Enrichment of Wind Turbine Health History for Condition-Based Maintenance

COX, ROGER

Enrichment of Wind Turbine Health History for Condition-Based Maintenance

Roger Cox

This research develops a methodology for and shows the benefit of linking records of wind turbine maintenance. It analyses commercially sensitive real-world maintenance records with the aim of improving the productivity of offshore wind farms.

The novel achievements of this research are that it applies multi-feature record linkage techniques to maintenance data, that it applies statistical techniques for the interval estimation of a binomial proportion to record linkage techniques and that it estimates the distribution of the coverage error of statistical techniques for the interval estimation of a binomial proportion. The main contribution of this research is a process for the enrichment of offshore wind turbine health history.

The economic productivity of a wind farm depends on the price of electricity and on the suitability of the weather, both of which are beyond the control of a maintenance team, but also on the cost of operating the wind farm, on the cost of maintaining the wind turbines and on how much of the wind farm's potential production of electricity is lost to outages. Improvements in maintenance scheduling, in condition-based maintenance, in troubleshooting and in the measurement of maintenance effectiveness all require knowledge of the health history of the plant. To this end, this thesis presents new techniques for linking together existing records of offshore wind turbine health history.

Multi-feature record linkage techniques are used to link records of maintenance data together. Both the quality of record linkage and the uncertainty of that quality are assessed. The quality of record linkage was measured by comparing the generated set of linked records to a gold standard set of linked records identified in collaboration with offshore wind turbine maintenance experts. The process for the enrichment of offshore wind turbine health history developed in this research requires a vector of weights and thresholds. The agreement and disagreement weights for each feature indicate the importance of the feature to the quality of record linkage. This research uses differential evolution to globally optimise this vector of weights and thresholds.

There is inevitably some uncertainty associated with the measurement of the quality of record linkage, and consequently with the optimum values for the weights and thresholds; this research not only measures the quality of record linkage but also identifies robust techniques for the estimation of its uncertainty.

Enrichment of Wind Turbine Health History for Condition-Based Maintenance

Thesis submitted towards the degree of Doctor of Philosophy

Roger William Cox

Department of Engineering

Durham University

2021

Contents

List of Tables

List of Illustrations

List of Abbreviations

ANN	Artificial Neural Network	page 35
AT	Absolute Tolerance	page 69
ATAE	Alternative Technique Assuming Equally representative samples	page 147
AUC	Area Under a ROC Curve	page 34
AW_{Fe}	Agreement Weight for a given Feature	page 81
B	Beta distribution	page 131
bin	binomial crossover scheme	page 68
BNB	Bernoulli Naïve Bayes	page 64
BoP/OFTO	Balance of Plant / OFfshore Transmission Owner	page 50
BS	Bootstrap Size	page 149
CBM	Condition-Based Maintenance	page 31
CDF	Cumulative Distribution Function	page 145
CE	Coverage Error	page 135
CI	Confidence Interval	page 125

CL	Confidence Level	page 125
CMU	Condition Monitoring Unit	page 27
CO	Count of POLRs	page 107
CP	Coverage Probability	page 129
CR	Crossover Rate	page 69
CTV	Crew Transfer Vessel	page 25
D	high cycle fatigue Damage	page 32
DE	Differential Evolution	page 67
De	'Description' feature	page 98
Di	Dithering rate	page 67
Δt_{Fe}	time difference for each time Feature	page 90
Δt_{Fi}	time difference between the WO Finish date and the outage	page 90
Δt_{No}	time difference between the WO Notification date and the outage	page 90
Δt_{Pa}	time difference between the Part posting date and the outage	page 90
Δt_{St}	time difference between the WO Start date and the outage	page 90
DW_{Fe}	Disagreement Weight for a given Feature	page 80
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortisation	page 28
EEHH	Edited version of the EHH	page 172
EHH	Enriched Health History	page 25
FL	Functional Location	page 53
FN	number of False Negatives	page 33
FP	number of False Positives	page 33
FPR	False Positive Rate	page 34
FSPRL	Fellegi and Sunter, 1969, Probabilistic Record Linkage	page 58
g	generation number	page 68
GSSLR	Sample "Gold Standard" Set of Linked Records	page 56
HHE	Health History Enrichment	Page 88
i	running index	page 68
I_{AB}	Increase from A to B	page 141

${}^{PPV}_{Th_{De}} I \, {}^{0.75}_{0}$	Increase in the *PPV* of the EHH that would by yielded by a Th_{De} of 0.75 over that yielded by a Th_{De} of 0	page 170
${}^{PPV}_{Th_{Pa}} I \, {}^{2}_{-0.5}$	Increase in the *PPV* of the EHH that would by yielded by a Th_{Pa} of 2 over that yielded by a Th_{Pa} of -0.5	page 177
${}^{\mathrm{PM}}_{\mathrm{Th_S}} \mathrm{I} \, {}^{-4}_{4}$	Increase in the *PM* of the EHH that would by yielded by a Th_S of -4 over that yielded by a Th_S of 4	page 210
${}^{PPV}_{Th_{SP}} I \, {}^{1.7}_{0}$	Increase in the *PPV* of the EHH that would by yielded by a Th_{SP} of 1.7 over that yielded by a Th_{SP} of 0	page 174
${}^{PPV}_{TTOA} I \, {}^{10}_{100}$	Increase in the *PPV* of the EHH that would by yielded by a *TTOA* of 10 minutes over that yielded by a *TTOA* of 100 minutes	page 166
${}^{PPV}_{We_{MF}} I \, {}^{0}_{1}$	Increase in the *PPV* of the EHH that would by yielded by a We_{MF} of zero over that yielded by a We_{MF} of 1	page 189
j	parameter index	page 68
κ	$1 - {}^{S}/_{2}$ quantile of the probit function	page 127
$L_{Interval}$	Lower limit of an interval	page 129
LO_{Alarm}	LO_{Class} of the “cabinet too hot” Alarm	page 106
LO_{Class}	Log Odds that each POLR belongs to a given Class	page 105
MLI	Material Consumption Line Item	page 54
MP	Mutation Probability	page 68
Mu	Mutation rate	page 68
n	sample size	page 126
NLP	Natural Language Processing	page 58
NP	Number of Parameters	page 68
NPPV	Negational Positive Predictive Value	page 77
NR	number of POLRs where the Health History would Not Recommend all the required parts	page 141
OFF	OverFitting Factor	page 201
OPEX	OPerational EXpenditures	page 28
OWT	Offshore Wind Turbine	page 19
P	Probability that any given POLR in the EHH is correct	page 72
P_{00}	Probability that a randomly sampled POLR would not recommend all the required parts for both case A and case B	page 144

P_{01}	Probability that a randomly sampled POLR would not recommend all the required parts for case A but would recommend all the required parts for case B	page 144
P_{10}	Probability that a randomly sampled POLR would recommend all the required parts for case A but would not recommend all the required parts for case B	page 144
P_{11}	Probability that a randomly sampled POLR would recommend all the required parts for both case A and case B	page 144
PCT	Process for Comparing Techniques for interval estimation for a change in a proportion	page 151
PDCE	Process for predicting the Distribution of the *CE*	page 153
PDF	Probability Density Function	page 148
PE	Population Energies	page 69
PEOHH	Process for the Enrichment of OWT Health History	page 45
PFT	Parts Frequency Technique	page 108
$\hat{p}$	point estimate of *p*	page 125
ϕ	Gaussian	page 126
Φ^{-1}	Probit	page 127
PIT	Process for the Identification of the Training data	page 104
PM	Proportion of POLRs in a set where the Health History would recommend all the Material needed	page 140
POLR	Pair Of Linked Records	page 56
POTS	Process for the Optimisation of Th_{SP}	page 172
PRL	Probabilistic Record Linkage	page 56
Prob	vector of Probabilities	page 144
PS	Parts Score	page 109
PSi	Population Size	page 68
PS_A	*PS* for Alarms	page 192
PS_O	*PS* for Outages	page 192
PVEHH	Process for the Validation of the EHH	page 85
r	randomly selected population vector	page 68
rand	uniformly distributed real number generated anew for every parameter of every vector	page 69

randg	uniformly distributed real number generated anew for every generation	page 68
Reps	number of Repetitions of the PCT	page 153
ROC	Receiver Operating Characteristics	page 34
RQ	Research Question	page 41
RT	Relative Tolerance	page 69
S	Significance	page 127
SS	Step size	page 146
SCADA	Supervisory Control and Data Acquisition	page 34
SOV	Service Operation Vessel	page 25
SPOLR	Set of Pairs Of Linked Records	page 81
S_{POLR}	Score for each POLR	page 82
SR	Similarity Ratio	page 100
TAE	Technique Assuming Equally representative samples	page 144
TAI	Technique Assuming Independence between PM_A and PM_B	page 143
TeS	Size of the Test data sample	page 181
Th_{De}	Description Threshold	page 100
Th_{Fe}	agreement Threshold for each time Feature	page 90
Th_S	Score Threshold	page 197
Th_{SP}	Parts Training Data Score Threshold	page 104
TN	Number of True Negatives	page 33
TNR	True Negative Rate	page 57
TP	number of True Positives	page 33
TPR	True Positive Rate	page 33
TrS	Size of the Training data sample	page 181
TTOA	Time difference between Outages and Alarms	page 97
TUB	Technique Using Bootstrapping	page 150
u	probability space	page 146
$U_{Interval}$	Upper limit of an interval	page 129
u_g^i	trial vector	page 68

v_g^i	mutant vector	page 68
VVT	Verification, Validation and Testing	page 73
We_{MF}	Minor Feature Weight	page 187
$W_{Fe_{POLR}}$	Weight for each Feature for each POLR	page 81
WO	Work Order	page 52
WR	number of POLRs where the Health History Would Recommend all the required parts	page 140
x_g^i	target vector	page 68
x_g^{best}	best vector	page 68

Acknowledgements

This research project was supported both by Durham University Department of Engineering and by renewable energy company Ørsted A/S.

Firstly, I want to thank my supervisors at Durham University, Dr Peter Matthews and Dr Christopher Crabtree and my supervisors at Ørsted, Miriam Marchante Jiménez, Jacob Juhl Christensen and Lars Høst Johansen for their very helpful advice and support.

This research benefitted from access to Ørsted's valuable and confidential records of wind turbine health history and to support from Ørsted's wind turbine experts, such as Robert Jakobsen and Neil MacDougall who helped with identifying a "gold standard" set of linked records. It also benefitted from visits to two of Ørsted's wind farms and I am grateful to the staff who I worked with at these sites whose names are not included in this thesis because the identity of these sites is confidential.

I would like to thank Rikke Meyer Pedersen, Maria Saxild-Laursen and Christopher Dam Jensen who introduced Ørsted's records of wind turbine health history to me, Simon Børresen who explained the economics of wind turbine maintenance to me and also Jan Frydendall and Mikael Sonne Hansen from Ørsted and Professor Toby Breckon from Durham University Department of Computer Science who advised me regarding data science and linguistic techniques.

I am grateful to Jonathan Owen and Dr Matthias Troffaes from Durham University Department of Mathematical Sciences who helped this project with their statistical expertise.

Lastly, I would like to thank my partner, my parents and my friends for their support.

1 Introduction

Section 1.1 will present a brief introduction to the offshore wind energy sector of the energy industry. It presents overall trends in the sector: trends in the rating of Offshore Wind Turbines (OWT) and in the distance to shore of wind farms. It is included to provide context for section 1.2, which will go into more detail about maintenance, and about maintenance record keeping, in this sector and will describe the history of attempts to improve the maintenance of wind turbines. It will explore the state of the art of the maintenance of OWTs, demonstrating a need to enrich the health history of the OWTs. These sections together will show that the challenges of maintaining OWTs, together with their growing importance in the energy industry, have brought innovations in OWT maintenance to the forefront of maintenance technology and that success in this sector relies on good information. Section 1.3 specifies this thesis area of research, section 1.4 describes the research process, section 1.5 identifies some key constraints, section 1.6 describes the thesis structure and section 1.7 details its original contributions.

1.1 Introduction to the Wind Energy Industry

This section will locate wind energy in a global context. Renewable energy can refer to a range of forms of energy commodities such as electricity, fuel or heating where the energy source is naturally replenished on a human timescale. Figure 1-1 shows global consumption of the most important sources of electrical energy: the major fossil fuels: oil, gas and coal, nuclear fission and the major renewables: hydroelectricity, solar and wind. Data in the BP Statistical Review of World Energy, 2021, shows that the total consumption of all these commodities together increased 3.53 times over the 50 years from 1970.

Data in the BP Statistical Review of World Energy, 2021, shows that wind energy made up only 2.58% of world energy consumption in 2020 but that it was only 0.43% in 2008 so its relative importance is increasing. Consumption of wind energy was 1.86 times higher than that of solar in 2020 but it was still well behind nuclear and fossil fuels, despite world consumption of oil and gas generated electricity dropping by 6% due to Covid 19.

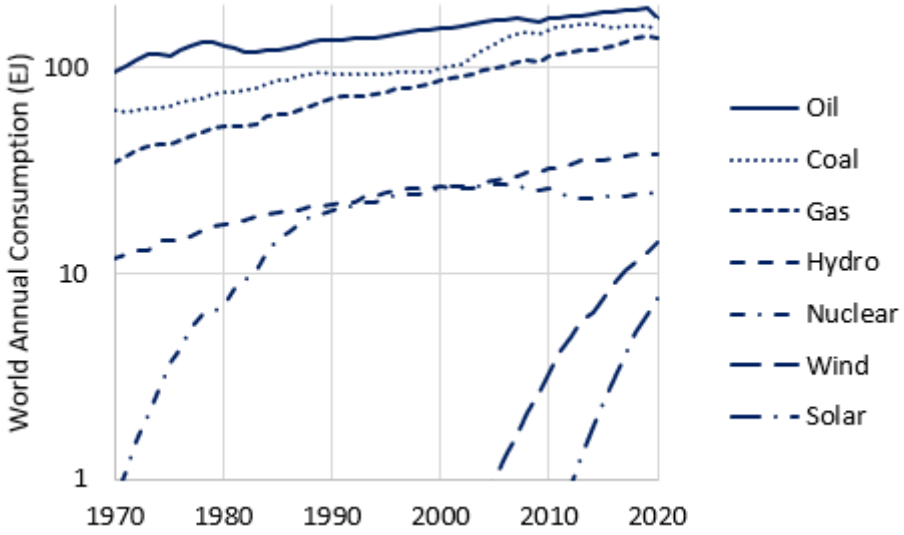


Figure 1-1, World Annual Consumption by Energy Type
from data in BP Statistical Review of World Energy, 2021.

1.1.1 Offshore Wind Energy

This section presents an overview of the OWT sector of the energy industry. It presents overall trends in the sector: trends in the rating of OWTs and in the distance to shore of wind farms. It is included to provide context for section 1.2, which will explore the state of the art of the maintenance of OWTs. It will describe the main advantages of offshore wind energy over onshore wind energy and the different costs of these two distinct technologies.

The main advantage of offshore is that the wind tends to be stronger and less turbulent offshore (Davis et al., 2019). Higher speed increases the power available proportionally to speed cubed, while the thrust increases proportionally to speed squared. A structure that can react against higher forces is more expensive to manufacture and to install but that is more than offset by the increased generation of electricity. Lower turbulence causes less fatigue damage to the structure such that it lasts longer or is cheaper to manufacture.

