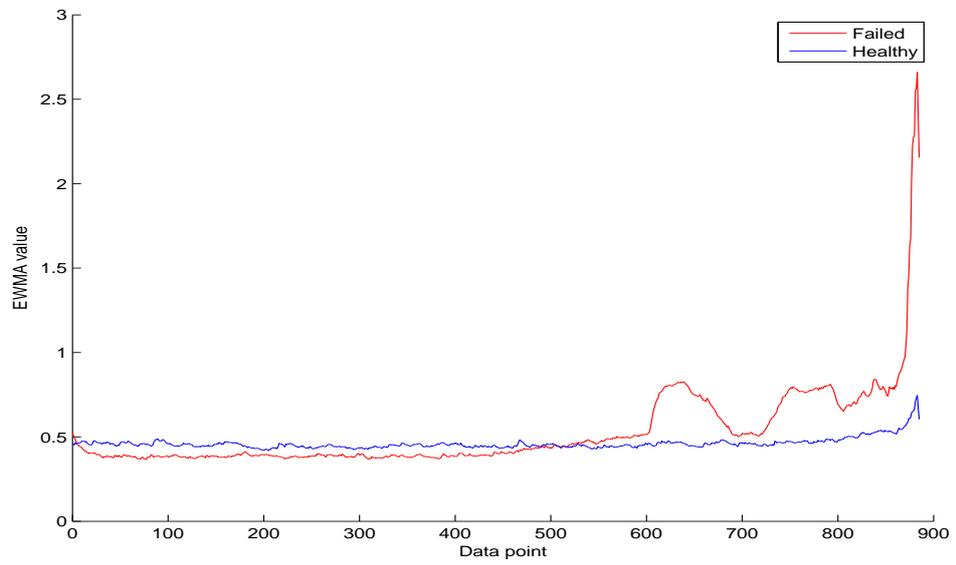
Figure B.36: Maximum over time of the failed and healthy bearing.
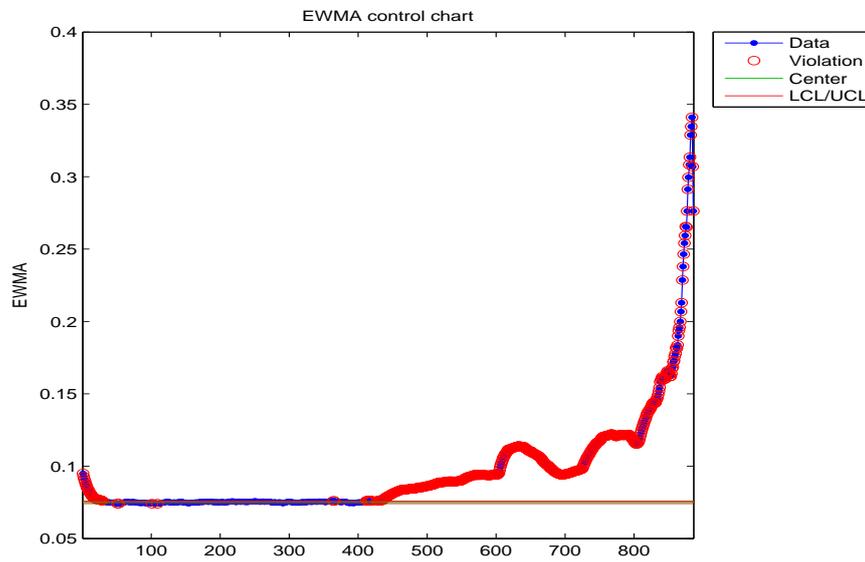
Figure B.37: EWMA chart for the median absolute deviation of the failed bearing.
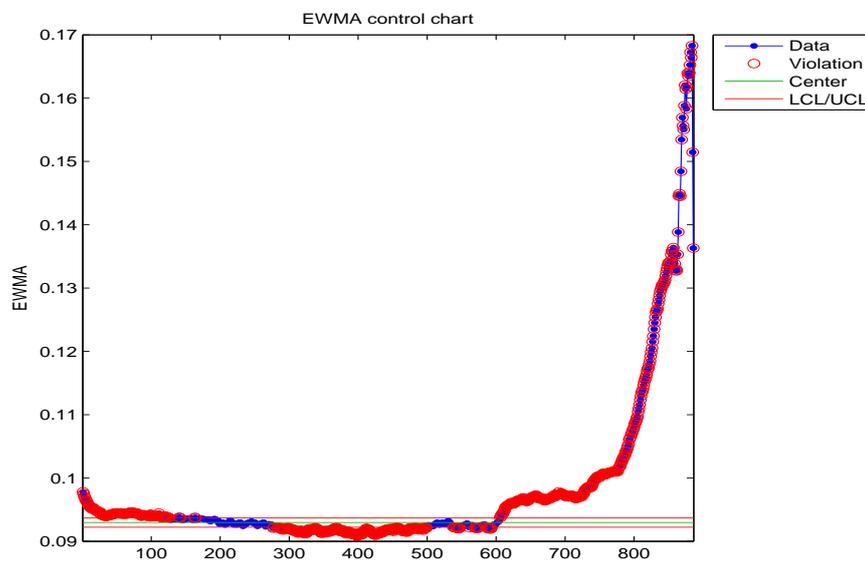
Figure B.38: EWMA chart for the median absolute deviation of the healthy bearing.
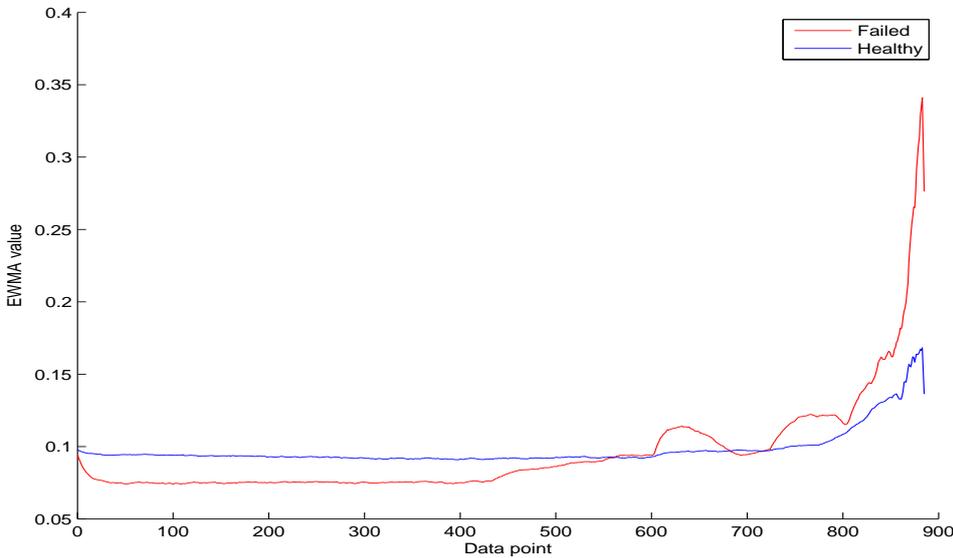
Figure B.39: Median absolute deviation over time of the failed and healthy bearing.

B.3   Section 4.2 results - Measures of shape
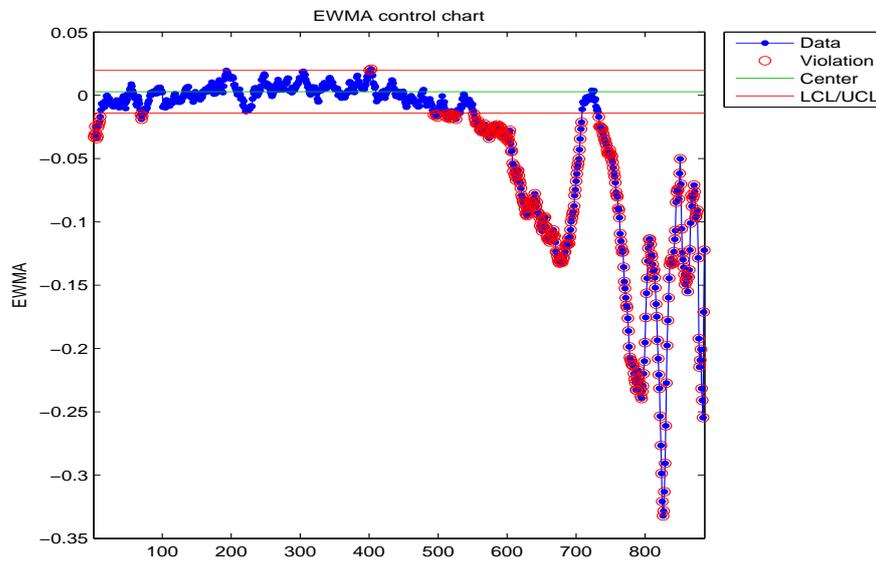

Figure B.40: EWMA chart for the skewness of the failed bearing.
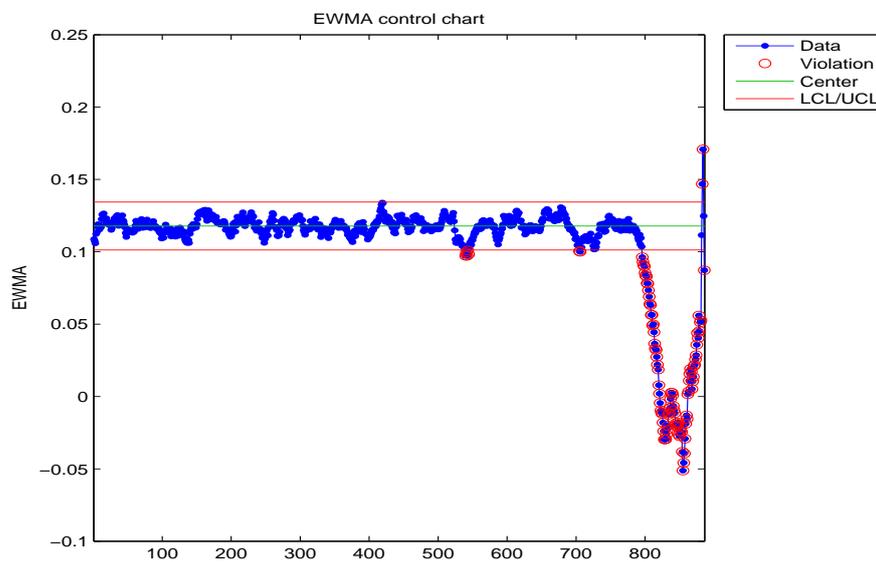
Figure B.41: EWMA chart for the skewness of the healthy bearing.