Wind farm developers are companies that develop, own and operate wind farms. Figure 1-2 shows mean water depth against mean distance to shore of bottom-fixed offshore wind farms in Europe, organised by development status. The size of the bubble indicates the overall capacity of the site. It shows that some new OWT farms, shown in yellow, are further offshore than those that are already generating electricity, shown in blue. Section 1.2 will discuss the consequences for the maintenance of offshore wind farms of them being developed further offshore.

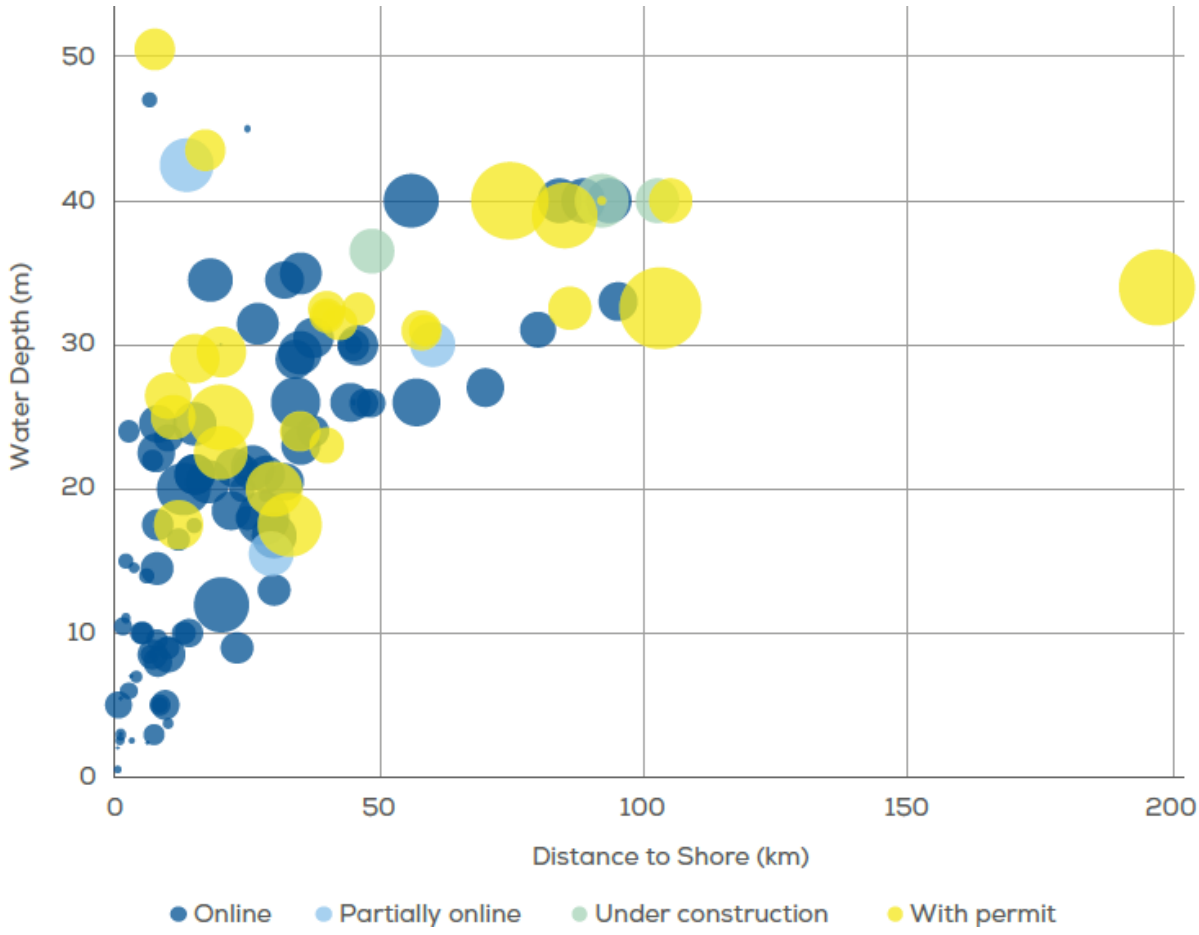


Figure 1-2, Mean Water Depth against Mean Distance to Shore of Bottom-Fixed, Offshore Wind Farms in Europe, by development status. The size of the bubble indicates the overall capacity of the site. Reproduced from WindEurope, 2018, who do not report a scale for the bubble size.

The main disadvantage of offshore compared to onshore wind power is that installation, maintenance and decommissioning tend to be more expensive offshore. Figure 1-3 illustrates that the rated power (rating) of new OWTs in Europe is tending to increase rapidly. The same trend is true in other world regions. Increasing the rating allows operators to take advantage of economies of scale, reducing operational expenditures per unit production of energy (Prässler and Schaechtele, 2012).

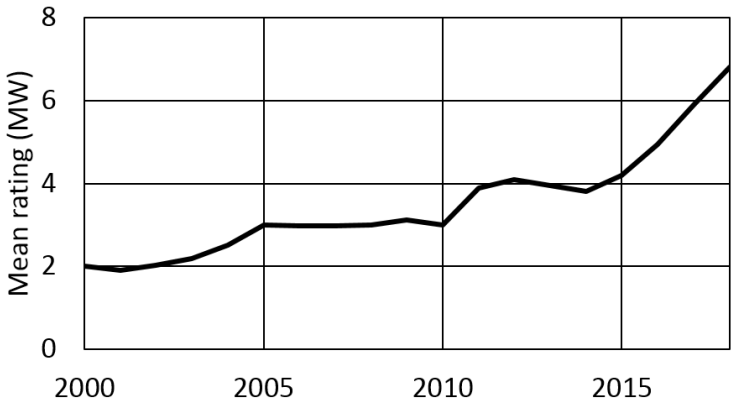


Figure 1-3, Mean Rating of Newly Installed OWTs in Europe. Based on Data from WindEurope

This research had access to wind turbine maintenance data of particularly good quality. In the early development period of wind farm development, the economic drivers for keeping excellent maintenance records were not felt strongly. Records were kept on paper and not computerised (Leahy et al., 2019). In contrast to this overview of the sector however, Ørsted recognise the importance of excellent maintenance record keeping. In the corrective maintenance of wind turbines, a work order is an instruction to carry out a maintenance activity. Figure 1-4 shows an example work order, number 80135873.

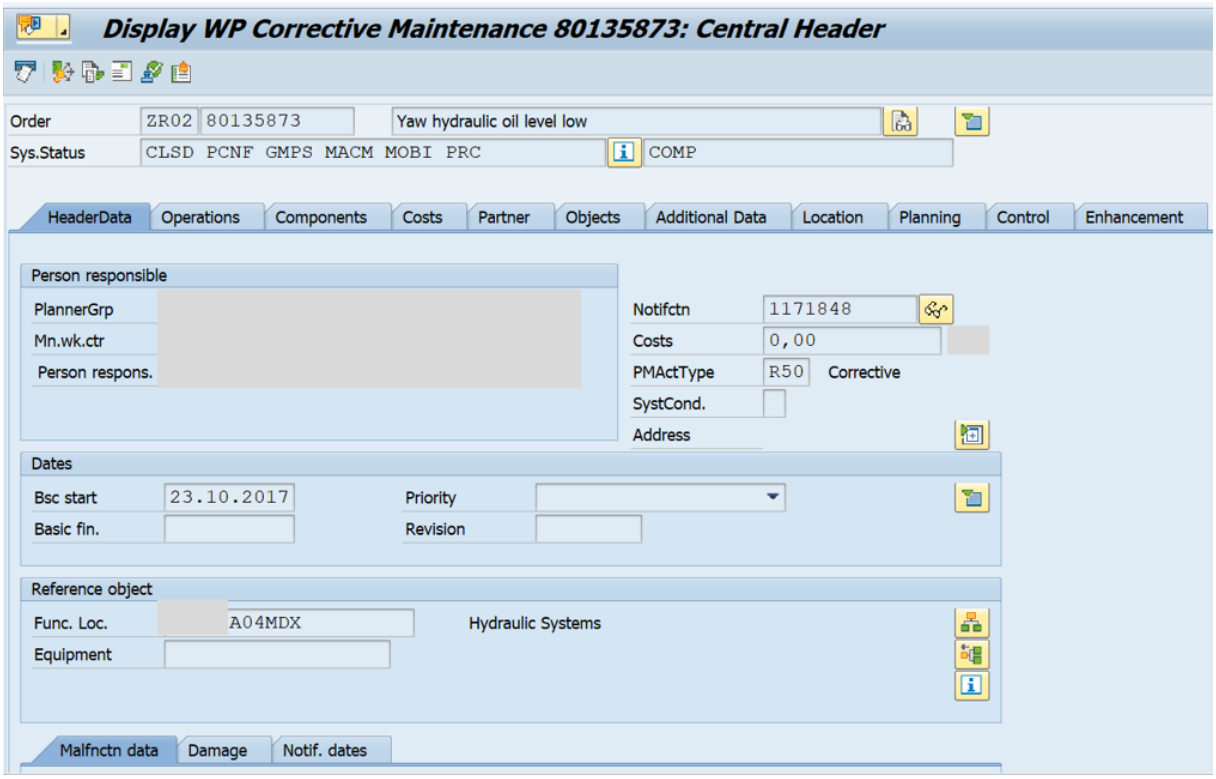


Figure 1-4, Example Work Order

After a work order has been issued, that work is added to the maintenance schedule. This prompts a maintenance team to visit the wind turbine. They asses what maintenance is actually required and, if they have brought the spare parts required to do the job, they do it. They often record any problems that they encounter on the work order.

1.2 The Maintenance of Offshore Wind Turbines

This section explores the state of the art of the maintenance of OWTs. It also samples applicable research on maintenance in other sectors and other industries. It explains the economics that drive OWT maintenance practices and it identifies gaps in the literature that, when filled, represent opportunities to aid in their improvement. This research has benefitted from access to maintenance records; existing records of OWT health history, and links those records together to determine an Enriched Health History (EHH). This section will identify gaps in the published literature that an EHH enables further work to fill.

Technicians are transported to OWTs by Crew Transfer Vessels (CTV), by Service Operation Vessels (SOV) or by helicopter. CTVs are typically 20m catamarans with 4 berths and a cargo capacity of 2 to 3 tonnes (Boote et al., 2015). SOVs are typically 90m monohulls with 60 cabins and a cargo capacity of 4000 tonnes. (Marine Traffic website). Vessel manufacturers have proposed new vessel concepts to service wind farms that are further from port. For example, Boote et al., 2015 propose a mothership for CTVs.

OWT maintenance activities are classified by operators as either preventive activities (such as the OWTs annual service), retrofit, inspections and surveys, condition-based or corrective. This chapter will show that health history information is already used to improve condition-based and corrective maintenance practices but that literature searches up to 2021 did not find health history, enriched by record linkage, described. As such, its use to further improve these maintenance practices is not described either.

OWTs are owned by energy companies that exist in a competitive business environment. The protection of their commercially sensitive data is important to them because they are indicative of commercial performance and consequently, they can affect deals between wind farm manufacturers, developers and owners. These data are also valuable because they embody learning that organisations can use to their commercial advantage and because their collection has required investment. Commercially sensitive records include failure rates, downtime, maintenance costs and maintenance working time.

Offshore wind farm operators list how much lost production each failure mode is causing and use this data to prioritise their maintenance activities. They typically keep this information confidential but despite this, researchers have been able to publish anonymised data on the failure rates of onshore and offshore WTs (Hahn et. al., 2007, Ribrant et. al., 2007, Tavner et. al., 2007, Spinato et. al.,2009, Feng et. al., 2010, Wilkinson et. al., 2011, Pinar Pérez et. al., 2013, Sheng et. al., 2013, Carroll et. al., 2015). Spinato et. al., 2009, published failure rate data for onshore WTs in a report that is of particular interest because it breaks the data down by component. This thesis presents further analysis of the data in Spinato et. al., 2009, shown in Figure 1-5, showing that the mean downtime per turbine per year was highest for the electrical system but that the gearbox caused the highest mean downtime per failure.

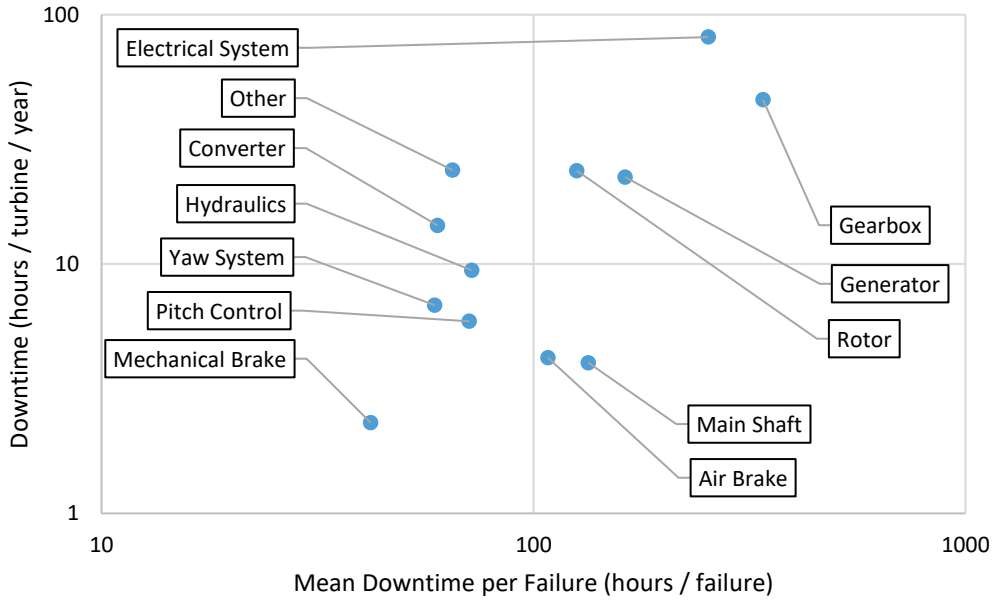


Figure 1-5, Downtime against Mean Downtime per Failure by Subassembly. Calculated from data in Spinato et. al., 2009.

This thesis presents further analysis of OWT maintenance data published by Carroll et. al., 2015, in Figure 1-6. The figure identifies that the gearbox was the component whose failures cost the most in terms of material and labour but this does not include the cost of lost production.