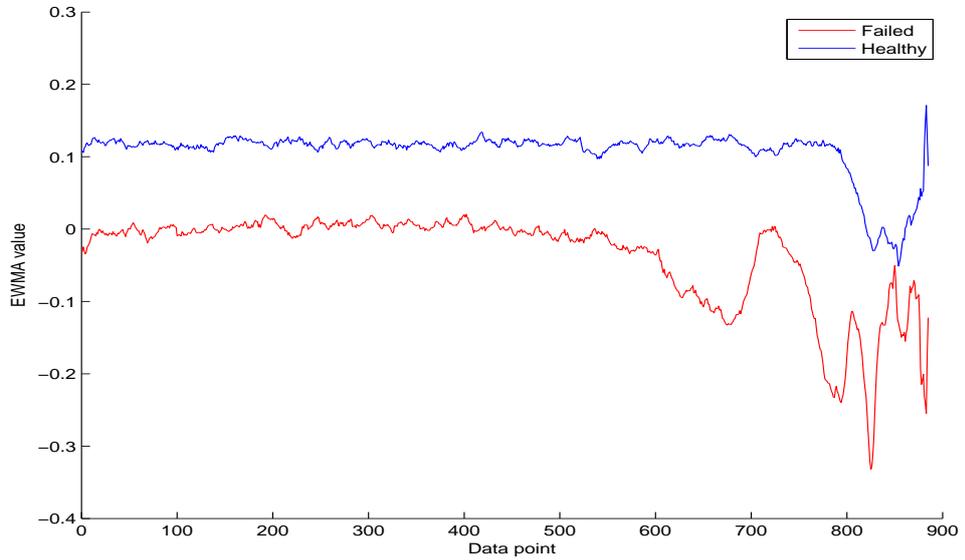
Figure B.42: Skewness over time of the failed and healthy bearing.
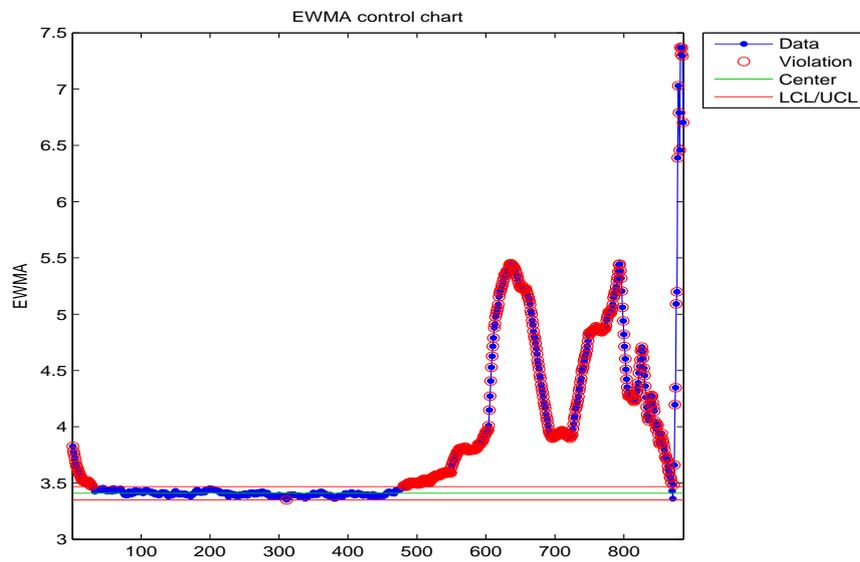
Figure B.43: EWMA chart for the kurtosis of the failed bearing.
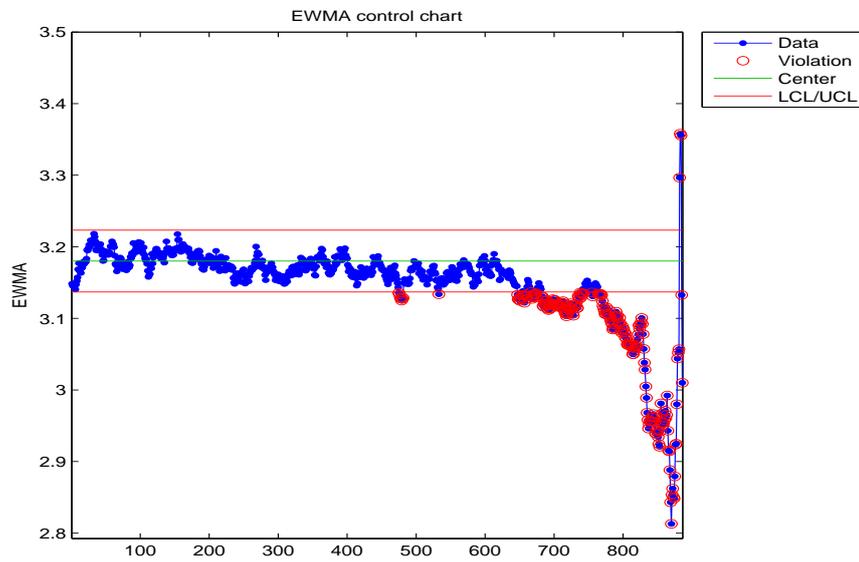
Figure B.44: EWMA chart for the kurtosis of the healthy bearing.
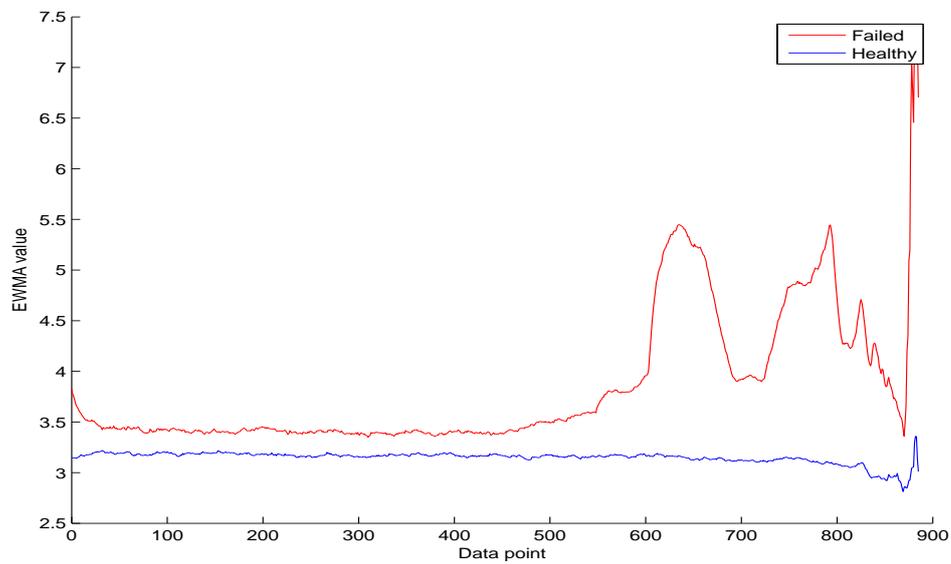
Figure B.45: Kurtosis over time of the failed and healthy bearing.
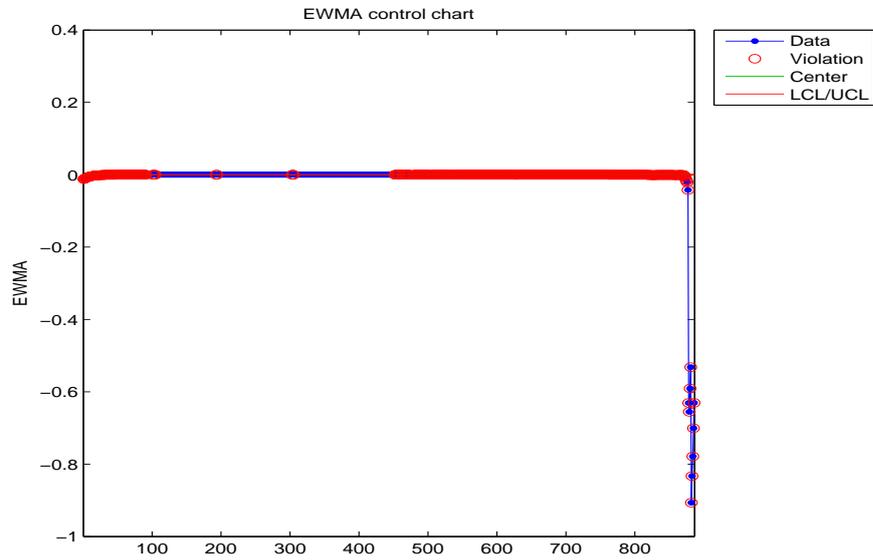
Figure B.46: EWMA chart for the hyperskewness of the failed bearing.
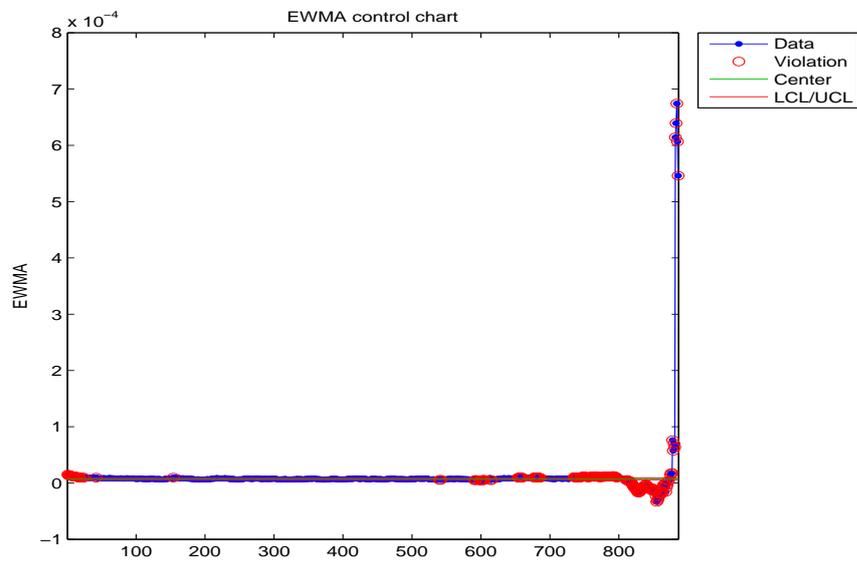
Figure B.47: EWMA chart for the hyperskewness of the healthy bearing.