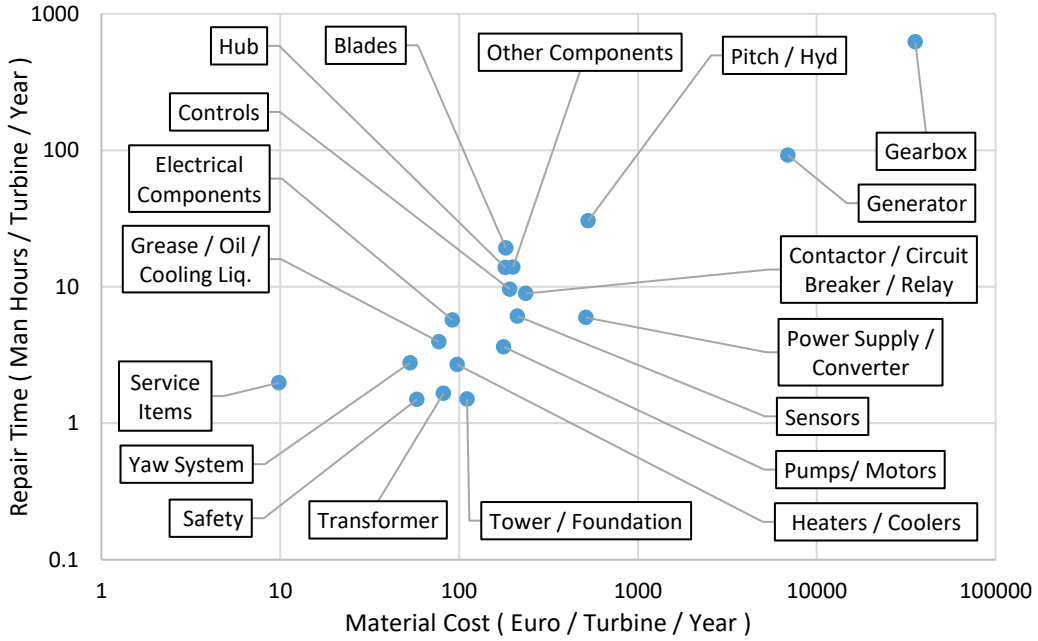


Figure 1-6, Total Work Hours against Material Cost of WOs by Subassembly. Calculated from data in Carroll et. al., 2015

Section 2.1.2 introduces outages and section 2.1.3 introduces Work Orders (WO). Of the two figures above, Spinato et. al., 2009, present power outage data whereas Carroll et. al., 2015, present data from WOs. The first publication in any sector that integrates these two sources of maintenance data was Papatzimos et al., 2017. Section 2.2 reviews Papatzimos et al., 2017 and this thesis builds on that state of the art research.

OWTs are fitted with hundreds of sensors (Qiu et al., 2012). Each OWT contains a Condition Monitoring Unit (CMU); the sensors send data to the CMU where it is recorded. The CMU is programmed with hundreds of condition models designed to identify faults, environmental conditions or operational conditions such as that the OWT is in local operation. When one of these models is triggered by one or more of the sensor signals, the CMU asserts an alarm and logs it in a database, known to operators as the alarm log (Kusiak and Li, 2011).

1.2.1 The Economics of OWT Maintenance

This section will present performance indicators used in OWT maintenance. Earnings Before Interest, Taxes, Depreciation, and Amortisation (*EBITDA*) is a financial metric used to assess productivity (Espinoza and Morris, 2013). It is defined by equation 1.1 where OPerational EXpenditures (*OPEX*) are the costs of maintenance and of operating the wind farm.

$$EBITDA = Revenue - OPEX \tag{1.1}$$

OWT operators use *EBITDA* to track the effect of maintenance strategies (Gonzalez et al., 2017, Pfaffel et al., 2019). In this research, the optimum OWT maintenance strategy is defined as that strategy that generates the maximum *EBITDA* for the operator.

OPEX summarises operators' decisions such as how many vessels to hire and how many technicians to employ. These decisions depend on the price of electricity and the cost of labour as shown below. The optimisation of OWT maintenance is outside of the scope of this research but this research does present methodologies that when they are applied will enable the more productive operation of OWTs.[1]

Equation 1.2 defines *revenue* over a given interval, where *price* denotes the price at which the electrical energy is sold to the operator's customer and *production* denotes the electrical energy generated in that interval. It disregards income from ancillary services such as the stabilisation of the frequency of the electricity grid and other external sources.

$$Revenue = Price \times Production \tag{1.2}$$

Operators refer to intervals when the OWT is not generating electricity as outages. An outage is associated with lost production if, during the interval of the outage, there was a wind resource that would, had the OWT been in operation, have been exploited. Outages that occur during intervals when the wind speed is too high or too low for the OWT to operate are not associated with lost production. To minimise lost production, there is a preference for maintenance activities to be carried out on days with low wind speed.

The *capacity factor* is defined as the ratio of *production* to theoretical *rated production* of the OWT where *duration* is the length of the time interval.

$$Production = Capacity\ Factor \times Rated\ Power \times Duration \tag{1.3}$$

[1] Applications of this research will be identified in sections 1.2.2 to 1.2.5.

Outages occur during intervals when the wind speed is too high or too low for the OWT to operate, while an OWT is being maintained and when the OWT is experiencing a fault that has caused it to stop. Operators plan maintenance activities with the intention of avoiding or reducing lost production from outages to increase the capacity factor. Substituting equation 1.2 and 1.3 into equation 1.1:

$$EBITDA = Price \times Rating \times Capacity\ Factor - OPEX \quad (1.4)$$

EBITDA does not track changes in working capital, capital expenditure, taxes, or the interest rate.

To maximise *EBITDA*, maintenance planners try to understand the relationship between capacity factor and *OPEX*. Alternative approaches to OWT maintenance optimisation, other than maximising *EBITDA*, include the strategy that generates the lowest cost of energy, the strategy that generates the highest capacity factor, the strategy that generates the highest *time-based availability* (the proportion of time that an OWT is available to generate) and the strategy that generates the highest *production-based availability* (the ratio of actual *Production* to the *Production* that would be expected of a fully available OWT) (DNV GL, 2017).

As *OPEX* increases, capacity factor increases but the detail of this relationship for each farm is confidential. This relationship, when combined with equation 1.4, has the following implications:

- If, rather than maximising *EBITDA*, operators were instead minimising the cost of energy, less *OPEX* would be required which would mean committing less resources to maintenance.
- If, rather than maximising *EBITDA*, operators were instead maximising the capacity factor, more *OPEX* would be required which would mean committing more resources to maintenance.
- If, rather than maximising *EBITDA*, operators were instead minimising *OPEX*, they could set *OPEX* to zero, committing no resources to maintenance. This would lead to the failure of all of their turbines, resulting in capacity factor of zero and consequently no *production* of electricity.
- If, rather than maximising *EBITDA*, operators were instead minimising *OPEX* for a specified minimum *time-based availability* or capacity factor, then the resources that they committed to maintenance would depend on what their contract specified. Such an arrangement has been used in the offshore wind sector but is not assumed in this thesis.
- Another option is to maximise *time-based availability*. *Capacity factor* has the advantage over *time-based availability* that it depends on *production*; and *production* takes account of the weather: there is no value in an OWT being available during intervals of either too low or too high wind speed and there is less value in it being available during intervals of wind speed lower than rated speed, where the *power output* is less than the *rated power*, than above rated speed, where the *power output* is the *rated power*.

- Another option is to maximise *production-based availability*. This is a useful and popular indicator, but it is not used in this thesis because Ørsted's Advanced Analytics Lab use *EBITDA*.

Nielsen et al., 2011, approach maintenance optimisation by minimising the expected total costs including the cost of lost production, equivalent to maximising *EBITDA*. As well as maximising *EBITDA*, operators also protect the future value of their assets by sufficient maintenance to avoid wear out. For readability, this thesis will refer to *EBITDA* as "productivity".

1.2.2 Maintenance Scheduling

This section will present the state of the art in the scheduling of wind turbine maintenance.

In contrast to corrective maintenance, preventative maintenance is performed regularly while the equipment is still operational to lessen the likelihood of it breaking down (Dao et al., 2018). OWT preventative maintenance is planned using scheduling tools developed from health history records that will be described in section 2.1.

Maintenance optimisation aims to determine maintenance plans that balance maintenance costs such as parts and labour against the consequences of not maintaining the plant such as loss of power production. It requires models of the uncertainties associated with wind farm inspection and maintenance such as dependencies among components, weather-dependent access to the wind turbines, stochastic demand for spare parts and availability of labour. (Shafiee and Sørensen, 2019, Seyr and Muskulus, 2019). Stock-Williams and Swamy, 2019, demonstrate that automated maintenance planning can significantly improve productivity.

Yürüşen et al., 2020, present a decision support system for the maintenance of onshore wind turbines. They consider a generator replacement operation that requires a crane. Safe working rules limit the wind speed and the wind gust speed at which work is allowed at different locations on the turbine. To find weather windows when the tasks could be performed, they use records of wind speed and wind gust speed with maintenance records that show how long each task takes. They identify all the possible scheduling combinations; an optimisation technique known as a brute force search. In order to calculate lost production, they combine this model with electricity price records and use the wind speed which enables them to identify the best times to carry out maintenance. These researchers simply use historical records of the duration of tasks and of the weather whereas a real-world planner must deal with the uncertainties in these forecasts.

There are opportunities to use data more intelligently to further improve maintenance scheduling. Work Orders (WO) will be discussed in section 2.1.3; they contain information on what work was done and they can be used to identify the cost of repairs. Outages will be discussed in section 2.1.2; when

labelled with a failure mode, they indicate the failure rate for that failure mode and can be used to estimate the lost production from each outage. Linking WOs to outages would help because the integrated information could provide valuable insights into historical costs of maintenance and of lost production. These insights could be used to further optimise maintenance scheduling from a logistical perspective by providing more robust information to the models. There is therefore the potential to develop better maintenance scheduling tools by linking WOs to outages. Literature searches up to 2021 did not find such techniques described and this thesis will not fill that gap, but it will present novel techniques that enable such further work by joining WOs to outages. An enriched health history of the machinery under study could be used by engineers and data scientists to develop better maintenance scheduling tools.

1.2.3 Condition-Based Maintenance

Condition-Based Maintenance (CBM) is a maintenance strategy that monitors the actual condition of an asset, such as a specific component in a wind turbine, to identify a requirement to carry out a maintenance intervention that is in addition to scheduled maintenance. CBM can reduce the requirement for scheduled maintenance while still avoiding delays and consequent production loss in repairing a failed turbine.

Machinery operators are attempting to move from planned and reactive maintenance to proactive CBM. CBM can help operators to avoid lost production by forecasting faults and making repairs prior to failure but, to predict the future, we need a detailed knowledge of the past.

This section reviews the state of the art for CBM. It concentrates on wind energy but it also samples other sectors and other industries. These samples indicate that wind energy is at the forefront of the development of CBM, but that it is also developing in solar energy (Mellit and Kalogirou (2011), Dagnely et al., (2015), Dong et al., (2017)), fossil fuels (Doostparast and Doostparast, 2018), electricity transmission (Sheng et al., 2018), naval engines (Cipollini et al., 2018), transport (Wang et al., 2018), capital goods (Arts et al, 2019) and aviation (Liu et al., (2018), Luo et al., (2018)).

Operators are introducing CBM both within the offshore wind energy sector and in other sectors. CBM uses wear out models to predict failures. It requires prognostic models and these require the determination of the health history of the machinery. Health history data such as breakdowns and other maintenance activities are required for the development of models that forecast breakdowns, facilitating preventative maintenance. Researchers have published wear out models that are based on vibration analysis (Crabtree et al., 2010, Bach-Andersen et al., 2015, Koukoura et al., 2017, Artigao et al., 2018, Carroll et al., 2019), on acoustic analysis (Wang et al., 2018), on oil particle analysis (Crabtree et al., 2010, Feng et al., 2013) and on data from temperature sensors (Garcia et al., 2006, Zaher et al., 2009, Feng et al., 2013, Kusiak et al., 2012, Godwin et al., 2014, Qiu et al., 2016, Bach-

Andersen et al., 2017). These models each predict failure with different confidence and over a different prognostic horizon, from a day to several months. These models are typically based on an understanding of a failure mode and that understanding is based on failure data. Applications of the health history for CBM include probabilistic models (Sheng et al., 2018, Wang et al., 2018), physics models (Qiu et al., 2016, Gray and Watson, 2010) and models developed using supervised machine learning methods (Godwin and Matthews, 2014, Hu et al., 2016).

A fault in a mechanical component tends to reduce its efficiency and consequently to cause its temperature to be higher, all else being equal, than it would be in a fault free component. Garcia et al., 2006 use wind turbine gearbox temperature data with data on whether the 2 cooling fans are on or off from a single gearbox fault. They predict that the fault was in the main bearing and also that the main shaft was misaligned, but they do not test this hypothesis.

Zaher et al., 2009, use temperature, power and fan data from 26 wind turbines over 2 years but do not have fault data to label what is happening to the wind turbines. They identify some abnormal behaviour that could be indicative of a fault, but they have no way of testing whether or not it is.

Crabtree et al., 2010, use wind turbine gearbox vibration and oil debris ferrous particle count data covering a 5-month interval. During this interval a gearbox bearing became damaged and was replaced and they plot the data against cumulative energy generation rather than against time to show the fault retrospectively. They show that such comparison of these signals indicates that the component was developing a fault 2 months ahead of its replacement. They present results that indicate that a prognostic model could be developed but because they only use a single instance of a fault, they cannot address whether such a model would work across multiple instances.

Gray and Watson, 2010 use wind turbine power and rotational speed data from 160 wind turbines, 6 of which experienced serious gearbox failure in quick succession, to present a physics of failure approach to wind turbine CBM that uses design data such as the dynamic load capacity of the gearbox bearings to estimate the fatigue damage for each failure mode. They calculate the high cycle fatigue Damage (D) for a specific failure mode. They refer to *reliability* as the probability that a turbine will survive at a damage exceeding D and they estimate the relationship between the reliability for the failure mode and D. They then use the failure rate to identify correction factors to their wear out model. Similarly, operators use failure, inspection and sensor data to update their wear out model for each failure mode for each model of wind turbine.

Feng et al., 2011 use wind turbine gearbox vibration, oil pressure and filter status, temperature, power and generator speed data. They report that an operator “achieved success in detecting a number of bearing faults in both gearboxes and generators” and that faults can be detected by using multiple

signals, which they plot against cumulative energy generation rather than against time to retrospectively show the faults.

Kusiak et al., 2012 use wind turbine temperature, voltage, current and generator speed data covering two instances of an over temperature fault on a generator bearing. They use data from 10 turbines that do not record this fault to develop a normal behaviour model that predicts the bearing temperature. They use these models to predict the over temperature events on average 1.5 hours ahead of the fault occurring. They train their model on turbines that they assume to be healthy and then test it on turbines that have experienced a fault. This does not constitute testing their model to see whether it correctly identifies faults on unlabelled data.

Godwin et al., 2014 use wind turbine temperature, power, wind speed and generator speed data covering one instance of a gearbox failure. They train a model on three turbines that do not suffer the fault and test it on another two turbines that also do not suffer the fault and on the single turbine that does. Because they only use a single instance of a fault, they cannot address whether such a model would work across multiple instances.

Bach-Andersen et al., 2015, use vibration data from 80 main bearing[2] failures on different turbines to compare two prognostic models. For each failure, they use the preceding 6 months of vibration data and an equal quantity of vibration data that they label as “non-fault”. They divide the 80 turbines into 40 for training, 15 for validation and 25 to test their models. Their models are a logistic regression and a convolutional network.

Bach-Andersen et al., 2015, use a robust statistical method for measuring the quality of prognostic models for wind turbine CBM. To present their method, this section defines the measures of the quality of classification that they use. These measures use the number of True Positives (*TP*), False Positives (*FP*), True Negatives (*TN*) and False Negatives (*FN*) as follows.

True Positive Rate (*TPR*), (otherwise known as recall, sensitivity or hit rate), defined by equation 1.5, measures the proportion of true matches that have been classified correctly.

$$TPR = \frac{TP}{TP + FN} \tag{1.5}$$

[2] The main bearing of an OWT is an important component and is typically fitted with vibration sensors. It mounts the turbines rotor on to its nacelle; that is the housing that holds all of the generating components.