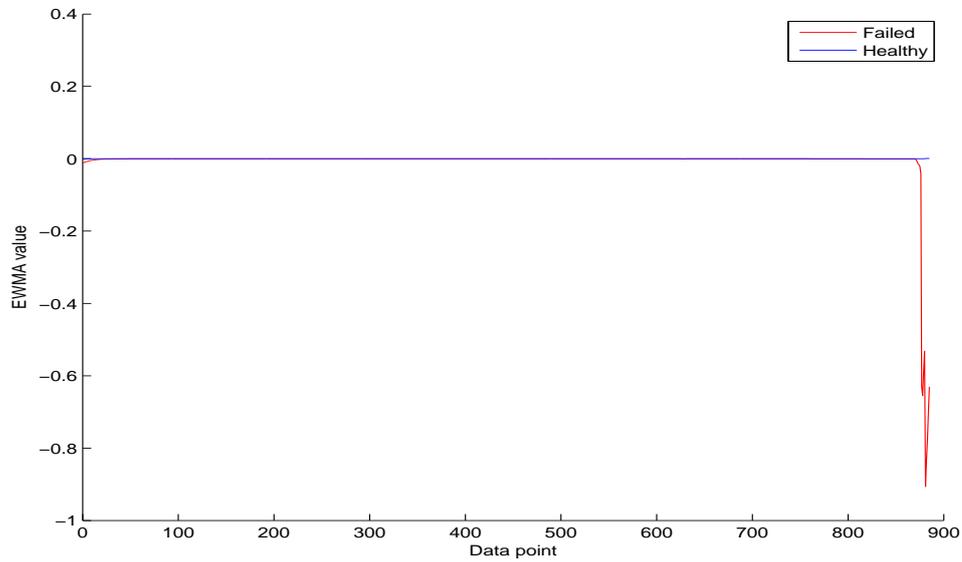
Figure B.48: Hyperskewness over time of the failed and healthy bearing.
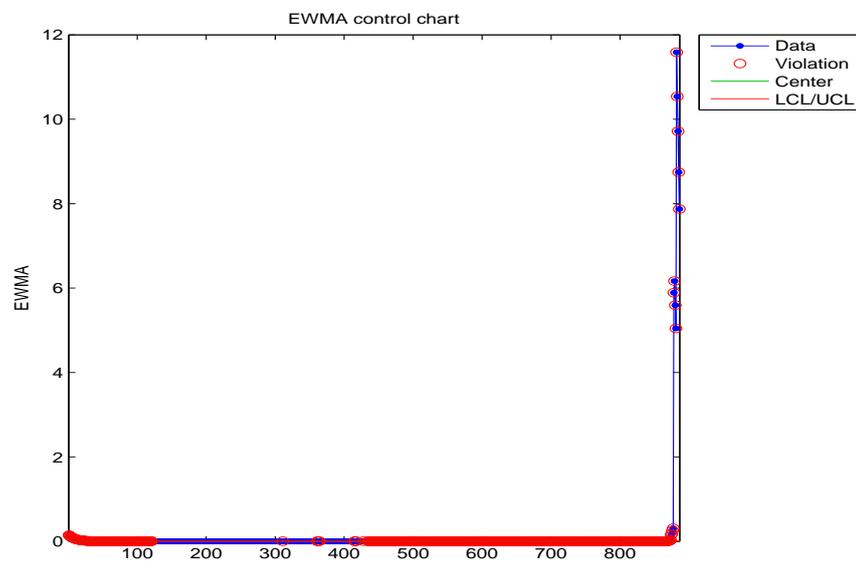
Figure B.49: EWMA chart for the hyperflatness of the failed bearing.
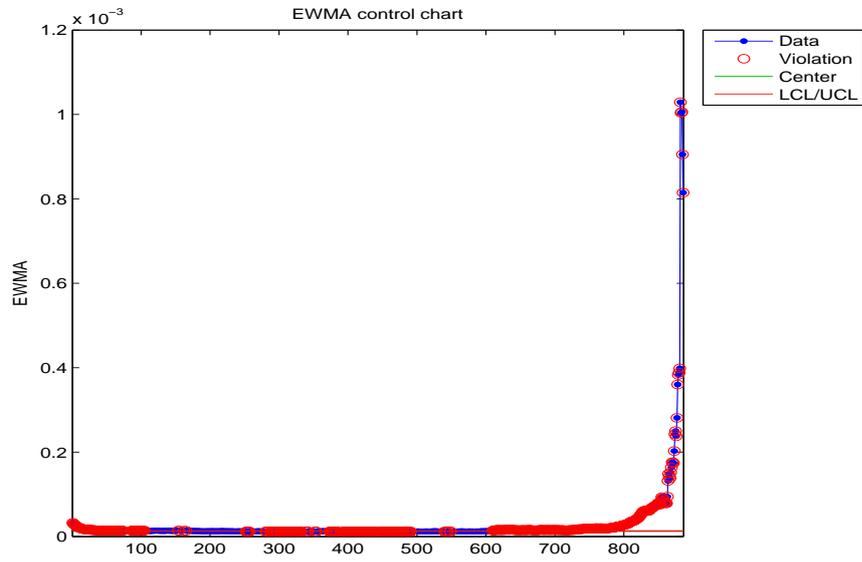
Figure B.50: EWMA chart for the hyperflatness of the healthy bearing.
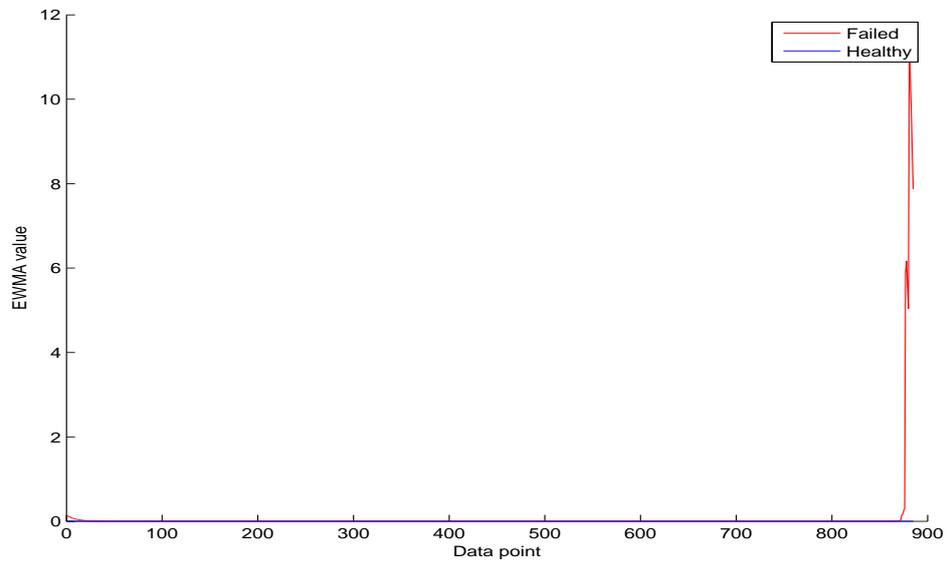
Figure B.51: Hyperflatness over time of the failed and healthy bearing.
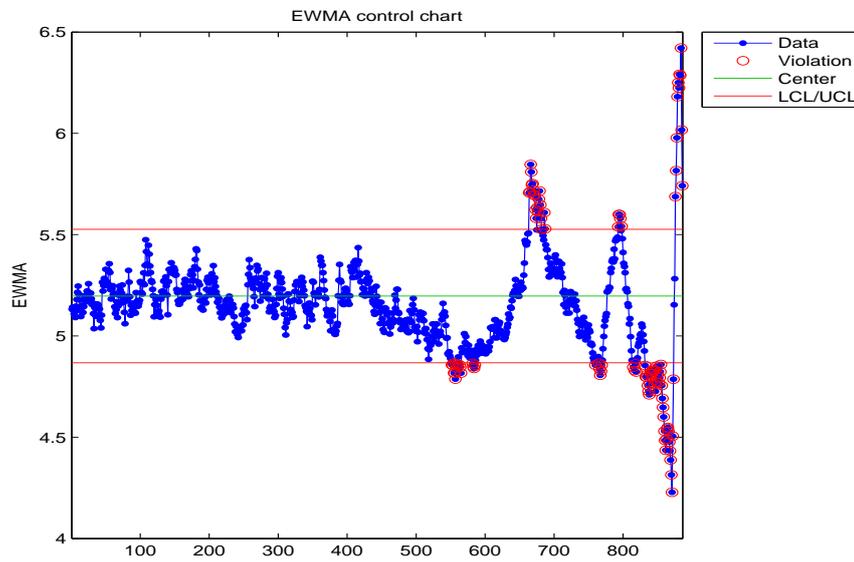
Figure B.52: EWMA chart for the crest factor of the failed bearing.
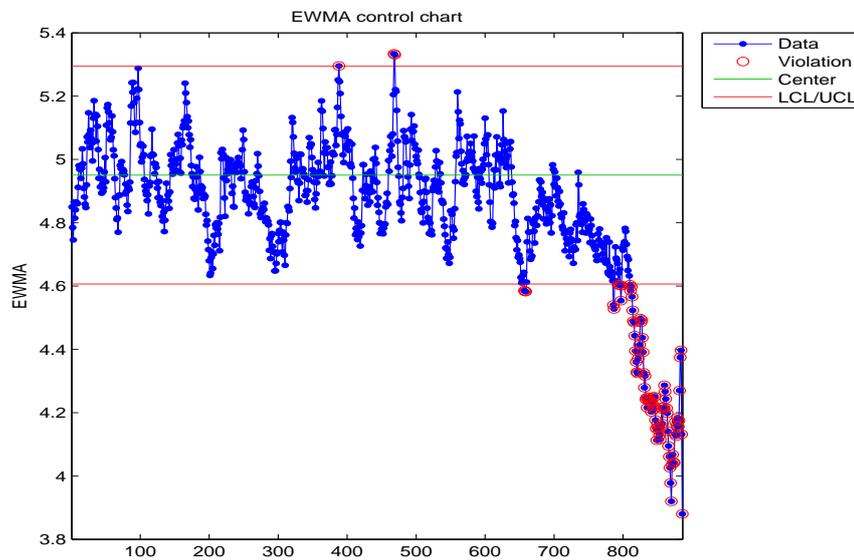
Figure B.53: EWMA chart for the crest factor of the healthy bearing.
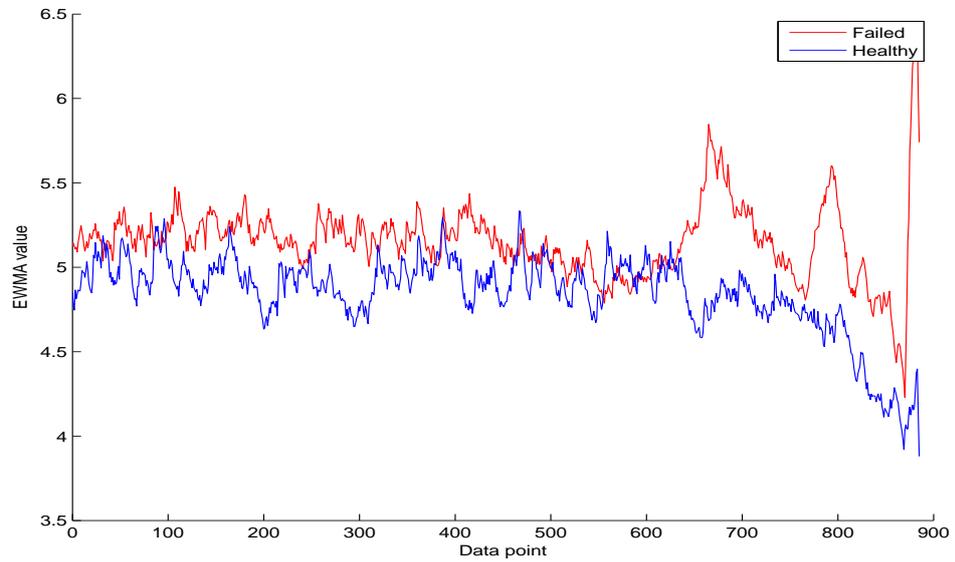
Figure B.54: Crest factor over time of the failed and healthy bearing.
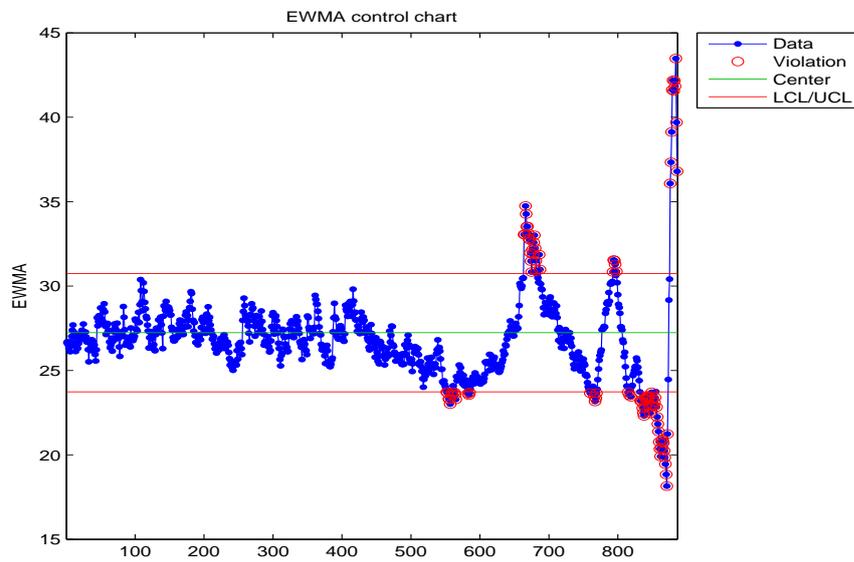
Figure B.55: EWMA chart for the peak to average power ratio of the failed bearing.
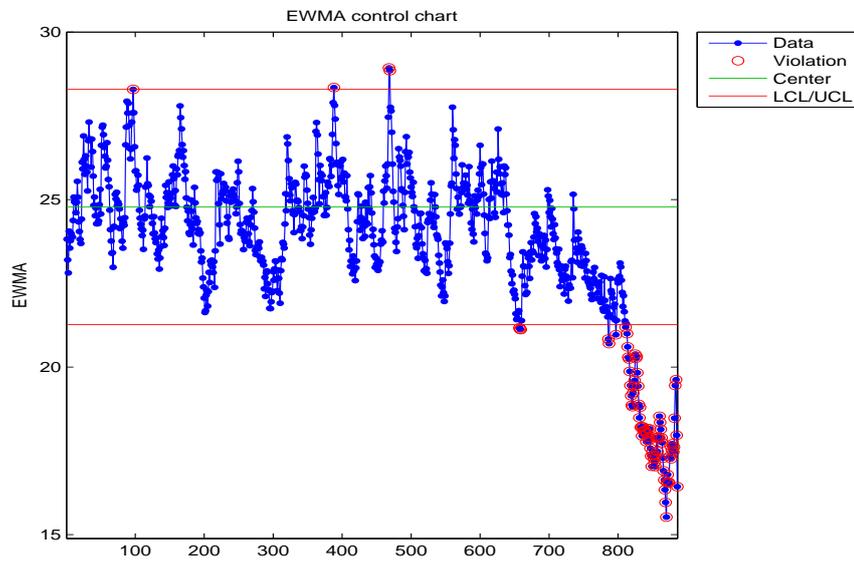
Figure B.56: EWMA chart for the peak to average power ratio of the healthy bearing.
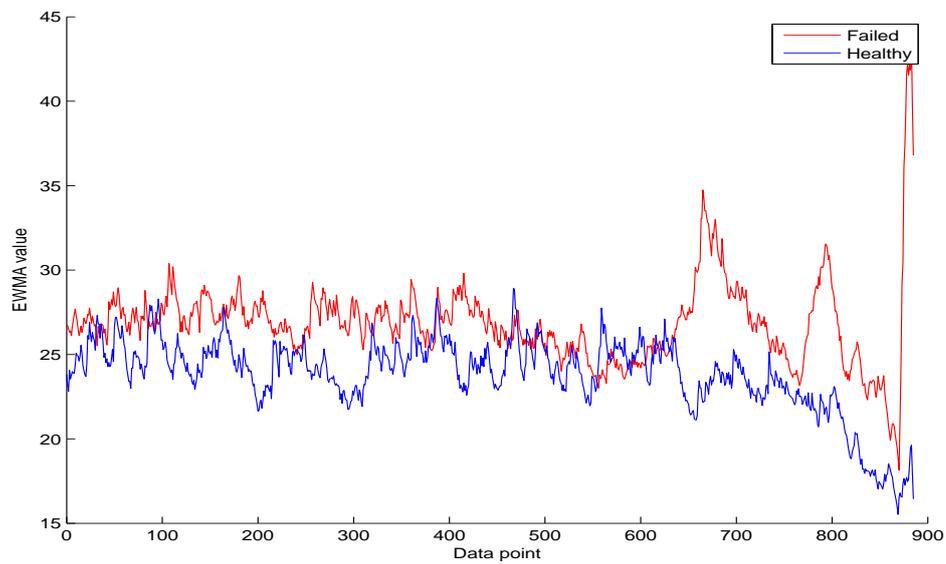
Figure B.57: Peak to average power ratio over time of the failed and healthy bearing.

B.4   Section 4.2 results - Measures of position
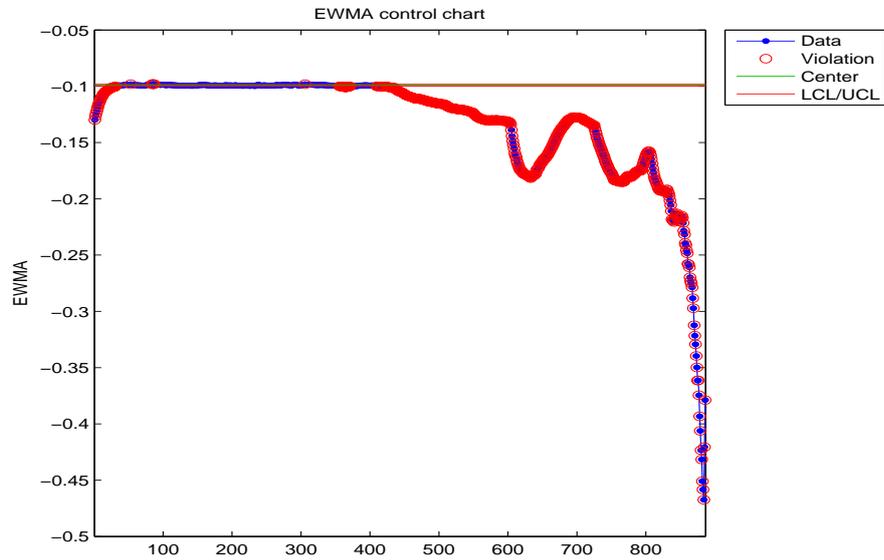
Figure B.58: EWMA chart for the 10th percentile of the failed bearing.
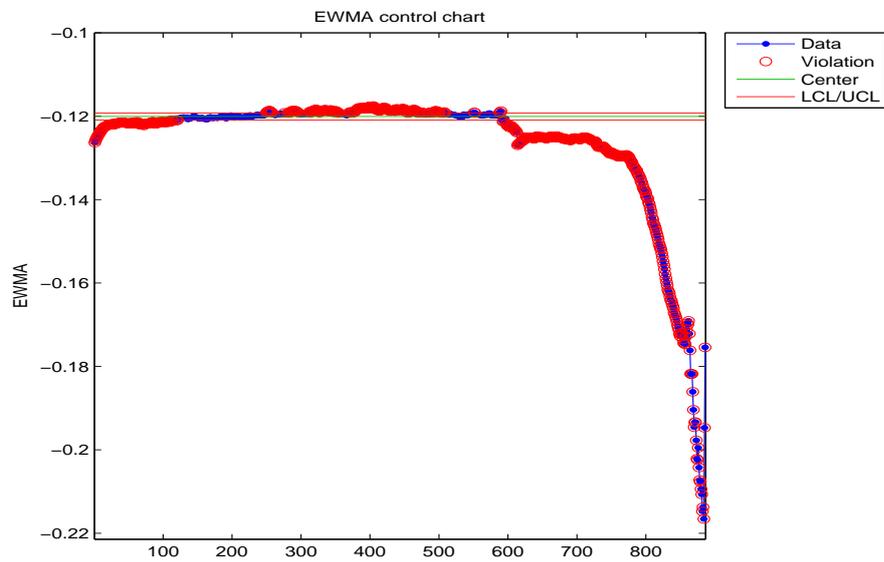
Figure B.59: EWMA chart for the 10th percentile of the healthy bearing.
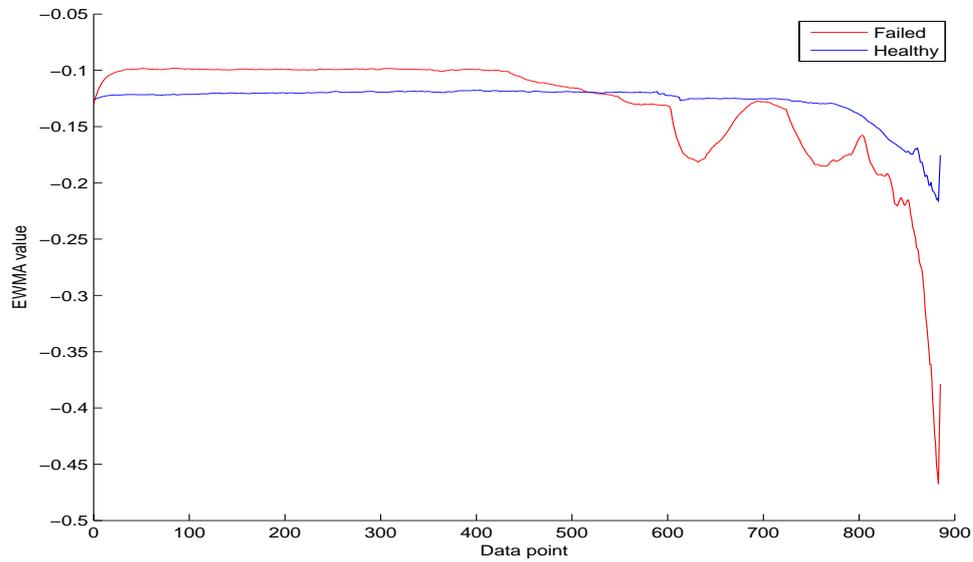
Figure B.60: 10th percentile over time of the failed and healthy bearing.