False Positive Rate (*FPR*), (otherwise known as fallout or false alarm ratio), defined by equation 1.6, measures the proportion of false positives in the total number of negatives.

$$FPR = \frac{FP}{TN + FP} \tag{1.6}$$

A Receiver Operating Characteristics (ROC) curve plots TPR against FPR and is robust against imbalanced classes (Christen, 2012). The Area Under a ROC Curve (*AUC*) predicts the probability that a classifier will rank a positive sample higher than a negative sample (Fawcett, 2006). Bach-Andersen et al., 2015, show that their convolutional network outperforms their logistic regression by plotting AUC against time to failure. Bach-Andersen et al., 2015, have access to vibration data that is labelled with rich information about a specific fault, a wind turbine main bearing failure. This enables them to train, validate and test a prognostic model for this fault. Bach-Andersen et al., 2015, is an example of health history data being used to develop a model that predicts main bearing failures.

Standard solar energy Supervisory Control and Data Acquisition (SCADA) systems only collect current and voltage data (Dong et al., 2017) and, unlike on OWTs, temperature sensors are not routinely fitted. Dong et al., 2017, combine SCADA data with weather data to determine a normal behaviour model that identifies the least healthy strings of solar panels.

Koukoura et al., 2017, use wind turbine gearbox vibration, power and generator speed data from one instance of a gearbox tooth issue on the pinion of the intermediate stage of the gearbox. To diagnose the state of the gearbox they calculate reference torque using the produced electrical power and generator speed and they also calculate health indicators on the second harmonic of the gear mesh frequencies of the vibration signals. They train a decision tree classifier, a supervised learning method, both on the health indicators and on the reference torque to distinguish signal variations due to loads from signal variations due to faults. The loads on a wind turbine fluctuate so this is an appropriate approach. They present results indicating that prognostic models could be developed.

Bach-Andersen et al., 2017, use wind turbine temperature data from 22 main bearing failures on different turbines to compare four prognostic models. They plot the area under the ROC curve against time to failure. They have access to temperature data which is labelled with information about a specific fault, a main bearing failure. This enables them to train, validate and test a prognostic model for this fault. They recognize that this model could be used together with vibration analysis.

Wang et al., 2018 use data from 4 acoustic sensors on a single set of railway points. By comparing new observations to a predicted baseline, they detect by a probabilistic approach that one of the rails has no damage but that the other has a crack in it. This result is verified by an urgent site visit and repair. Such acoustic sensors are not appropriate for fitting to the whole railway as they require a power source. They might alternatively be fitted to a train.

Doostparast and Doostparast, 2018, present a methodology for forecasting corrosion in oil pipelines. They forecast points of corrosion, approximating the rate of corrosion using a non-homogeneous Poisson process.

Artigao et al., 2018, use high frequency (1.5 kHz) current data from both the stator and the rotor of a wind turbine generator with vibration, power and generator speed data from a single instance of a fault caused by an imbalance in the shaft between the gearbox and the generator that the operator had misdiagnosed as a generator bearing fault. They use their vibration data to validate their result from the current data but they do not combine the current data with the vibration data. They present results indicating that prognostic models could be developed.

Carroll et al., 2019, use wind turbine gearbox vibration, temperature, power, wind speed and generator speed data. They use operational logs and failure logs to identify two gearbox failure modes that occur on different models of wind turbine: 200 instances of a gearbox bearing issue on the low speed planetary stage of the gearbox and 28 instances of a gearbox tooth issue on the pinion of the intermediate stage of the gearbox. For both failure modes, they use 70% of their data to train a multi-class Artificial Neural Network (ANN) and 30% to test it. They use the ANN with the vibration data to predict whether the time to failure is 1, 2 or 3 months or whether it is more than 3 months. They also use the ANN with the vibration data to predict whether the time to failure is 1 to 2 or 5 to 6 months or whether it is more than 6 months. They show that accuracy is higher with the vibration data than with the other data but this is not a direct comparison because they have different classes for the two data sets. They do not combine the vibration data with the other data.

Bach-Andersen et al., 2018, use wind turbine vibration and generator speed data from gearbox faults that they break down into 85 rotor bearing faults, 63 planetary stage bearing faults and 103 helical stage bearing faults across two models of wind turbine to compare five prognostic models. They do not report which faults are on which of their two models of wind turbine. For each model of wind turbine, they use 65% of these faults for training, 10% for validation and 25% for testing. Their prognostic models are a logistic regression, an ANN and a deep ANN, meaning an ANN with multiple layers to extract abstract features. For both the deep and shallow ANN, they compare the results of solving tasks separately and of applying multi-task learning (in which multiple learning tasks are solved at the same time to exploit commonalities and differences across tasks). They plot the area under the ROC curve against time after retrospective detection by a human expert, showing that their deep ANNs recognised features in the vibration data at the same time as retrospectively identified by their human expert. Bach-Andersen et al., 2018, have access to vibration data that is labelled with rich information about these three faults. This enables them to train, validate and test a prognostic model for this fault.

Lei et al., 2019, use vibration data from a model wind turbine in a wind tunnel. They simulate eleven mechanical faults in the drive train. They compare a variety of methods for the classification of their simulated faults from the sensor data. These are: support vector machines; supervised learning models that construct a set of hyperplanes in a high-dimensional space, multilayer perceptrons; a class of feedforward neural network, 2 layer recurrent neural networks; a class of neural networks where connections between nodes form a directed graph along a temporal sequence and convolutional neural networks; a class of deep neural networks, used for analysing visual imagery. They present the best results using long short term memory networks; a recurrent neural network architecture with feedback connections used in handwriting and speech recognition. Their method of fault classification could be of interest to implement on real wind turbines.

Wang et al., 2020, use vibration data from a single wind turbine gearbox bearing failure. They combine physical knowledge of the bearing's characteristic frequencies of typical defects relative to shaft speed with Bayesian inference to quantify uncertainty. Bayesian inference is a statistical technique that updates the probability for a hypothesis as more information becomes available. They present an approach that estimates the remaining useful life of such bearings as a probability distribution. Operators can plan preventative maintenance more effectively when they better understand the uncertainty that exists about the condition of their plant.

The literature shows a high level of focus on wind turbine gearboxes and main bearings, identified as important components for reliability in Figure 1-5 and Figure 1-6 but not as the only important components. Operators use outage data (which will be discussed in section 2.1.2) to identify which other components are also worthy of study.

This section has shown that even existing health history data, such as that available to this research, described in section 2.2, was not available to many other researchers and that this lack of labels has held back research into CBM. There are opportunities to use data more intelligently to further improve CBM. When outages are labelled with a failure mode, they indicate the failure rate for that failure mode. Work Orders (WO) contain information on what work was done and will be discussed in section 2.1.3. There is therefore the potential to develop new CBM models and improve existing ones by linking WOs to outages. Literature searches up to 2021 did not find such techniques described and this thesis will not fill that gap, but it will present novel techniques that enable such further work by joining WOs to outages. An enriched health history of the machinery under study could be used by engineers and data scientists to develop both more accurate prognostic models and a better understanding of the confidence of these forecasts.

This section has reviewed the literature on CBM; the prognosis of faults. Troubleshooting requires the diagnosis of faults. Both CBM and troubleshooting involve the classification of the failure mode and, consequently, the literature on CBM also pertains to troubleshooting.

1.2.4 Troubleshooting

Troubleshooting is the activity of repairing a faulty OWT by the replacement of components (Walford, 2006, Tang et al., 2019). This section will present the state of the art in OWT troubleshooting.

Data scientists analyse failure histories and derive troubleshooting guides from them. These guides are key to the technicians' diagnosis of faults, alongside their expert knowledge. Technicians refer to a trouble shooting guide for advice on how to diagnose the failure mode and the repair activity that is most likely to be effective.

There are opportunities to use data more intelligently to further improve troubleshooting. Work Orders (WO) will be discussed in section 2.1.3; they contain information on what work was done including what materials were consumed. Outages will be discussed in section 2.1.2 but, when labelled with a failure mode, they indicate the failure rate for that failure mode.

Each troubleshooting guide applies to a single alarm code, rating and manufacturer. It contains a fault tree, representing the possible failure modes, their predicted probability, and the recommended repair activities. Operators derive the information that they use to generate the fault tree from WO records. WOs are sometimes labelled with an alarm code indicative of the failure mode and it is only these WOs that are currently used to generate the fault tree.

This research identified two gaps in the literature on troubleshooting. Firstly, literature searches up to 2021 did not find Enriched Health History (EHH) information described and as such its use to further improve troubleshooting fault trees is not described either. The health history enrichment process presented in this thesis uses alarm data to identify the failure mode of each WO and this increases the amount of information available to the authors of the troubleshooting guides. The use of an EHH (specific to the WT's rating and manufacturer) provides data that can increase the accuracy of the probability assigned to each branch of the fault tree. Increasing the accuracy of the fault tree probabilities for a failure mode can avoid unsuccessful repairs, reducing maintenance work and avoiding lost production.

Troubleshooting guides also contain a list of spare parts that might be required to repair the specific failure modes indicated by the alarm code. The second gap identified in the literature is that techniques to increase the length of these parts lists to improve the probability that the repair team has brought the correct part to repair the fault are not described. Extending the troubleshooting guide parts list can avoid the lost production caused by the right part not being available. Bringing the correct parts to the OWT will be of increasing importance as the distance to shore increases and it could help to avoid the cost of an offshore spare parts store. Health history enrichment would fill this gap in the literature.

To repair a fault, technicians must identify which components to replace and, while it is not strictly necessary to identify the failure mode, this is desirable because the root cause can lie in a component other than those that are not working. For example, on one occasion, a fault on one wind turbine damaged an adjacent turbine so that the service team was sent to the wrong turbine (personal conversation with an Ørsted technician, 2019). Troubleshooting is not straightforward and finding the root cause requires experience.

Improving either the parts list or the fault tree reduces the duration of outages, reducing lost production and increasing productivity. There is therefore the potential to improve troubleshooting by linking WOs to outages. Literature searches up to 2021 did not find such techniques described and this thesis will not fill that gap, but it will present novel techniques that enable such further work by joining WOs to outages. An enriched health history of the machinery under study could be used by engineers and data scientists to develop improved troubleshooting guides. This research elicited the opinion from OWT experts that troubleshooting would be the most valuable application of an EHH.

1.2.5 Maintenance Effectiveness

The purpose of maintenance is to increase the reliability of plant. Each maintenance activity should increase the mean time to failure of the subsystem that was maintained relative to the subsystem's mean time to failure without that maintenance activity. This thesis will refer to the effect of reducing the mean time to failure of the subsystem that was maintained as the effectiveness. Garcia et. al., 2006, predict that operators could use effectiveness to optimise their planning and troubleshooting, increasing productivity. This section will present the state of the art in the measurement of maintenance effectiveness.

Lin et al., 2019, present a model that combines performance deterioration together with maintenance effectiveness and that would be applicable to a wide variety of assets including wind turbines. Sewers are flushed (a maintenance activity) to remove blockages (a failure mode). Lin et al., 2019, use dated sewer pipe flushing records, dated sewer pipe condition data (camera inspection reports that rate the condition of the pipe at five discrete grades and record the type of defects), sewer pipe attribute data (material, diameter, length, slope) and sewershed area data (the land area that drains into the pipe).

A Markov chain model is a statistical technique that describes a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. Lin et al., 2019, model the deterioration process using a continuous-time Markov chain model with a deterioration intensity matrix that estimates the probability that, during a time step, a pipe will transition from each of the condition grades to each of the other condition grades. They model the maintenance process using a discrete-time Markov chain model with a maintenance effectiveness matrix that estimates the

probability that, during a flush, a pipe will transition from each of the condition grades to each of the other condition grades.

Bayesian inference is a statistical technique that updates the probability for a hypothesis as more information becomes available. A Markov chain Monte Carlo simulation is a statistical technique that takes samples from a probability distribution. Lin et al., 2019, estimate the parameters for their performance deterioration and maintenance effectiveness models using Bayesian inference with Markov chain Monte Carlo simulations.

Lin et al., 2019, compare two methods that predict the probability of reaching a given grade at a given time. The first method uses the mean of the estimate of each parameter value while the second method uses the full distribution of the estimate of each parameter value. They show that using the full distribution of the estimate of each parameter, rather than using the mean, predicts far lower probabilities that, under a given maintenance regime, a given pipe will get blocked. Using the full distribution is a more robust approach than using the mean. They recognise that this huge difference could result in totally different strategies for flushing programs. Wind turbine operators both plan maintenance and measure maintenance effectiveness using the same type of deterioration and maintenance effectiveness models as those presented by Lin et al., 2019.

There are opportunities to use data more intelligently to further improve the measurement of maintenance effectiveness. Work Orders (WO) will be discussed in section 2.1.3; they contain information on what work was done. Outages will be discussed in section 2.1.2 but, when labelled with a failure mode, they indicate the failure rate for that failure mode. If a repair was successful, then the time to failure for the modes effected by the repair would tend to increase. There is therefore the potential to measure maintenance effectiveness by linking WOs to outages. Literature searches up to 2021 did not find such techniques described. This thesis presents techniques that link WOs to outages and this will enable further work measuring maintenance effectiveness. An enriched health history of the machinery under study could be used by engineers and data scientists to develop more accurate measures of maintenance effectiveness and a better understanding of the confidence of these measures.

1.2.6 Conclusion to the Maintenance of Offshore Wind Turbines

The literature search found more literature on CBM than on maintenance scheduling, troubleshooting or on maintenance effectiveness. This does not imply that CBM is more important to operators than the other aspects of maintenance listed; operators have tended to collaborate with academic researchers and to publish their own papers on CBM, while they have tended to keep their work on the other aspects of maintenance more in-house. Improvements in troubleshooting are seen as

tending to avoid more lost production than improvements in CBM (personal conversation with an Ørsted wind turbine data analyst, February 2019).

Figure 1-7 shows an OWT CBM architecture. The figure is based on the approach of Garcia et. al., 2006, but this approach is typical of how operators plan OWT maintenance today. Sensor data from each turbine is interpreted by comparing it to sensor data from nearby turbines, to meteorological data and to models trained on data from the past. This comparison identifies anomalies that could be indicative of faults and these are interpreted by OWT experts to diagnose whether a repair activity should be planned. The health condition of each turbine is assessed to predict faults and if a fault is considered likely then the component can be replaced prior to failure. The effectiveness of maintenance activities is appraised by statistical techniques that address whether down time was avoided by the maintenance activity.