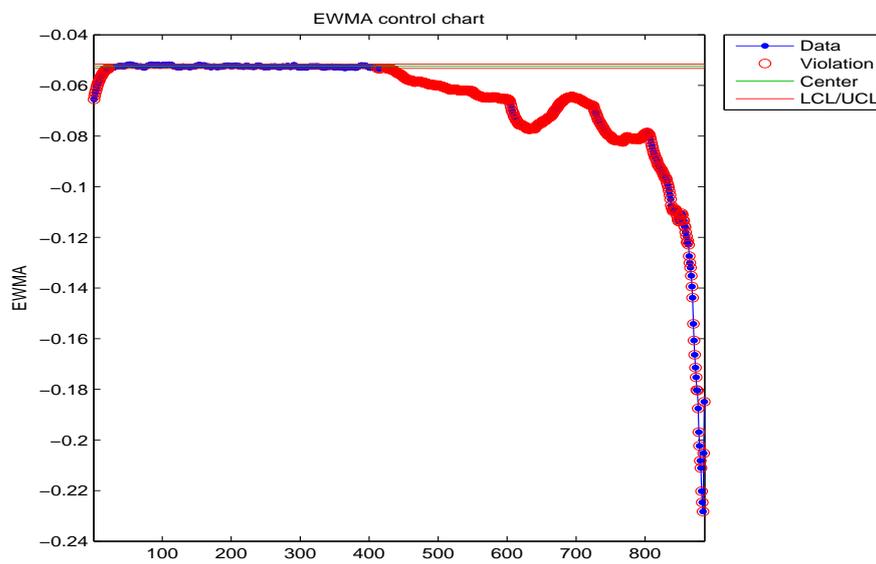
Figure B.61: EWMA chart for the 25th percentile of the failed bearing.
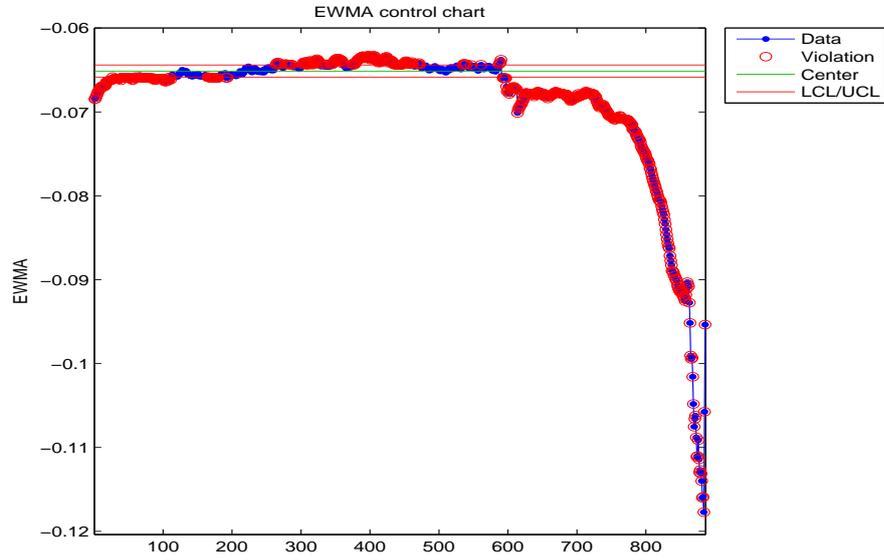
Figure B.62: EWMA chart for the 25th percentile of the healthy bearing.
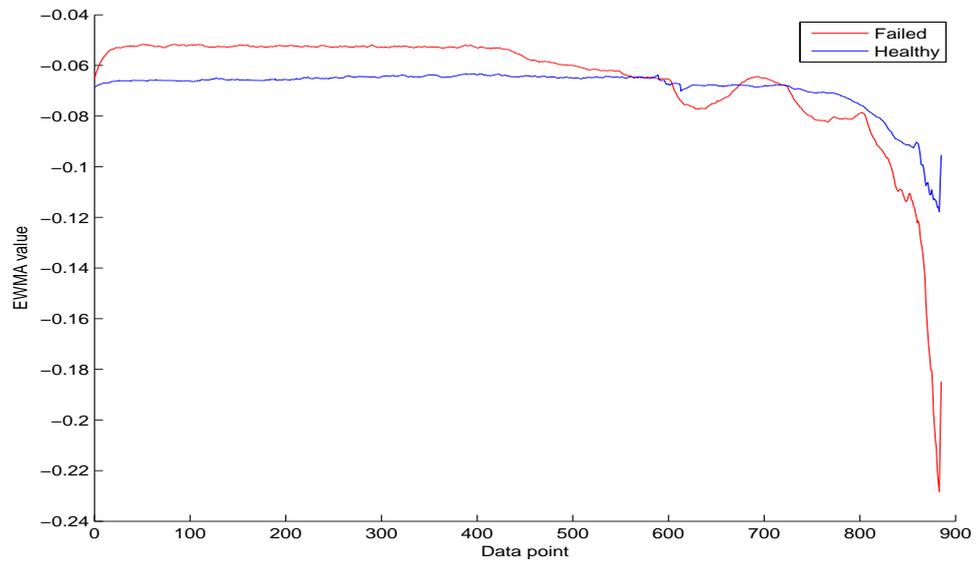
Figure B.63: 25th percentile over time of the failed and healthy bearing.
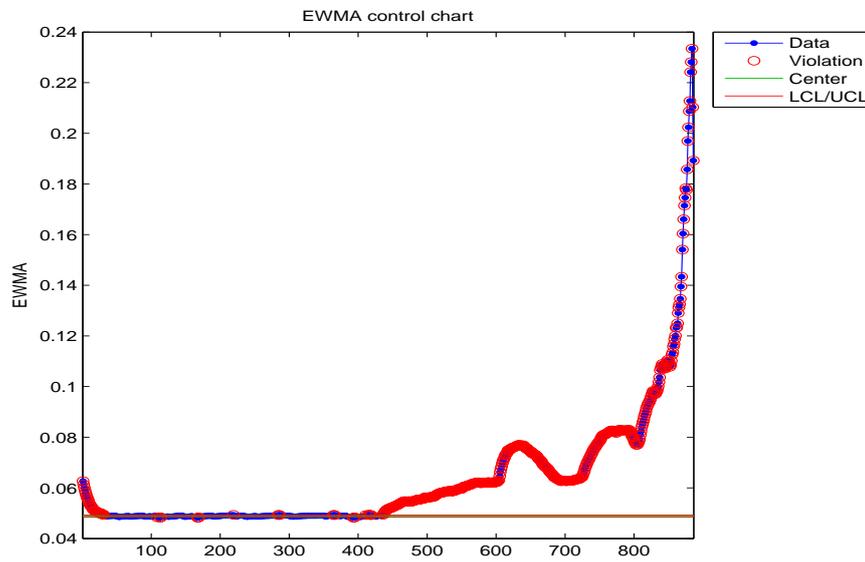
Figure B.64: EWMA chart for the 75th percentile of the failed bearing.
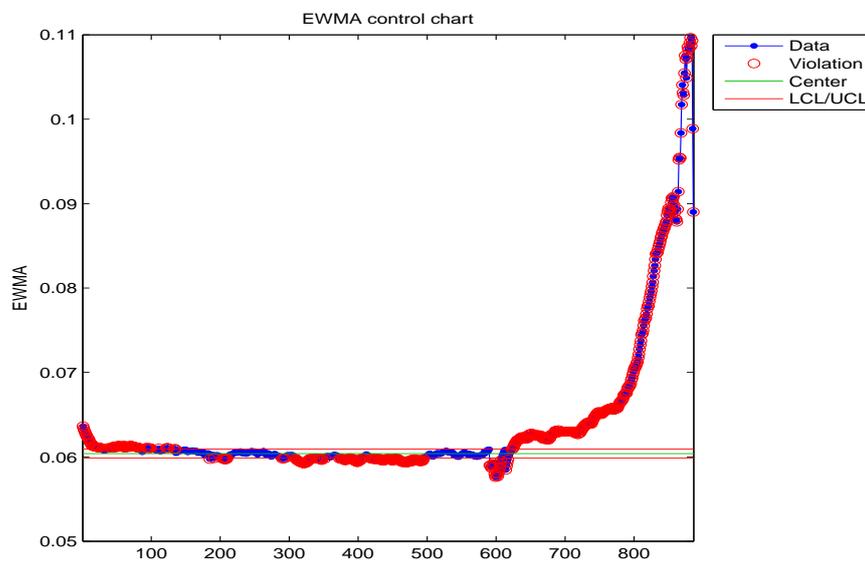
Figure B.65: EWMA chart for the 75th percentile of the healthy bearing.
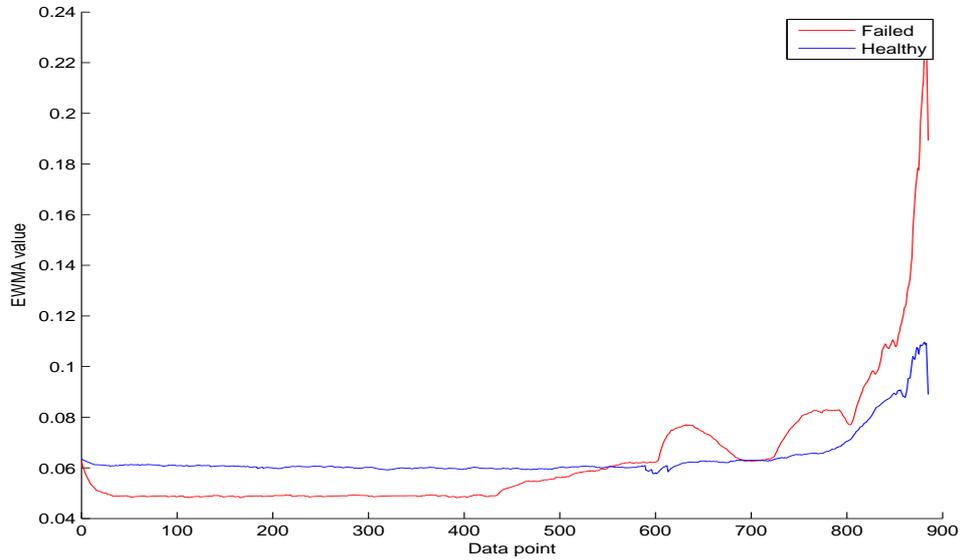
Figure B.66: 75th percentile over time of the failed and healthy bearing.