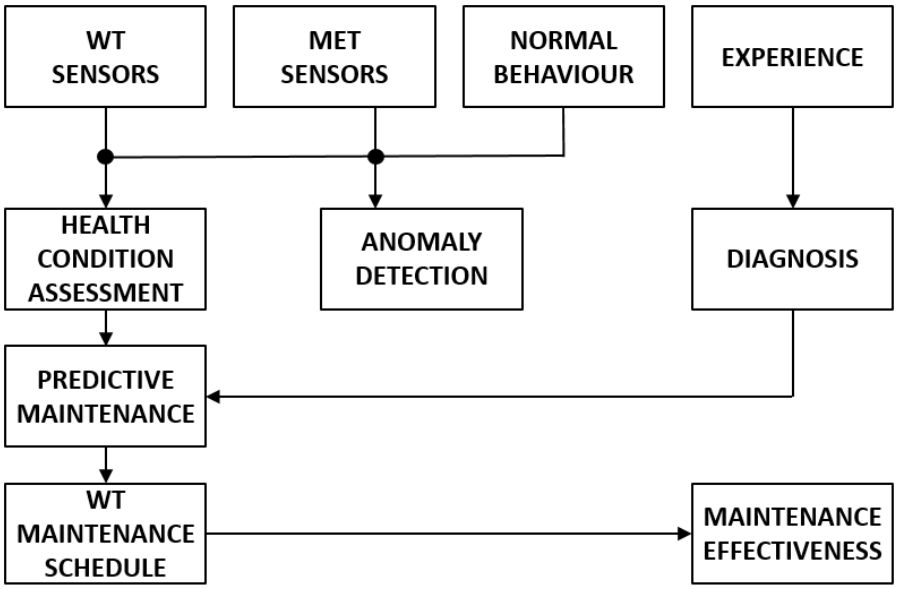


Figure 1-7, Typical OWT CBM Architecture. Adapted from Garcia et. al., 2006

This section has identified the gap in the literature that multi feature record linkage techniques have not previously been used to improve wind turbine maintenance strategies; or to improve maintenance strategies for any other type of machinery either. Section 2.2 will present the state of the art for record linkage techniques.

1.3 Research Questions

This research benefitted from access to Ørsted's valuable and confidential records of wind turbine health history. It applies record linkage techniques to these records to improve maintenance. The Research Questions (RQ) that this thesis will address are:

RQ1 To enable improvements in WT CBM & troubleshooting, how can WT health history be enriched?

RQ2 How can the quality of the Enriched Health History (EHH) be validated?

RQ3 How can the richness of historical data on wind turbine health be measured?

This thesis will present new techniques that link together existing records of the health of OWTs to determine and validate an EHH. In this thesis, the statement that the health history has been 'enriched' means that the records have been made mode useful for specific maintenance applications such as maintenance scheduling, troubleshooting, condition-based maintenance or for the measurement of the effectiveness of maintenance activities.

Figure 1-8 describes how, by linking WOs to outages, the EHH will facilitate improvements to the diagnosis of WT faults, helping with repairs[3], and to their prognosis, helping to avoid faults[4]. The dotted circle indicates the scope of this thesis. Methods developed in this project will be applicable to other record linkage applications as well. The EHH will also facilitate improvements to the scheduling of maintenance activities and to the measurement of the effectiveness of maintenance activities.

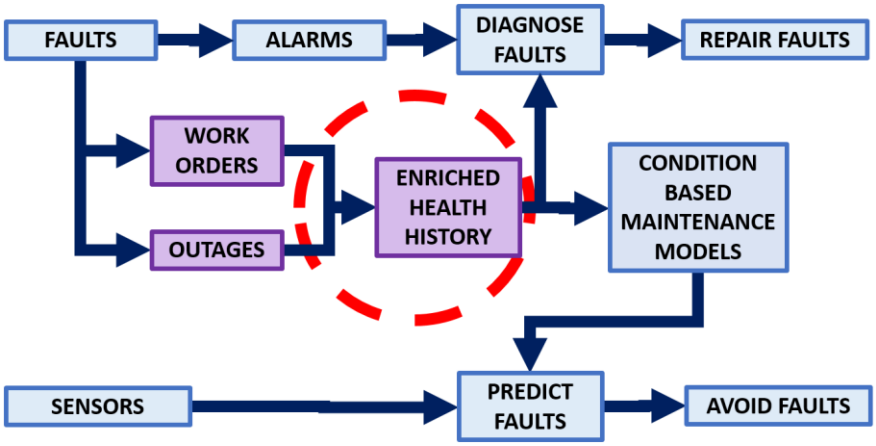


Figure 1-8, Enrichment of Health History for CBM & Troubleshooting

[3] Troubleshooting for repairs was described in section 1.2.4

[4] Avoiding faults by condition based maintenance was described in section 1.2.3

1.4 Research Process

This research project involved presentations to conferences and research placements as detailed in Table 1-1.

Discussions with Ørsted data scientists during placement 1 in their office in Gentofte, Denmark, identified that a process for the enrichment of offshore wind turbine health history would be a useful contribution. Some of the record linkage techniques that will be presented in chapter 4 were developed by these discussions.

By presenting at conferences, this research benefitted from questions from the audience that facilitated improvements in how this thesis is presented. Attendance was also a method, alongside monitoring the published scientific literature, of identifying other researchers working in related areas. The most important of these was Eric Salo, whose research into data mining wind turbine work orders (Salo et al., 2019) is of interest to developers but covers a different area to this research.

Event	Subject	Location	Date	Duration
Research visit		Gentofte	May 2017	1 week
Placement 1		Gentofte	August 2017	18 weeks
Research visit		Wind Farm	January 2018	1 day
Placement 2		Gentofte	January 2019	1 week
Wind Energy Science Conference	Wind Turbine Enriched Health History	University College Cork	June 2019	3 days
European Academy of Wind Energy PhD Seminar	Validation of Wind Turbine Health History	Centrale Nantes	October 2019	2 days

Table 1-1, Placements and Conferences

1.5 Constraints

This research benefitted from access to Ørsted's valuable and confidential records of wind turbine health history. Ørsted's policy is that their confidential data may not be copied on to non-Ørsted machines and, to comply with this policy, this research was done entirely using a laptop computer supplied by Ørsted.

The process for writing this thesis involved the identification of confidential material and the contents of this thesis are as approved for publication by Ørsted.

1.6 Thesis Structure

The structure of this thesis is as follows. The introduction chapter (1) presents a brief introduction to the offshore wind energy sector. The main literature review is chapter 2, Background, which presents the state of the art for OWT maintenance including the motivation for the enrichment of OWT health history and a description of the existing records of OWT health history that this project has benefitted from access to. It also presents the state of the art for record linkage. The next chapter (3) presents techniques for evaluating health history enrichment and poses, in detail, the research questions that this thesis will address.

The first technical chapter (4) presents an overview of the Process for the Enrichment of OWT Health History (PEOHH) developed in this research and of how the PEOHH will be validated. It presents twelve record linkage techniques organised into four sub sections; four timestamp based techniques, a technique that uses the records of visits to the wind turbine, a technique that considers the recorded type of maintenance and lastly six failure mode based techniques. For each record linkage technique, the chapter presents its method and the hypothesis that it might be effective as part of an ensemble of techniques; it does not test these hypotheses.

The second technical chapter (5) offers two example work orders and uses them to illustrate the PEOHH by linking them to records of OWT power outages.

Chapter 6 reviews statistical techniques that quantify the uncertainty of measures that will be used in chapters 7 and 8:

- Section 6.1 will review techniques for estimating the uncertainty of the extent to which a binomial proportion from of a sample is representative of a population.
- Section 6.2 will consider the estimation of the uncertainty of a difference between two such uncertain estimates.
- Section 6.3 will investigate the probability that one such an uncertain estimate is greater than another.

Applying these techniques offers, for the first time, an understanding of the uncertainty of the quality of record linkage. Understanding uncertainty informs maintenance decision making which can improve productivity.

The primary results from this research are presented in chapters 7 and 8. Chapter 7 reviews techniques for global optimisation and presents the optimisation of the weights and thresholds used by the PEOHH. It tests the hypotheses that were presented in chapter 4. Chapter 8 demonstrates that the PEOHH has indeed succeeded in enriching the health history. Chapter 9 is a critical review of this

research and the final chapter (10) presents the conclusions and indicates some useful directions for further work.

1.7 Contribution of this Thesis

Outages can be due to routine maintenance, to environmental conditions, to problems with the grid or to faults. This thesis will show that the major gap in knowledge impeding progress in this field is matching the information available from maintenance logs to an existing database of outages. The research contribution made by this project will be to address this gap in knowledge with the advantage of academic access to commercially sensitive real-world fault data.

The original contributions of this research are:

- The application of multi-feature record linkage techniques to maintenance data
- The application of statistical techniques for the interval estimation of a binomial proportion to record linkage techniques
- The estimation of the distribution of the coverage error of statistical techniques for the interval estimation of a binomial proportion

The main contribution of this research is a process for the enrichment of offshore wind turbine health history.

2 Background

This background chapter is divided into 5 sections. This research has benefitted from access to existing records of OWT health history and they are described in section 2.1. This research links those records together to determine an Enriched Health History (EHH) and the state of the art for record linkage is presented in section 2.2.

Classification techniques predict the class of a binary dependent variable by analysing a dataset. Section 2.3 reviews standard classification techniques in preparation for section 4.4.4.3.2, which experiments with the application of a classification technique for record linkage.

Global optimisation is a branch of applied mathematics that attempts to find the global minima or maxima of a function by choosing the system parameters. Section 2.4 will review two appropriate techniques for global optimisation in preparation for section 7.6, which will use the most appropriate of these techniques for the optimisation of the weights and thresholds used in the PEOHH.

Section 2.5 is the conclusion to the background.

2.1 Existing Records of Wind Turbine Health History

This research shows that an Enriched Health History (EHH) can be determined by combining four existing OWT health history databases. These are databases of alarms, outages, Work Orders (WO) and material consumption. This section describes each of these four databases, their salient features (also known as tags) and how each database is generated. It reviews literature on similar databases and this comparison shows that wind turbine records are at the forefront of maintenance technology.

This section describes Ørsted's records (insofar as confidentiality agreements allow). This project has not had direct access to other operators' records which, like Ørsted's records, are also confidential. To apply the Process for the Enrichment of OWT Health History (PEOHH) developed in this research, other wind turbine operators, as well as operators in other sectors, would need to adapt it to their records.

2.1.1 Database of Alarms

Many of the components of an Offshore Wind Turbine (OWT) are fitted with sensors that monitor their physical parameters such as temperature, vibration or contamination of the oil. Each OWT contains a Condition Monitoring Unit (CMU) which collects data from the sensors. The CMU is programmed with a set of models that interpret the sensor data with reference to thresholds to identify environmental, operational and fault conditions on the basis of which they assert alarms. Each such model is uniquely identified by an alarm code. A free text description, such as "cabinet too hot", is linked to each alarm code.

The CMU maintains a database known as the alarm log; for each alarm event it records the alarm code, start time and end time. These two timestamp features define the duration over which the alarm was asserted (Qiu et al., 2012).

When technicians visit an OWT to carry out corrective maintenance, they first check the database of alarms which offers a rich record of the recent health history of the OWT and informs the technicians of what seems to be the problem. This information helps the technicians to decide which tests or repair activities to use to find the fault and to repair it.

Dagnely et al., 2015, present a problem faced by data scientists attempting the analysis of alarm data from photovoltaic plants in which the inverters and the monitoring systems originate from different manufacturers, models and versions: in this situation the alarm labels can differ. They present a methodology for the integration of heterogeneous, machine generated alarm data. The OWT health history records that this thesis has had access to do not suffer from this problem of heterogeneous data.

Alarm data is used to automatically diagnose the failure mode (Qiu et al., 2012, Gonzalez et al., 2016).

Kusiak and Li, 2011, use alarm data to identify the health history of four wind turbines, enabling them to develop CBM models. Their approach is like that used, on a larger scale, by Ørsted. They observe that better prediction performance can be achieved with higher quality data. The techniques presented in this thesis will enable operators to make better use of their data to enable such improvements.

Leahy et al., 2018, showed that alarm data can be used to automatically identify intervals of faulty operation.

Papatzimos et al., 2019, showed that the frequency of alarms for each OWT subassembly can be predicted from the wind speed and the turbulence intensity.

2.1.2 Database of Outages

Operators refer to intervals when the OWT is not generating electricity as outages. Outages occur during intervals when the wind speed is too high or too low for the OWT to operate, when the OWT is being maintained and when the OWT is experiencing a fault that has caused it to stop.

Ørsted's database of outages is automatically generated from the alarm log. It contains two timestamps: the outage start time and the outage finish time. These define the duration of the outage. Ørsted have validated their database of outages' timestamps against the active power signal. These timestamps are very important to operators because they are indicative of the availability of the wind turbine, a key performance indicator, and so operators see them as very accurate; checking their accuracy is outside of the scope of this thesis.

Ørsted label each outage with an alarm code indicative of the failure mode using a combination of automatic and manual methods. Their automatic system selects an alarm code from the database of outages using a confidential algorithm. Ørsted's data scientists sometimes later manually adjust these labels to better reflect the failure mode by discussing what happened with the technical team involved in the repair. Literature searches up to 2021 did not find techniques to further automate and to further validate this diagnosis by joining work orders to alarms and to outages described.

The 'type' feature classifies each outage as either corrective, predetermined, 'Balance of Plant / OFfshore Transmission Owner' (BoP/OFTO), condition-based, environmental or unknown. BoP/OFTO refers to those outages caused by faults that are outside of the wind turbine. 'Environmental' refers to those outages caused by the weather: either too low or too high wind speed. This classification is useful because breaking the records of power outages down into these categories enables operators to identify the performance of different parts of their operation. If some outages were classified incorrectly then commercial decision makers would not have reliable data to work on, and so the accuracy of this classification is important. Literature searches up to 2021 did not find techniques to

investigate the accuracy of this classification by joining work orders to alarms and to outages described.

It is a safety requirement that when an OWT is visited to carry out maintenance, it is brought under local operation; a setting in which it cannot produce electricity. The ‘reset’ feature classifies each outage that has been classified as corrective as either a visit, a remote reset or an automatic reset. The ‘number visits’ feature is either zero or a positive integer. It is Ørsted’s estimate of how many times the OWT was visited during the outage. It is calculated using the ‘wind turbine in local operation’ alarm.

The database of outages records the duration of each outage and estimates its lost production as the average of energy generated by the farm’s working wind turbines.

2.1.3 Database of Work Orders

This section describes the database of Work Orders (WO). Each WO has an order number: a unique identifier for the WO. Figure 2-1 shows an example WO, number 80135873. Each WO is also labelled with a notification number: a unique identifier for the notification. There is one notification per WO and the notification is the example is number 1171848. The notification is the instruction to the wind farm planner to create the WO. Each WO is an instruction to a technical team to undertake a maintenance activity; in the example the hydraulic oil needs topping up. WOs often contain free text entries that describe maintenance activities in detail. Examples include when routine maintenance is provided, such as oil changes, as well as when more significant maintenance is performed, such as replacing key components. This data is often entered by the technical team undertaking the maintenance.