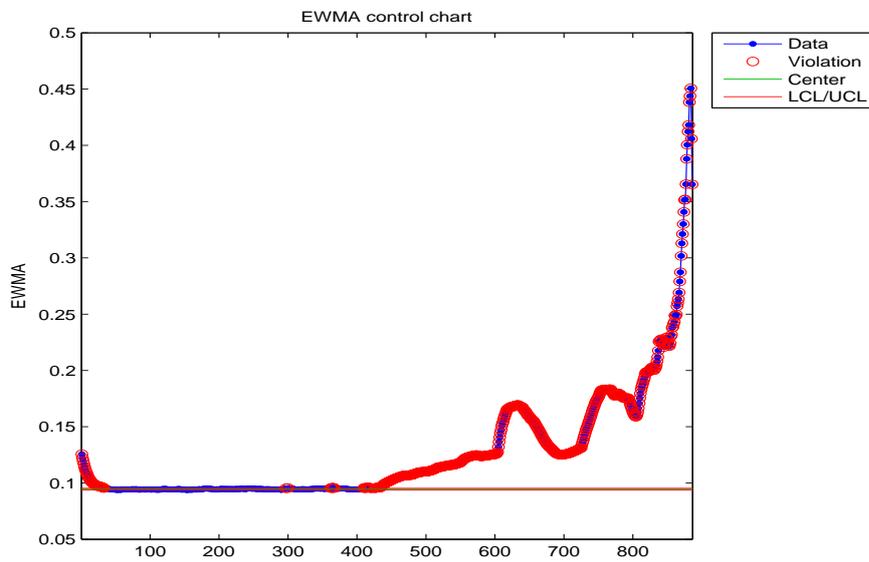
Figure B.67: EWMA chart for the 90th percentile of the failed bearing.
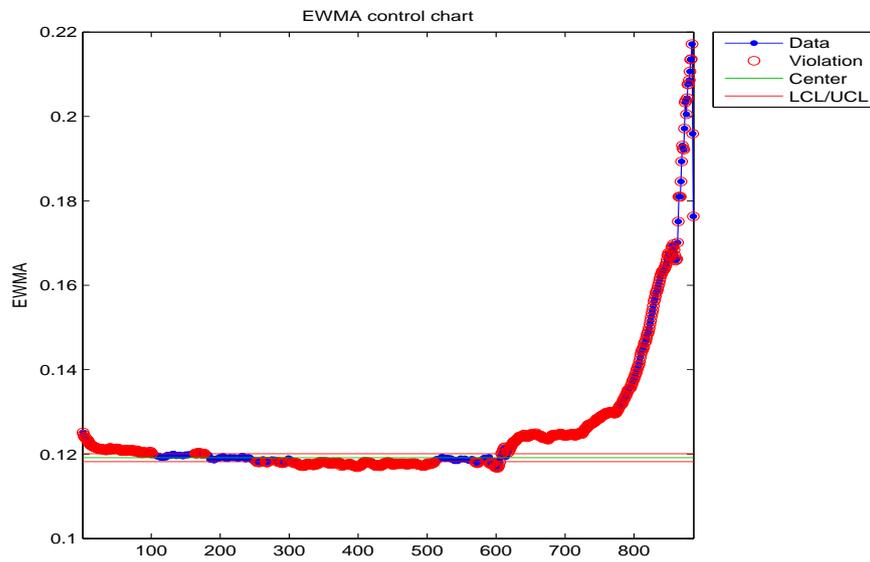
Figure B.68: EWMA chart for the 90th percentile of the healthy bearing.
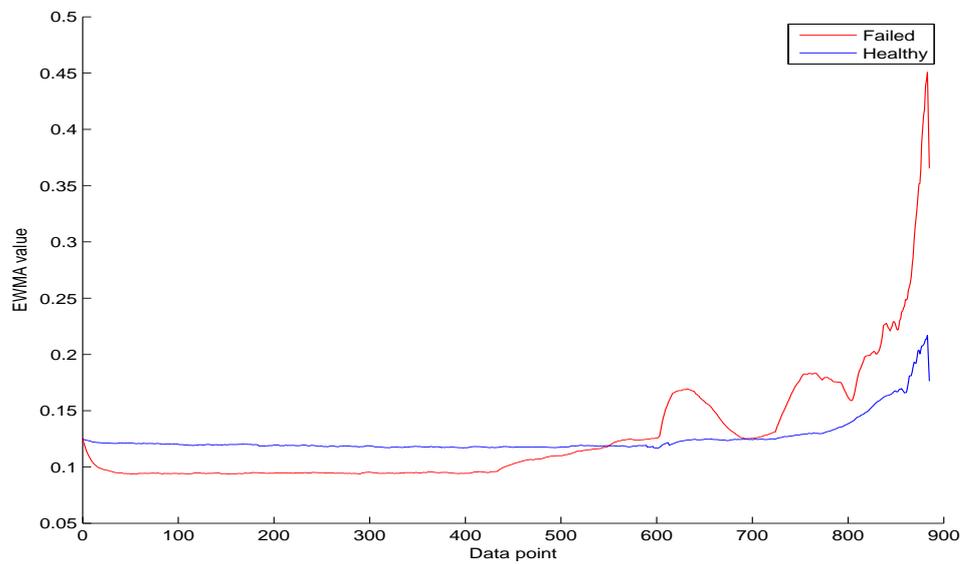
Figure B.69: 90th percentile over time of the failed and healthy bearing.

B.5   Section 4.2 results - Measures of impurity
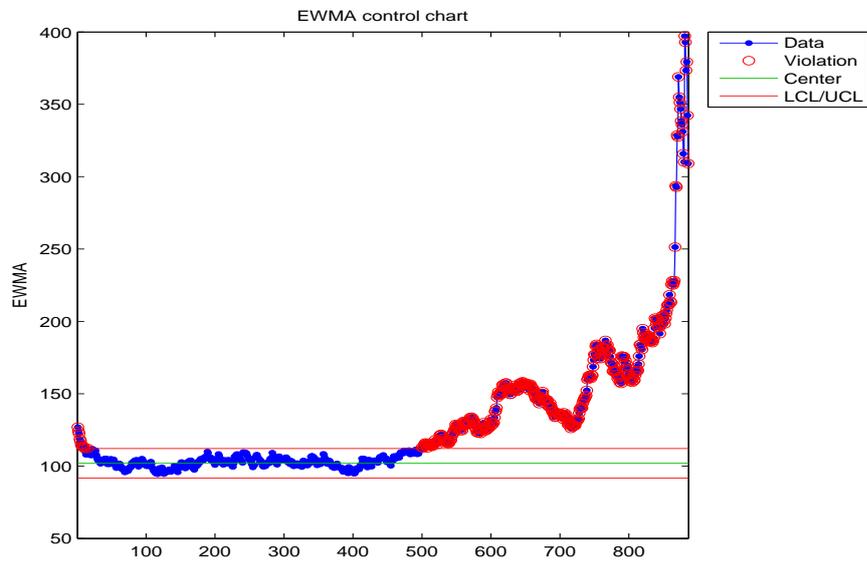
Figure B.70: EWMA chart for the Shannon entropy of the failed bearing.
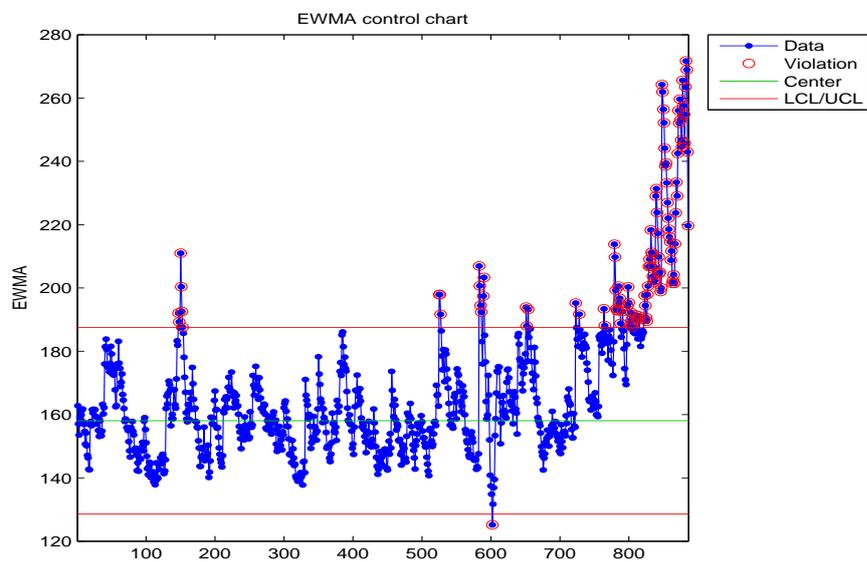
Figure B.71: EWMA chart for the Shannon entropy of the healthy bearing.
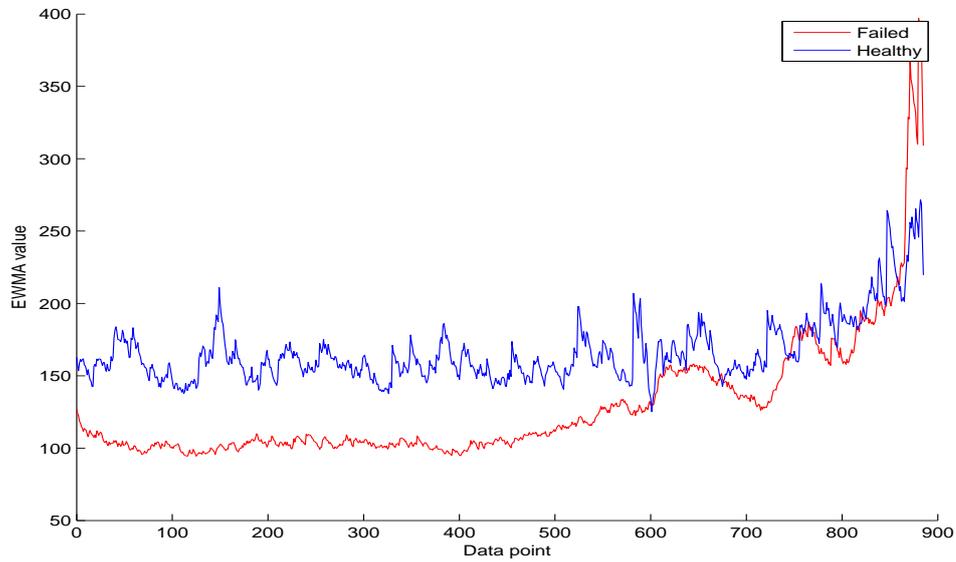
Figure B.72: Shannon entropy over time of the failed and healthy bearing.
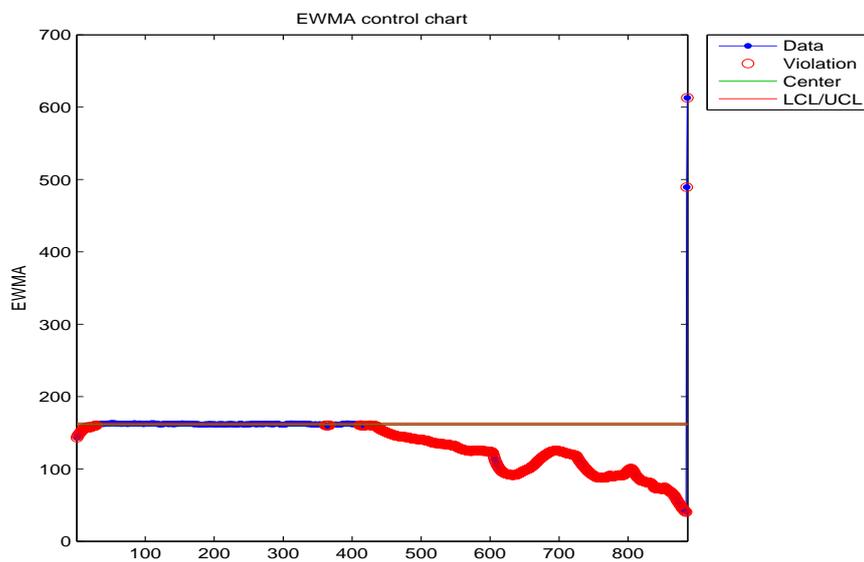
Figure B.73: EWMA chart for the Chao-Shen entropy of the failed bearing.
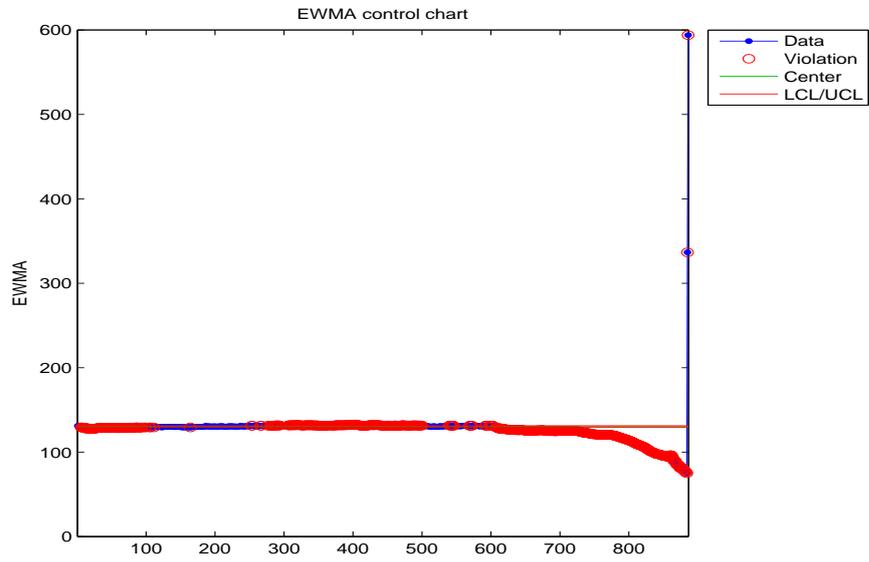
Figure B.74: EWMA chart for the Chao-Shen entropy of the healthy bearing.
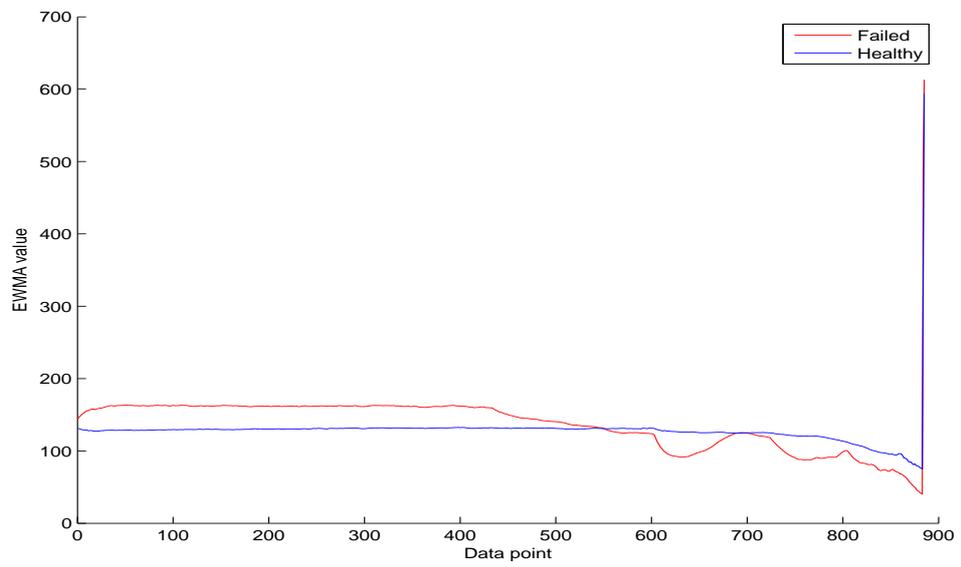
Figure B.75: Chao-Shen entropy over time of the failed and healthy bearing.
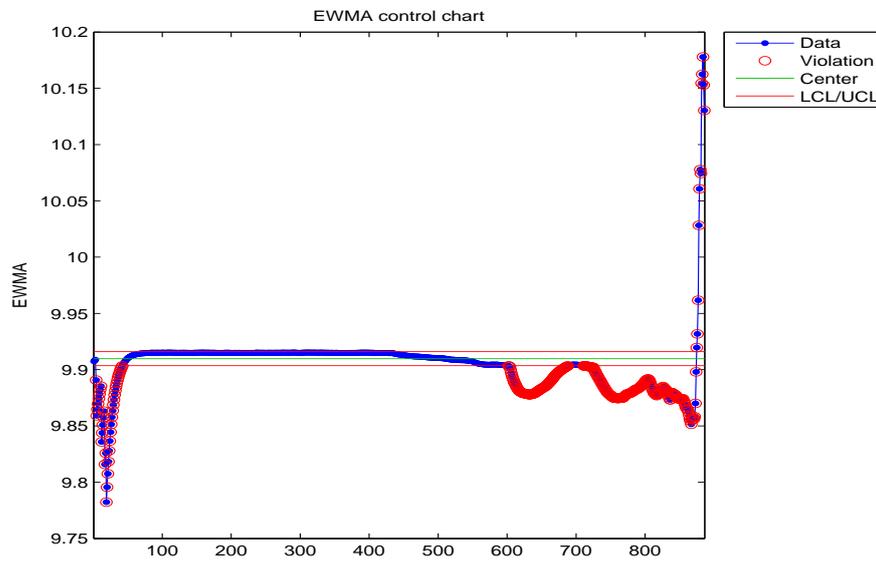
Figure B.76: EWMA chart for the KT entropy of the failed bearing.
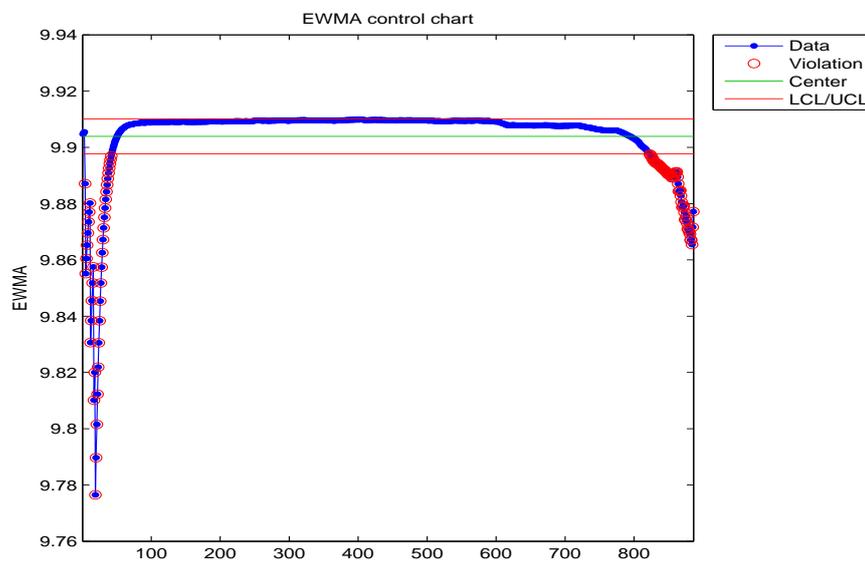
Figure B.77: EWMA chart for the KT entropy of the healthy bearing.
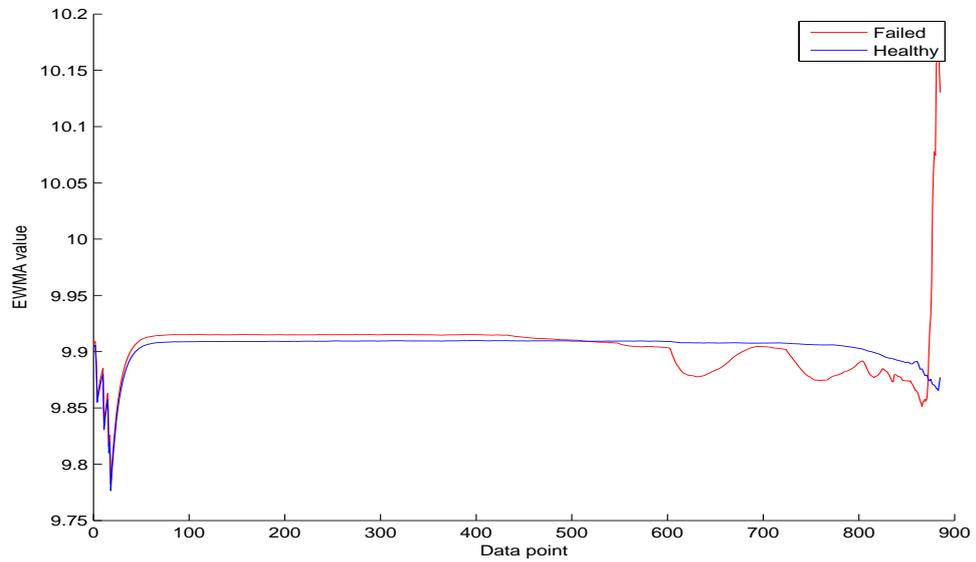
Figure B.78: KT entropy over time of the failed and healthy bearing.