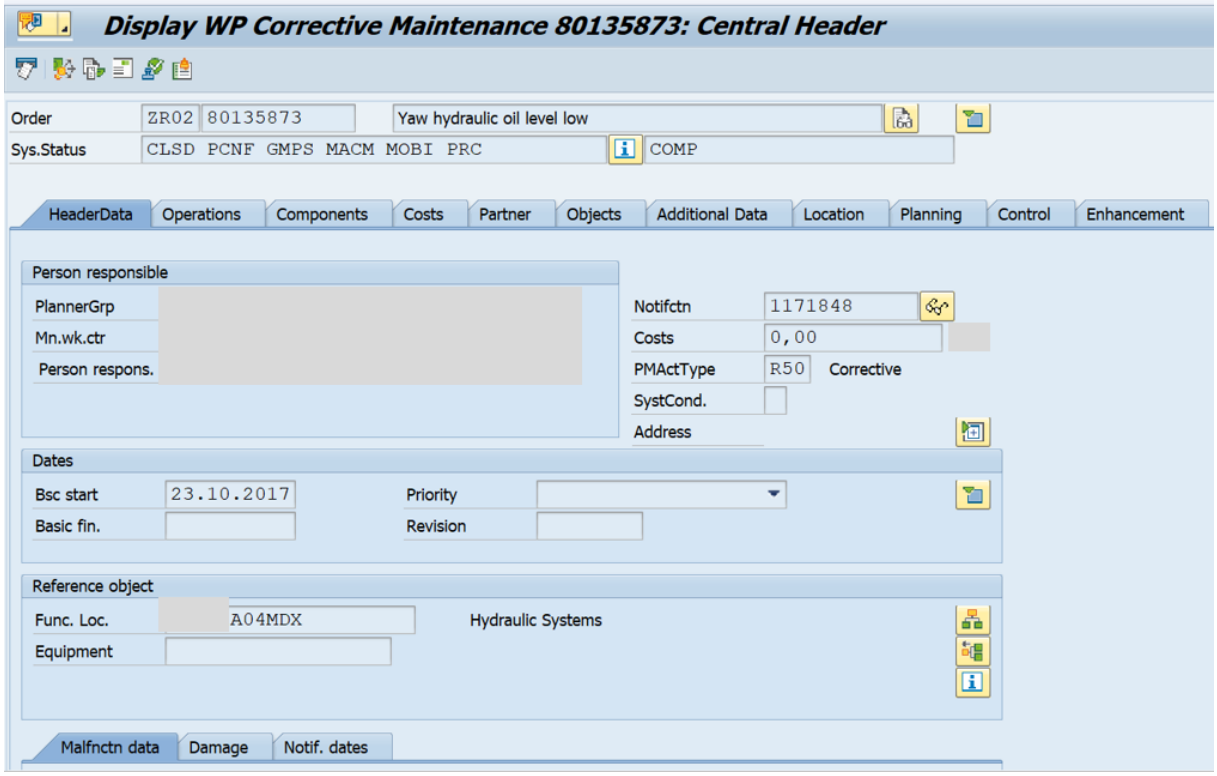


Figure 2-1, Example Work Order

WOs contain structured information about the turbine acted upon, semi-structured information about the subsystem acted upon and unstructured information about the activities performed. The unstructured information is in the form of free text. The free text fields can document alarm codes and / or specific actions undertaken by the technical team. The semi-structured and the unstructured data can both contain typographical or other errors (inevitable with a field entry system) and these

errors present a challenge when linking them together and for other valuable data mining applications that were discussed in section 1.2.

This research involved placements with Ørsted's data scientists who explained the key features of the WO data. It also included visits to wind farms, where conversations with the managers, planners and technicians who generate the data validated these descriptions and put them in a more detailed context.

2.1.3.1 Timestamps

The WO records contain various timestamps including the start date, the finish date, the notification date, the created on date and the malfunction start date. The start date refers to when the fault started. The finish date refers to when the fault ended. The notification date refers to when the notification was created. The created on date refers to when the WO was created. The malfunction start date refers to when the malfunction started. This feature was too difficult to extract from the database because, while it is accessible when extracting data on an individual WO, it is not accessible when extracting data on a list of WOs. Further work should find a way to extract it and investigate its usefulness. The advice elicited from technicians and from data scientists was that malfunction start date would be very useful. That is because the timestamps are all human generated and can be approximate or erroneous. The malfunction start date, on the other hand, is automatically generated from the alarm log. All the experts agreed that there would be no relationship between when a WO was created and when its matching outage occurred. These experts included a Planner Scheduler who creates WOs.

2.1.3.2 Functional Location

The 'Functional Location' (FL) feature is a standard taxonomy; a human generated feature entered by the planner; a string specified by the Reference Designation System for Power Plants standard for wind turbines (V.G.B. PowerTech, 2014). It contains a unique identifier for the wind farm. It identifies which OWT the WO relates to by giving the number of the row within the wind farm in which the OWT is located and the OWTs position along that row. It can include detailed information about which subsystem the WO relates to and even which component. The FL will always be accurate about which OWT the WO refers to because if an incorrect OWT ID had been entered by the planner then the technical team would visit the wrong OWT. Literature searches up to 2021 did not find descriptions of techniques to enrich the FL by joining work orders to alarms and to outages. This thesis will not fill that gap, but it will present novel techniques that enable such further work by joining WOs to outages. An enriched health history of the machinery under study could be used by engineers and data scientists to enrich the FL.

2.1.3.3 WO Type

The ‘type’ feature classifies each WO as either preventive (such as annual servicing), retrofit, inspections and surveys, condition-based or corrective. The feature is human generated by the planner who selects a type from a list of options.

The WO free text fields are the ‘description’ and the ‘long text’. These fields can be written in any natural language but farms where they are not in English are outside of the scope of this thesis. Further work could apply automatic translation techniques to WO data.

2.1.3.4 Description

The ‘description’ feature, otherwise known as ‘short text’, is a short, free text description of the WO. For corrective maintenance it often refers to an alarm code. It may contain the alarm code but it more often contains an abbreviated reference to the standard text description of the alarm code.

Salo et al., 2019, show that this feature can be exploited to cluster WOs by their failure mode using manual or automatic, text mining methods. This application might benefit from the addition of other features of the health history data.

2.1.3.5 Long Text

The ‘long text' feature is a free text description of the notification and of the WO of unlimited length. For corrective maintenance it often contains semi structured recent entries from the alarm log that include alarm codes. These alarm log entries are automatically copied in when the notification is created and are typically error free. It can also contain unstructured notes made by the maintenance team relating to faults or to maintenance activities, particularly if these are considered unusual.

2.1.3.6 Alarm Code

Some WOs are labelled with an alarm code[5]. This label is used to identify the failure mode. Ørsted decide which alarm code to use by a combination of automatic and manual methods, where wind turbine experts’ identification of the most important alarm code supplements automatic methods.

2.1.4 Database of Material Consumption

The material consumption database lists what parts were used in the maintenance of the OWTs. Each Material consumption Line Item (MLI) refers to a single part number and is assigned to an order number. Some WOs have no material consumption line items assigned to them while others have many. Materials include replacement parts as well as consumables such as oil, grease or paint. The

---

Alarm codes were described in section 2.1.1.

‘material’ feature is the part number: the identifier of the design of the part. The ‘description’ feature describes the part. The material consumption records contain various timestamps: the ‘posting date’ and the ‘reserved date’ which are both typically input by the planner. The ‘reserved date’ refers to when the part was reserved for use and the ‘posting date’ refers to when the part was used.

2.1.5 Conclusion to Existing Records of Wind Turbine Health History

This section has described four offshore wind turbine databases that this research has had the benefit of access to. Each of these four databases relates to the same maintenance activities on the same wind turbines and was generated as part of the maintenance of Ørsted’s wind farms. The database of WOs and the database of material consumption are linked together by order numbers. The database of outages is generated from the database of alarms and so linking these two together will be fairly trivial. The gap in knowledge is how to link these two sets of maintenance data together. This thesis will address this gap by presenting techniques that join WOs to outages to identify an enriched health history. It will do this using features of all four of the databases that have been presented in this section.

2.2 Existing Record Linkage Techniques

Record linkage techniques aim to determine whether pairs of data records describe the same entity. This thesis presents new record linkage techniques and applies these as well as existing record linkage techniques in the field of Offshore Wind Turbine (OWT) maintenance. This section explores the state of the art of record linkage techniques. Record linkage becomes non-trivial when the records do not share a unique key. OWT outages are not currently labelled with an order number, making it uncertain which records relate to the same event. Methodologies for linking records under uncertainty are known as Probabilistic Record Linkage (PRL) techniques.

This review included Durham University Library searches of the literature of for terms including “wind turbine”, “maintenance”, “record linkage”, “work order” and “condition-based maintenance”. It found that there is very little literature on linking wind turbine maintenance records to records of alarms or of outages. The one published reference to it is in Papatzimos et al., 2017, in which OWT WOs are linked to alarms using a single feature; a timestamp. Papatzimos et al., 2017, do not publish any measures of the quality of the classification of this technique, however the literature reviewed in this section and the results presented in this thesis show that the quality of record linkage can be increased by using more features. Record linkage techniques have been used to link medical records (Sayers et al., 2015, Nasseh and Stausberg, 2016, Oliveira et al., 2016), address data (Churches et al., 2002, Comber et al., 2019, Lin et al., 2019), census data (Jaro, 1989, Smith et al., 2016) and genealogical records (Wilson, 2011) and they have been used to detect duplicate internet search results (Hajishirzi et al., 2010). This thesis will present new record linkage techniques and will apply these as well as existing record linkage techniques in the field of offshore wind turbine maintenance.

PRL techniques join two databases together to create a new database in which each row represents one Pair Of Linked Records (POLR). They compare ensembles of partially-identifying, non-unique data items between pairs of records (Dunn, 1946, Churches et al., 2002). They could compare each record in one database with each of the records in the other database but this would be computationally expensive and can lead to inaccuracy (Sadinle & Fienberg, 2013) so they instead split the data into smaller blocks to disregard very unlikely POLRs. They compare features in the data sets being linked together to generate a comparison vector containing comparison features. For example, in address matching, comparison features typically include “house number matches” and “post code matches”. PRL techniques combine an ensemble of comparison features by weighting each dimension of the comparison vector to give an overall score indicative of the probability that each POLR is a true match. To compute a score for a given POLR, a weight is added for each feature. If the two records agree on the feature, a so called ‘agreement weight’ is added. If they disagree, a ‘disagreement weight’ is added. If one or both records have no data for the feature, then neither weight is added. PRL techniques require a sample “Gold Standard” Set of Linked Records (GSSLR) to

determine the optimum weighting. Such a sub-sample can be determined by clerical review (Dunn, 1946).

2.2.1 Measures of the Quality of Classification

This section will define selected measures of the quality of classification, some of which are used in the record linkage literature and others that are not but are included to illustrate why they are not used. Christen, 2012, is a key reference book that reviews record linkage techniques. It explains measures of the quality of classification. These measures use the number of True Positives (*TP*), of False Positives (*FP*), of True Negatives (*TN*) and of False Negatives (*FN*). Many record linkage applications have a large and unimportant number of *TN*s and so the record linkage literature uses those measures that do not use true negatives (Christen, 2012).

TPR and *FPR* were defined by equation 1.5 and 1.6 respectively.

$$TPR = \frac{TP}{TP + FN} \tag{1.5}$$

$$FPR = \frac{FP}{TN + FP} \tag{1.6}$$

True Negative Rate (*TNR*), (otherwise known as specificity or selectivity), defined by equation 2.1, measures the proportion of actual negatives, that is true negatives and false positives, that are correctly identified as such and it obviously takes account of true negatives.

$$TNR = \frac{TN}{TN + FP} \tag{2.1}$$

Positive Predictive Value (*PPV*), defined by equation 2.2, measures the proportion of classified matches that are correctly identified as such. In the record linkage and classification literature it is referred to as *precision,* however it is not analogous to the closeness of a set of measurements to each other and so to avoid confusion this thesis will refer to it as *PPV*.

$$PPV = \frac{TP}{TP + FP} \tag{2.2}$$

True negatives are important in medical research and in CBM because it is significant if a patient does not have a condition or if a device does not have a fault. In such cases *TPR*, *FPR* and *TNR* are appropriate measures of the quality of classification. On the other hand, true negatives are numerous but unimportant in record linkage, in pattern recognition and in information retrieval and so they should not be considered in these disciplines. *PPV* and *TPR* are appropriate measures of the quality of classification in these disciplines because they do not use true negatives. (Christen, 2012).

2.2.2 Probabilistic Record Linkage

This section will review PRL techniques. Techniques described by Fellegi and Sunter, 1969, for Probabilistic Record Linkage (FSPRL) are commonly used in record linkage to determine what weight to put on each feature (Christen, 2012). They make the simplifying assumption that each feature is independent of the other features, but despite this they achieve useful results in practice. Wilson, 2011, showed that performance can be improved by not making this assumption.

Wilson, 2011, used neural networks to adjust the agreement and disagreement weights to overcome errors caused by Fellegi and Sunters independence assumption. By observing the effect of weights on *TPR* and *PPV* and adjusting weights accordingly, neural networks avoid assigning too much weight to those features that correlate with other features. If some of the inputs are correlated, the neural networks training algorithm will tend to adjust weights to account for this whereas the FSPRL approach will not.

Wilson, 2011, also use what they refer to as ‘full features’, where each feature, such as ‘given name agrees’ or ‘given name disagrees’, is replaced by a range of features such as ‘given name agrees well’, ‘given name agrees weakly’ and so on. They show that neural networks and full features yield dramatic improvements in the quality of record linkage.

2.2.3 Natural Language Processing

This section will present Natural Language Processing (NLP) techniques, (otherwise known as linguistic analysis techniques) used in record linkage (Rubenstein and Goodenough, 1965, Mikolov et al., 2013, Sayers et al., 2015, Comber et al., 2019, Lin et. al., 2019).

The *Levenshtein distance* between two text strings is the minimum number of single-character edits required to change one string into the other. Sayers et al., 2015, use the *Levenshtein distance* to compare text strings for the linkage of medical records. This thesis uses the *Levenshtein distance* and further work could use the more complex techniques described in the remainder of this section.

Rubenstein and Goodenough, 1965, present what linguists refer to as the distributional hypothesis: That the more semantically similar two words are, the more distributionally similar they will be, and thus the more they will tend to occur in similar linguistic contexts. One application of the distributional hypothesis is word vectorization; a methodology in natural language processing that maps words or phrases from vocabulary to a corresponding vector of real numbers. The vectors are often compared by their cosine similarity (for example, Comber et al., 2019). Unlike scalar string comparison techniques such as the *Levenshtein distance*, vector techniques have been shown to find semantic and syntactic relationships between words without training.

In the information retrieval literature (Mikolov et al., 2013), *accuracy*, defined by equation 2.3, measures the proportion of correctly classified instances. It is not analogous to the closeness of a set of measurements to a specific value.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.3}$$

Mikolov et al., 2013, published Word2vec, a word vectorization algorithm intended to maximise *accuracy* and to minimise computational complexity. Low computational complexity made it possible to compute high dimensional word vectors from a large data set[6]. They presented two neural probabilistic language model architectures: The continuous bag of words architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word. These were trained using stochastic gradient descent and backpropagation. They used the Google News corpus for training the word vectors. This is a very large corpus, containing about 6 billion words, but they restricted the vocabulary size to the 1 million most frequent words.

Figure 2-2 shows *accuracy* (defined in equation 2.3) on a subset of their semantic-syntactic word relationship test set using word vectors from the continuous bag of words architecture with limited vocabulary. Semantics is the study of meaning in language and syntax is the set of rules, principles, and processes that govern the structure of sentences. Their models do not have any input information about word morphology: the analysis of the structure of words and of parts of words. This limits the *accuracy* as a question is assumed to be correctly answered only if the closest word to the vector computed is the same as the correct word in the question; synonyms are counted as mistakes. Mikolov et al., 2013, said that their data showed that adding more dimensions or adding more training data provides diminishing improvements. This thesis presents their results on a logarithmic scale in Figure 2-2. It shows that with high dimensionality there are improvements in *accuracy* from using more training data. This did not show up on the linear scale in their paper. This additional observation does not contradict those of the authors.

[6] Figure 2.6 shows the size of the data set used by Mikolov et al., 2013.