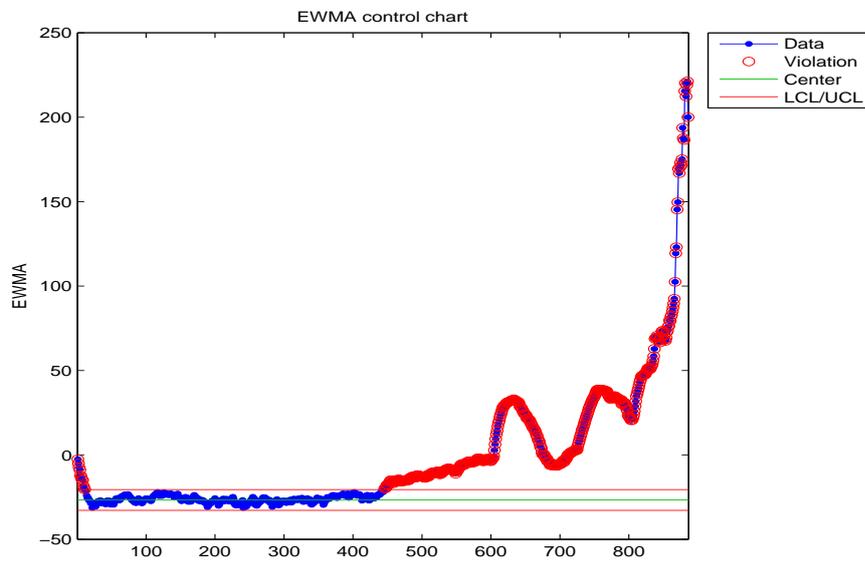
Figure B.79: EWMA chart for the Miller-Madow entropy of the failed bearing.
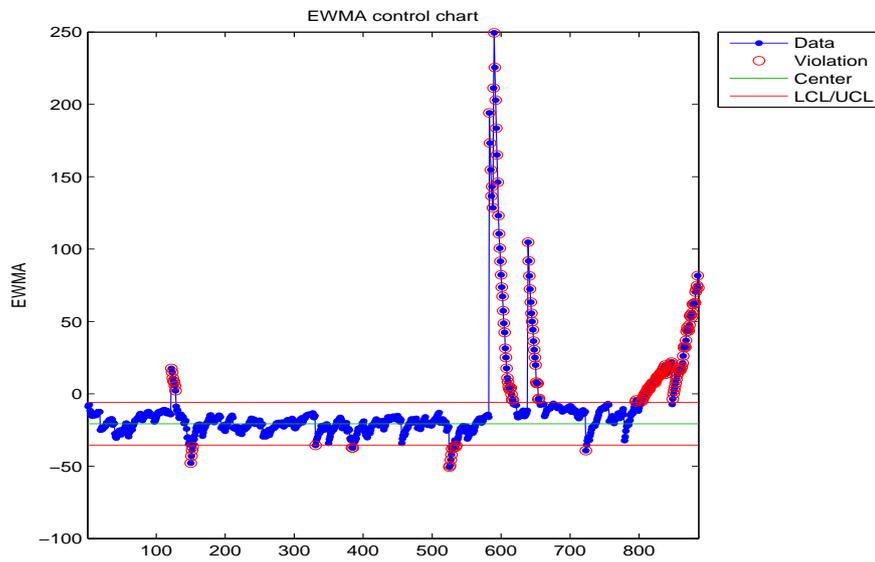
Figure B.80: EWMA chart for the Miller-Madow entropy of the healthy bearing.
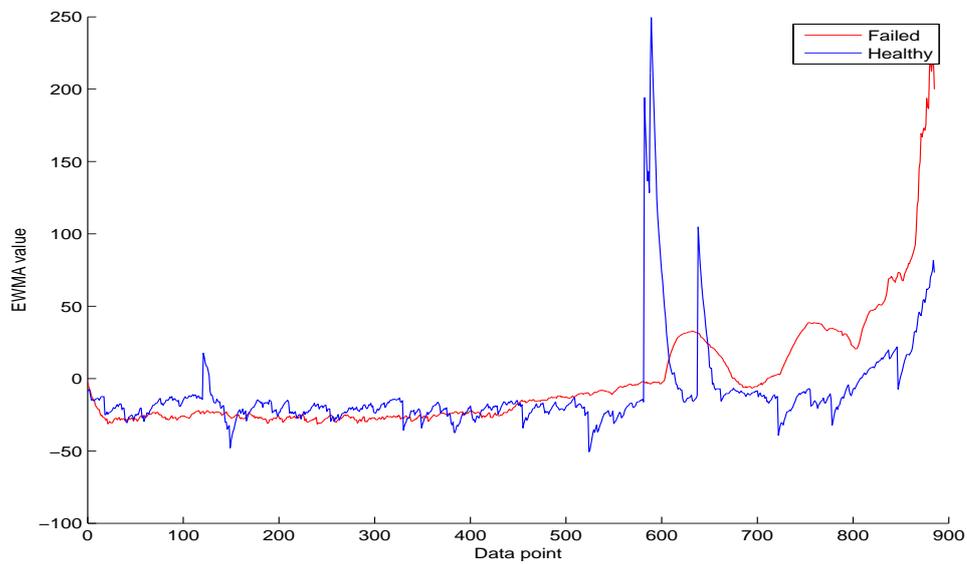
Figure B.81: Miller-Madow entropy over time of the failed and healthy bearing.
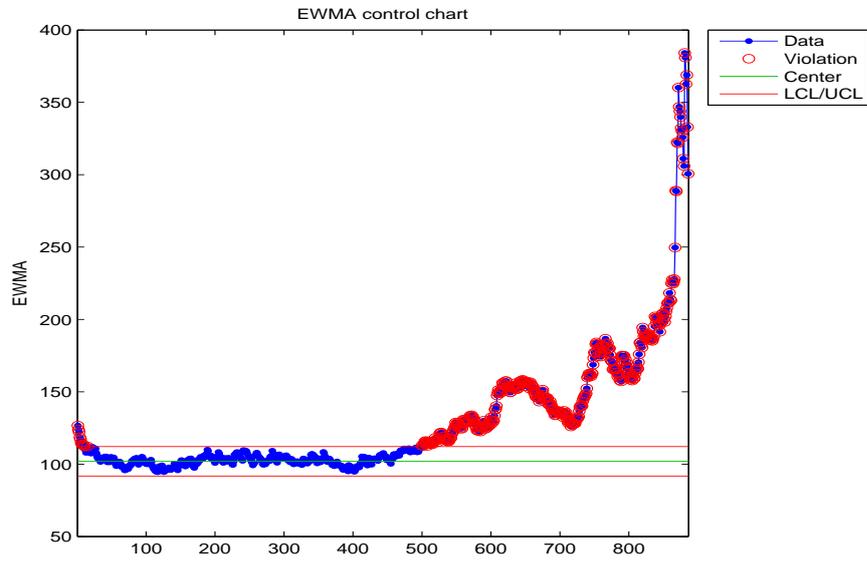
Figure B.82: EWMA chart for the James-Stein shrinkage entropy of the failed bearing.
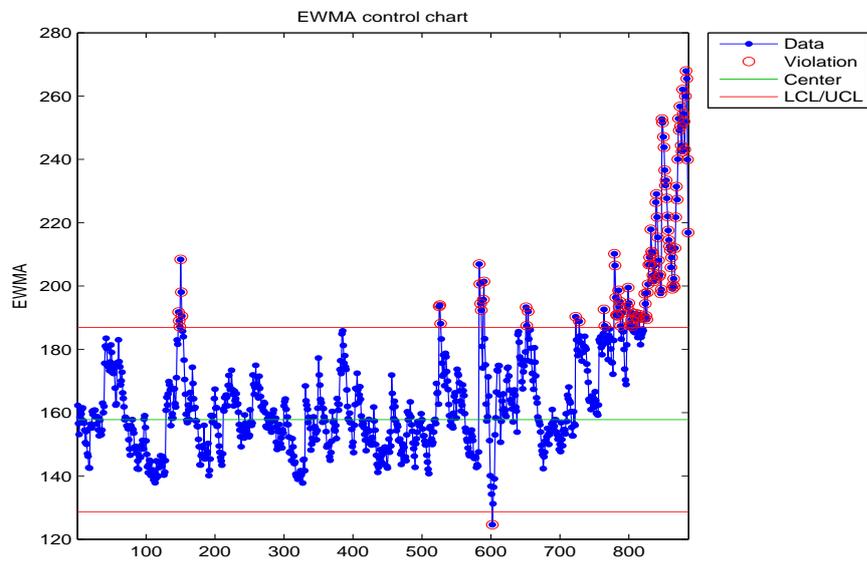
Figure B.83: EWMA chart for the James-Stein shrinkage entropy of the healthy bearing.
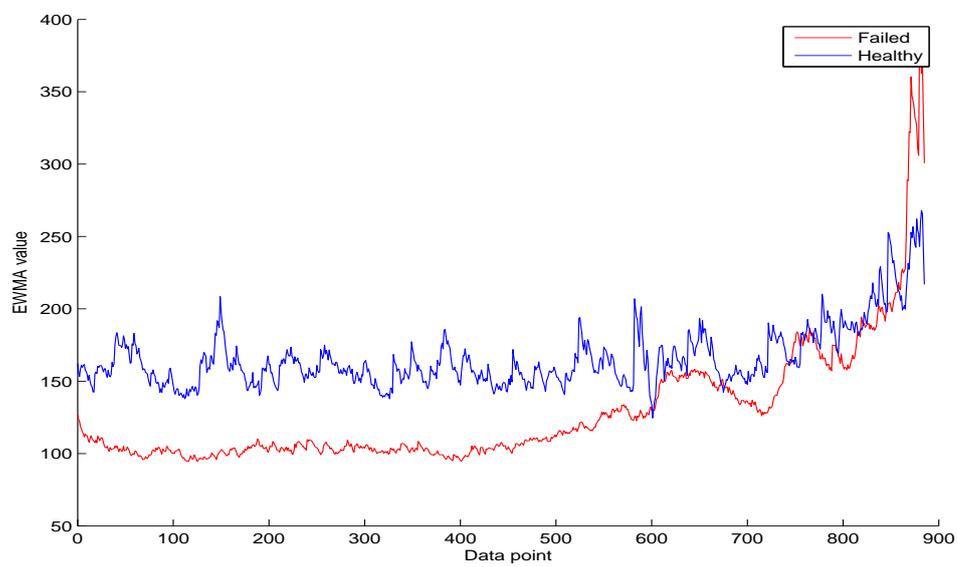
Figure B.84: James-Stein shrinkage entropy over time of the failed and healthy bearing.