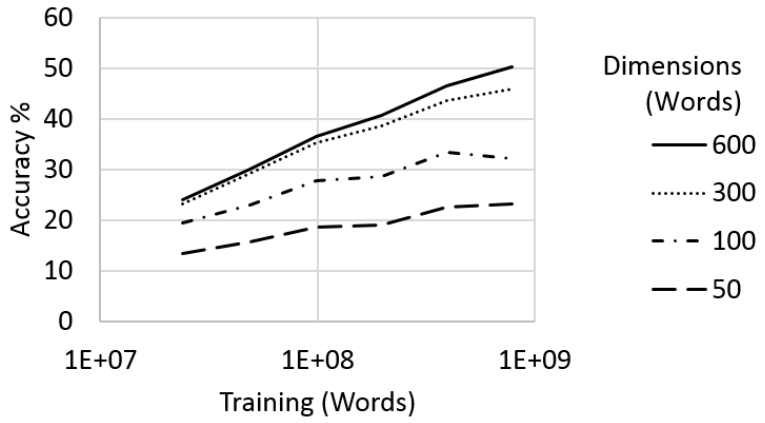


Figure 2-2, Accuracy on a Subset of the Semantic-Syntactic Word Relationship Test Set, using Word Vectors from the Continuous Bag of Words Architecture with Limited Vocabulary. Based on data from Mikolov et al., 2013

Comber et al., 2019, parse address sequences into features to convert their raw data into a structured format for comparison. They train word2vec on 29.6 million parsed postal addresses. Their algorithm maps words sharing the same context closer together in the vector space by feeding successive address features into the model.

The technique used by Comber et al., 2019, for measuring the quality of record linkage does not use a GSSLR. They instead measure the quality of record linkage using a questionable technique: They test their results against an automatically generated ‘ground truth’ dataset consisting of a hundred thousand pairs of ‘correctly linked’ address records that they generate by copying a list of addresses. They simulate incorrectly linked records by introducing error characteristics to the correct records. The match status of these synthetic non-matches is always set to false, meaning that their machine learning techniques learn the representations of non-matched addresses for what they call “highly nuanced cases”. However, typographic errors are not proof of an incorrect POLR. Incorrectly linked records could alternatively have been simulated by randomly linking the records. Comber explained this in a personal note:

> “We found the problem with simulating random links between addresses is that the trained classifier loses discriminative power. i.e. its performance actually diminished in cases where we were comparing two highly similar addresses that were only differentiated between, for example, "2A, West street, SW14 8RF" and "2B, West street, SW14 8RF"”

Lin et. al., 2019, train a word2vec model with dimensionality of 256 words to transform address records into vector representations. They then apply an enhanced sequential inference model; they

achieve local inference between two compared address records using a modified decomposable attention model: a neural architecture and they achieve global inference between two compared address records using a bidirectional long short-term memory: a recurrent neural network architecture. They simulate incorrectly linked records using the same method as Comber et al., 2019.

Neither Comber et al., 2019 nor Lin et. al., 2019, have access to a GSSLR to compare their results against and so the measurements of the quality of classification that they present are compromised.

2.2.4 Conclusion to Existing Record Linkage Techniques

This section reviewed the literature on record linkage. All the publications cited (except for Papatzimos et al., 2017) measure the quality of record linkage, but none of them assesses the confidence that can be placed in those measures. This means that the uncertainty caused by the size of the gold standard is not quantified. Chapter 6 will address this gap with a review of the mathematical literature on interval estimation for a binomial proportion.

2.3 Classification Techniques

This research required a statistical technique that can model the probability that a dependent variable that can take only certain discrete values (otherwise known as a label or as the outcome) belongs to a specific class (otherwise known as a category). Classification algorithms categorise data into a given number of classes, predicting the class of a discrete dependent variable by analysing a dataset (otherwise known as covariates, predictors, or explanatory variables). Most of the publications that this research found on such problems address the automatic detection of spam emails (unsolicited messages sent in bulk).[7]

This section presents a variety of methods for probabilistic classification; classification methods that use statistical inference to find the best class for a given instance. Probabilistic classification techniques model the probability that an observation belongs to a specific class, select the class with the greatest probability and assign the observation to that class.

2.3.1 Linear Regression

Linear regression models the relationship between variables by fitting a linear equation to continuous data. It is the linear form of regression analysis, the statistical processes that estimate the relationships between a dependent variable and one or more independent variables (otherwise known as predictors, covariates, explanatory variables or features). While it is not used for classification, because continuous data do not reflect the probability that a binary or multiclass dependent variable belongs to a specific class, it is a very popular statistical technique. It is used to assess the degree of correlation between two properties of a sample or to compare such a correlation to relationships predicted by theory. It has been used, for example, in medicine, to forecast life expectancy (Aalen, 1989), in astronomy, to study the structure of the universe by determining the distance to celestial objects (Isobe et al., 1990) and for facial recognition, technologies that measure and match people's facial characteristics for their identification or surveillance (Naseem et al., 2010).

---

[7] Section 4.4.4.3.2 will present a technique that uses BNB classification to link WOs to outages by the exploitation of material consumption records. It will present results showing that BNB classification is not an appropriate technique for this application; testing the hypothesis that an the relationship between an OWT failure mode and whether each class of spare part was used for the repair is analogous to the relationship between whether or not an email is spam and what words are in the email and finding that this analogy didn't work. It will then present a simplified technique, developed as part of this research, that has some of the characteristics of BNB classification.

Linear regression algorithms fit a trendline to a data set with 2 dimensions, a plane to a data set with 3 dimensions or a hyperplane to a data set with more dimensions. They use an estimator, a measure of the 'goodness of fit' of a statistical model, to assess how closely the trendline fits the data and they maximise this estimator by optimising the intercept and gradient of a straight line. A popular estimator used by regression algorithms is the ordinary least squares estimator (Gross and Groß, 2003), denoted r^2, which measures the distance, in the direction of the dependent variable, between the trendline and each data point. r^2 is one minus the ratio of the sum of the squares of these distances to the variance of the dependent variable; it results in a value between zero and one and a higher value implies a tighter correlation. A linear regression algorithm fits a trendline to data by maximising the estimator. Linear regression is not a classification technique but is included here to differentiate it from logistic regression.

2.3.2 Logistic Regression

The logistic function is defined by equation 2.4. It results in a value between zero and one. Logistic regression uses the logistic function to model the probability that a dependent variable belongs to a specific class, given continuous data. The dependant variable can either be binary or multiclass.

$$logistic(x) = \frac{1}{1 + exp(-x)} \tag{2.4}$$

The likelihood function is the probability that the logistic function would produce the training data. Maximum likelihood estimation maximises the likelihood function. Logistic regression fits a logistic function to training data by maximum likelihood estimation.

Logistic regression is widely used; for example, in medicine, it has been used to test the significance of the null hypothesis that two alternative treatments are equally liable to induce an unwanted side effect (Cox, 1958, Hosmer et al., 1991) and to identify risk factors for diseases (Tierney et al., 1985, Festa et al., 2005) and it has been used in behavioural ecology to analyse the behaviour of animals (Koster and McElreath, 2017).

This technique can model the probability that a dependent variable belongs to a specific class, as required for the application of predicting the probability that a WO matches a particular outage using material consumption data.

2.3.3 Support Vector Machines

Hyperplanes are decision boundaries used to classify data points. They have *n* dimensions where *n+1* is the number of dimensions of the data. Data points falling on either side of the hyperplane are modelled as belonging to different classes. Support Vector Machines use those data points that fall close to the hyperplane to determine its position and orientation (Cortes and Vapnik, 1995). They are

used for text classification (Joachims, 1999, Rezaeian and Novikova, 2020), image classification (Lin et al., 2011) and handwriting recognition (Bahlmann et al., 2002).

This technique can model the probability that a dependent variable belongs to a specific class, as required for the application of predicting the probability that a WO matches a particular outage using material consumption data.

2.3.4 K-Nearest Neighbours

K-nearest neighbours classification identifies the distances between a query and all the examples in the data, it selects the specified number of examples (K) closest to the query, and then votes for the most frequent label. (Fix and Hodges, 1989, Altman, 1992). Similarly to Support Vector Machines, it is also used for text classification (Dharmadhikari et al., 2011), image classification (Imani, 2021) and handwriting recognition (Lee, 1991).

This technique can model the probability that a dependent variable belongs to a specific class, as required for the application of predicting the probability that a WO matches a particular outage using material consumption data.

2.3.5 Decision Trees

A decision tree is a branching model of decisions. Classification by decision trees fits a model to a dataset using a cost function (Doyle, 1973). Similarly to Support Vector Machines and K-nearest neighbours, they are also used for text classification (Sakakibara et al., 1993), image classification (Xu et al., 2021) and handwriting recognition (Takagi, 2006).

This technique can model the probability that a dependent variable belongs to a specific class, as required for the application of predicting the probability that a WO matches a particular outage using material consumption data.

2.3.6 Bernoulli Naïve Bayes Classification

A multi-variate Bernoulli model is a probabilistic model trained using binary data. Naïve Bayes methods are commonly used for text classification, typically spam filtering, where binary data are generated by checking whether or not a word is used in the email, rather than counting how many times the word is used (Kim et al., 2006, Almeida et al., 2011, Jiang et al., 2012, Zhang et al., 2016, Xu, 2018, Rezaeian and Novikova, 2020). Bernoulli Naïve Bayes Classification has also been used in authorship attribution (Altheneyan and Menai, 2014), sentiment analysis (Abbas et al., 2019), medical diagnosis (Al Aidaroos et al., 2012), veterinary diagnosis (Kuncheva 2006) and software defect prediction (Arar and Ayan, 2017).

The naïve Bayes classifier makes the so called naive Bayes assumption: this is that all of the attributes of the examples are independent of each other given the context of the class; for example, that the probability of each word occurring in an email is independent of the occurrence of other words, given that it is, or is not, a spam email.

The multi-variate Bernoulli model considers both the probability of occurrence for attributes that do occur in the event (such as the probability of occurrence for words that do occur in the email), and, importantly, also the probability of non-occurrence for attributes that do not occur in the event.

Bayes' theorem is stated in equation 2.5. It identifies the posterior probability of A given B $\big(P(A|B)\big)$, the conditional probability of event A occurring given that B is true, using:

- The likelihood of A given a fixed B $\big(P(B|A)\big)$, that is the conditional probability of event B occurring given that A is true
- The prior probability $\big(P(A)\big)$, that is the unconditional probability of observing A
- The marginal probability $\big(P(B)\big)$, that is the unconditional probability of observing B.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{2.5}$$

A Bayesian learning framework can be used for supervised learning with labelled training examples. It uses training data (such as emails labelled either as spam or as not spam) to estimate model parameters (such as the probability of a word being used (or not being used) in an email, given that the email either is or is not spam). It next classifies new events using Bayes' theorem to calculate the posterior probability that a class would have generated the test event in question (such as the probability that a spam email would contain certain words and would not contain certain other words). Finally, it performs classification by selecting the class with the highest probability (such as predicting that an email is or is not spam).

This technique can model the probability that a dependent variable belongs to a specific class, as required for the application of predicting the probability that a WO matches a particular outage using material consumption data.

2.3.7 Conclusion to the Classification Techniques

This section reviewed standard classification techniques. Any of the techniques reviewed in sections 2.3.2 to 2.3.6 can be used for classification. Section 4.4.4.3.2 will present two techniques: a technique using a Bernoulli Naïve Bayes (BNB) classifier and a frequency-based technique. It will show that BNB classification does not support useful interpretation with the unbalanced health history data but that the frequency-based technique does. This research selected BNB classification because it is designed to work with binary input data, and assumed that the use of a particular part would indicate the failure more rather than how many of that part were used. Further work could test the frequency-based technique and the various standard classification techniques against each other in a side by side comparison.

2.4 Techniques for Global Optimisation

Global optimisation is a branch of applied mathematics that attempts to find the global minima or maxima of a function by choosing the system parameters. Most techniques for optimisation can find a minimum but not all can find the global minimum of systems where the variable space contains multiple distinct minima (Olson 2012). The following sections will review two appropriate techniques for global optimisation. Section 7.6 will use the most appropriate of these techniques for the optimisation of the weights and thresholds used in the PEOHH.

2.4.1.1 Brute Force

The “brute force” method computes a function’s value at each point of a multidimensional grid of points. The process for global optimisation by brute force defines this multidimensional grid of points as a list of values for each of the system parameters, for example a minimum, maximum and step. This method has the advantage that it makes no assumptions about the structure of the system that it is investigating and so if the optimum value of each parameter is listed, then the optimum value will be found. The optimal parameter value can be missed if it lies outside of the range or if it lies between the steps. If the Number of Parameters is *NP* and the Number of listed Values of each parameter is *NV* then the number of grid points to evaluate is NV^{NP}. For example, to try 11 values for each of the 28 parameters would require 11^{28} evaluations. This issue, referred to as the ‘curse of dimensionality’, means that the brute force method is unfortunately too computationally expensive for this application.

2.4.1.2 Differential Evolution

An evolutionary method is an optimisation method that uses mechanisms inspired by some features of evolution; mutation, recombination and selection (Schwefel, 1995). This section will present Differential Evolution (DE); an evolutionary method that optimises a function by maintaining a population of candidate solutions and creating new candidate solutions by combining existing ones (Storn and Price, 1997). It is a popular technique, used for example in electronic circuit design (Storn and Price, 1997), in web services (Rodriguez-Mier et al., 2010), in building energy management (Rodriguez-Mier et al., 2019), in wind turbine power curve modelling (Lydia et al., 2013), and in interplanetary trajectory design (Labroquere et al., 2014).

The steps in a DE algorithm are the initialisation of the population, mutation, recombination, replacement and evaluation. Unlike most techniques for global optimisation, it does not use derivatives and so it is appropriate for discontinuous step functions; for example, in the application of this thesis, the optimisation of the Positive Predictive Value (*PPV*).

DE methods have control variables that adjust the population size, the mutation rate, the variation in the mutation rate (or ‘dithering’) and the recombination rate. This section will present a basic DE

strategy. Section 6.2.4 will use this DE strategy to optimise a statistical technique and section 7.6 will use it to optimise the thresholds and the agreement and disagreement weights of the PEOHH.

The literature on DE (Storn and Price, 1997, Abbasa et al., 2017) describes an array of parameters; for example, in the application of this thesis, the thresholds and the agreement and disagreement weights of the PEOHH expressed as a vector. The number of dimensions of the vector is the Number of Parameters to be varied (*NP*).

DE generates a set of vectors referred to as population vectors. In each generation, DE operates on each vector, referred to as the current vector or target vector (x_g^i). It identifies each vector of each generation by a running index (*i*), (otherwise known as the vector index), it identifies each parameter by a parameter index (*j*) and it identifies each generation of the evolutionary process by a generation number (*g*). The Population Size (*PSi*) is the number of population vectors produced in each generation. DEs mutation operation creates *PSi* mutant vectors (v_g^i), (otherwise known as donor vectors). This section will present the most popular DE mutation scheme, called '*best/1*' because it selects the 'best' vector and because it uses one mutation term. Storn and Price, 1997 and Abbasa et al., 2017 define various DE strategies but for brevity this thesis will only present one of them.

Equation 2.6 defines the *'best/1'* strategy using the Mutation Probability (*MP*) (Storn and Price, 1997, Abbasa et al., 2017). The best vector in a population is the vector that yields the best fitness, that is the lowest cost function and in the case of the PEOHH, it is the parameter values that yield the highest $\hat{p}$. The *best/1* strategy selects the best vector (x_g^{best}) from the set of population vectors and it randomly selects two of the population vectors by their index (r_2, r_3). $r_2, r_3 \in \{1,2,3 \dots PS\}$ and $r_2, r_3 \neq i$. It adds the weighted difference between the two random population vectors to the best population vector using equation 7.9.

$$v_g^i = x_g^{best} + MP\left(x_g^{r_2} - x_g^{r_3}\right) \tag{2.6}$$

Dithering is the randomisation of the Mutation Probability (*MP*). It can increase the speed of convergence (Price et al., 2005, Dawar and Ludwig, 2014). Dithering is defined by equation 2.7 using the mean Mutation rate (*Mu*), the Dithering rate (*Di*) and a uniformly distributed real number in the range $(-0.5, 0.5)$ generated anew for every generation ($randg(-0.5, 0.5)$).

$$MP = Mu + Di \times randg(-0.5,0.5) \tag{2.7}$$

This section will present the most popular DE crossover or recombination scheme; called *binomial*. A trial vector (u_g^i) is a vector that is created in the crossover operation in which DE mixes the mutant vector (v_g^i) with the current vector (x_g^i). Equation 2.8 presents the binomial crossover scheme (*'bin'*)

using the Crossover Rate (*CR*), (otherwise known as the crossover probability or recombination constant), the vector index (*i*), the parameter index (*j*), the generation number (*g*), a uniformly distributed real number in the range $(0, 1)$ generated anew for every parameter of every vector ($rand_j^i[0,1]$) and a random integer between 1 and *NP* (j_{rand}) generated anew for every parameter of every vector. (Storn and Price, 1997, Dawar and Ludwig, 2014, Abbasa et al., 2017). A larger *CR* increases the mixing of the parameters to perturb the population more and can help the search to get out of multi-dimensional 'dips' to find whether there is a lower minimum elsewhere.

$$u_{j,g}^i = \begin{cases} v_{j,g}^i, & \left(rand_j^i[0,1] \leq CR \; or \; j = j_{rand}\right) \\ x_{j,g}^i, & else \end{cases} \tag{2.8}$$

In each generation and for each vector of the population, DE selects the fittest option between the trial vector (u_g^i) and the current vector (x_g^i). Equation 2.9 presents DEs selection operation where the fitness function calculates the value of the objective function, for example, in the application of this thesis, it calculates the *PPV* from a vector of thresholds and the agreement and disagreement weights of the PEOHH (u_g^i or x_g^i).

$$x_g^{i+1} = \begin{cases} u_g^i, & fitness\left(u_g^i\right) < fitness\left(x_g^i\right) \\ x_g^i, & else \end{cases} \tag{2.9}$$

Convergence criteria are the criteria at which the solving stops because DE has converged sufficiently. DE uses the convergence criteria defined in equation 2.10. The Population Energies (*PE*) are a table of the function evaluation for each population vector (Sandell, 2020). The Absolute Tolerance (*AT*) is set to the default value of 0. The Relative Tolerance (*RT*) is set to the default value of 0.01. These are the default settings for the convergence criteria of DE.

$$\sigma(PE) \leq AT + RT.|\overline{PE}| \tag{2.10}$$

DE solving stops either when the convergence criteria have been met or when the maximum number of iterations, set at the default value of 1000, is reached.

2.4.1.3 Conclusion to the Techniques for Global Optimisation

This section reviewed the "brute force" method of global optimisation and DE and it showed that to yield a computable result in the multi-dimensional application of this research it will be necessary to use DE. Chapter 7 will present its application.

2.5 Conclusion to the Background

This background chapter has presented a description of the existing records of OWT health history that this project has benefitted from access to, the state of the art for record linkage (used in chapters 3 and 4), a classification technique (used in section 4.4.4.3.2) and a global optimisation technique (used in section 7.6).

This chapter has identified a gap in the literature that techniques joining wind turbine WOs to alarms and to outages are only described in one publication. This publication, Papatzimos et al., 2017, links offshore wind turbine WOs to records of control system alarms. It uses a single feature; a timestamp, but this literature review has shown that multi-feature record linkage techniques outperform single-feature record linkage techniques.

This thesis will address that gap in the literature by presenting new techniques that link together existing records of WOs and outages to determine and validate an Enriched Health History (EHH). Section 1.2 predicted that this will enable improvements in the maintenance of OWTs in the areas of maintenance scheduling, troubleshooting, CBM and in the measurement of the effectiveness of maintenance activities.

The literature on record linkage does include measures of the quality of record linkage but Papatzimos et al., 2017, do not measure the quality of their record linkage. The literature does recognise that a small GSSLR can only yield an uncertain estimate of the quality of record linkage but it does not quantify this uncertainty. The literature also measures the quality of record linkage but it does not assess the confidence that can be placed in those measures. When applied to maintenance, understanding this uncertainty informs maintenance decision making; which, as section 1.2 showed, can improve productivity. Chapter 6 will review statistical techniques that will be used in chapters 7 and 8 to assess that confidence.

3 Evaluating Health History Enrichment

This chapter discusses the aims behind the methodology used in the Process for the Enrichment of OWT Health History (PEOHH) developed in this research and the assessment of the success of the PEOHH. It discusses which of the measures of the quality of classification presented in section 2.2.1 are appropriate for this research and it develops a new measure based on these.

Chapter 2 identified a gap in the literature that multi feature record linkage techniques have not been used to improve maintenance strategies. This thesis presents new record linkage techniques and applies these as well as existing record linkage techniques in the field of OWT maintenance, as described in the following Research Question (RQ):

RQ1 To enable improvements in WT CBM & troubleshooting, how can WT health history be enriched?

Chapter 4 will present a process for linking together existing records of OWT health history to identify an Enriched Health History (EHH). These records, all described in section 2.1, are the:

- database of alarms
- database of outages
- database of Work Orders (WO)
- database of material consumption

Ørsted generated the database of outages from the database of alarms and so linking these together is trivial since the records timestamps can be used as an index. The database of material usage is labelled with WO numbers and so linking these together is also trivial, since the order number can be used as an index. For reasons that were discussed in chapter 1, it is important to link WOs to records of outages and literature searches up to 2021 did not find such techniques described. In this thesis, a Pair of Linked Records (POLR) is generated when a single WO is linked to a single outage record. Record linkage is not a new area of research but applying it to maintenance records is.

Each POLR is either correct (the WO does refer to the same event as the outage) or it is incorrect (the WO refers to a different event to the outage).

The probability that any given POLR in the EHH is correct (*P*) is the probability that the WO does refer to the same event as the outage. If *P* can be predicted, then the EHH could be filtered by *P*. This would identify a filtered EHH in which each record was of at least a minimum value of *P*. If the quality of record linkage is not always perfect, then filtering would be beneficial for the applications detailed in section 1.2. For example, in the application of troubleshooting, a developer might filter the EHH to only include higher scoring POLRs to increase the quality of the EHH. With a high quality EHH its quantity is lower. The optimum depends on the application, where some applications benefit from a higher quality EHH and other applications benefit from a larger quantity of EHH data. For example, it is important to operators that maintenance technicians have confidence in the analytic tools that are made available to them. An analytic tool that recommended that technicians take parts that they would be unlikely to need might not win the confidence of technicians. In that case, a higher quality EHH would help to win the confidence of the technicians. On the other hand, if the tool were well understood by the technicians then they might choose a larger quantity EHH, so as to bring parts that, while unlikely, might turn out to be required for repairing their OWTs. As new wind farms are constructed at increasing distance from shore, the optimum set of spare parts will tend to get larger.

3.1 Validation

If the EHH were to contain incorrect information then this would be detrimental to any application of it. Verification, Validation and Testing (VVT) are techniques that together assess whether a product, service, or system meets requirements and specifications and fulfils its intended purpose. This section will introduce the processes that later chapters use for the VVT of the PEOHH. This thesis uses VVT terminology from Balci, 1994, Robinson, 1999, Sargent, 2013 and Hicks et al., 2015. VVT will measure the quality of the EHH and this measurement will enable the maximisation of that quality and is addressed in the following RQ:

RQ2 How can the quality of the EHH be validated?

Figure 3-1 and Figure 3-2 illustrate the techniques that this thesis will use for the VVT of the PEOHH. This thesis breaks this process down into four sub processes:

- Data validation: validation of the existing records of health history against the EHH, best carried out after the validation of the EHH. It results in measures of the quality of the existing records of health history down to a level of detail that can for example predict the confidence that a WO start date is accurate.
- Conceptual model validation: validation of the conceptual record linkage models against the existing records of health history, testing the assumptions behind each model.
- Computational model verification: verification of the computational models against conceptual models. This involved the checking of each line of code to identify errors and to check logic as well as running a manual test on each algorithm, checking that it works correctly and that it does not contain coding errors.
- Operational validation: Operational validation is the comparison of the model (or in this case the EHH) to reality. Reality is represented in this case by a gold standard sub-sample. Section 2.2 presented existing record linkage techniques. It was described in that section that "PRL techniques require a sample "Gold Standard" Set of Linked Records (GSSLR) to determine the optimum weighting. Such a sub-sample can be determined by clerical review (Dunn, 1946)." A clerical review by a wind turbine expert at Ørsted and this author[8] identified the GSSLR.

[8] The clerical review will be described in section 4.3, as will a semi random method that will be used to identify a set of WOs that is representative of the more important WOs, as only these will be included in the GSSLR.

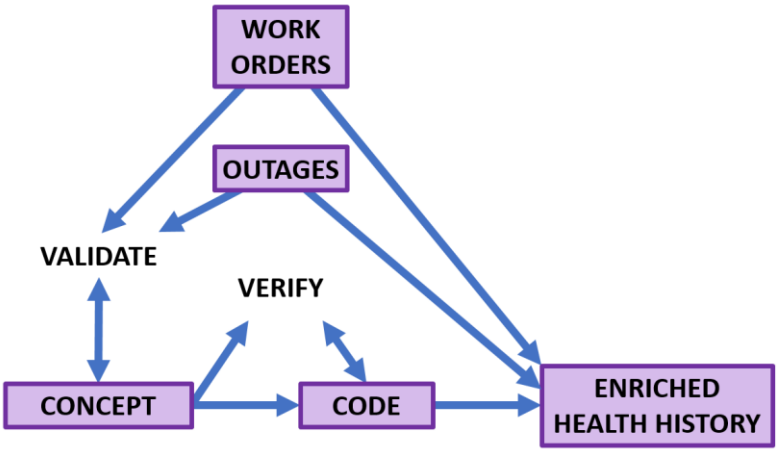


Figure 3-1, Conceptual Model Validation and Computational Model Verification

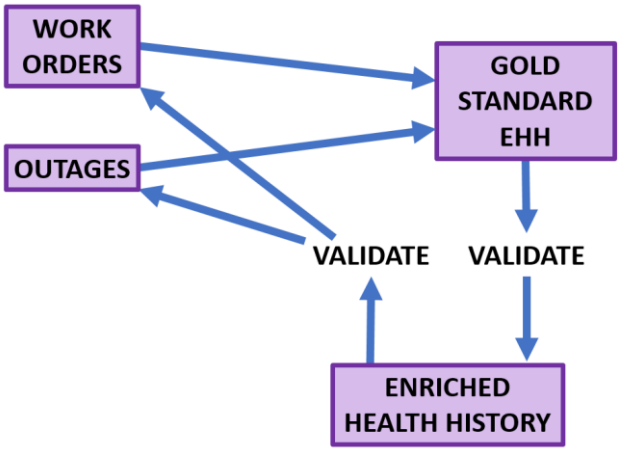


Figure 3-2, Data Validation and Operational Validation

All four of these forms of validation are essential but operational validation is the most important.[9]

The PEOHH uses an ensemble of record linkage techniques to compare features of the WOs to features of the outages. Section 2.2 showed that, in the fields of medicine, census data, genealogy and in the detection of duplicate internet search results, using an ensemble of record linkage

[9] The operational validation of the EHH will be discussed in section 4.3 and its results presented in chapters 7 and 8.

techniques gives a better quality result than using a single record linkage technique. This research will test the hypothesis that this also applies in the field of offshore wind turbine maintenance and chapters 7 and 8 will present the result that it does.

Section 4.3 will present the techniques that were used by this research for the validation of the EHH.

3.2 Measures of the Quality of Record Linkage

Each of the record linkage techniques that will be presented in chapter 4 is based on a set of assumptions; this thesis, however, does not rely on these assumptions but instead it tests them as hypotheses. These hypotheses are tested by the validation of a sample from the Enriched Health History (EHH) against the Gold Standard Set of Linked Records (GSSLR). Consider a set of WOs, each one linked correctly to its corresponding outage. This will be referred to as the 'true EHH'. All of these hypothesis tests rely on the assumption that the GSSLR is representative of the true EHH. The GSSLR was randomly selected from the important WOs[10] so it will be reasonably representative but a larger GSSLR would tend to be more representative[11]. Later chapters will present measures of the uncertainty of how representative the GSSLR is and will compare techniques that use this understanding to identify the confidence of the correctness of the EHH.

3.2.1 True Positive Rate and True Negative Rate

Measures of the quality of classification were discussed in in section 2.2.1 but as a recap, True Positive Rate (*TPR*), defined by equation 1.5, measures the proportion of true matches that have been classified correctly.

$$TPR = \frac{TP}{TP + FN} \tag{1.5}$$

Similarly, True Negative Rate (*TNR*), defined by equation 2.1, measures the proportion of actual negatives that are correctly identified as such.

$$TNR = \frac{TN}{TN + FP} \tag{2.1}$$

In other record linkage applications, each record might not link to any other records or it might link to multiple other records. The PEOHH assumes that each WO links to one outage. This means that unlike other record linkage applications, the PEOHH does not generate negative links which means that *TPR* and *TNR* are unknown. That the PEOHH does not generate negative links does not mean that it does not generate false links. This is neither an assumption nor an advantage of the PEOHH; it is one of its features.

---

[10] The selection of the GSSLR will be described in section 4.3.

[11] Section 4.3.2 will present the method used in this thesis for the validation of the techniques for health history enrichment and will explain that the size of GSSLR, 29 WOs, was constrained by the amount of expert time that was available.

3.2.2 Positive Predictive Value

As discussed in section 2.2.1, the Positive Predictive Value (*PPV*), defined by equation 2.2, measures the proportion of classified matches that are correctly identified as such. It indicates the quality of record linkage.

$$PPV = \frac{TP}{TP + FP} \tag{2.2}$$

This thesis will compare a sample from the EHH to the GSSLR. It will calculate the *PPV*, referred to as the *PPV* of the GSSLR or just as the *PPV*. A point estimate is a single value estimate of an unknown population parameter. The *PPV* of the GSSLR is a point estimate of the *PPV* of the EHH.

The binomial distribution describes the behaviour of a count variable for a fixed number of observations where each observation is independent, where each observation has one of two possible outcomes and where the probability of each outcome is the same for each observation. The binomial proportion is the number of successes divided by the number of trials. The *PPV* of the GSSLR will be modelled as a binomial proportion and chapter 6 will present methods for the estimation of the uncertainty of a binomial proportion. To identify the confidence of the correctness of the EHH, section 7.6.2 will construct confidence intervals (defined in chapter 6) for the *PPV* of the EHH.

3.2.3 Negational Positive Predictive Value

Section 2.2 presented the existing Probabilistic Record Linkage (PRL) techniques:

> PRL techniques combine an ensemble of comparison features by weighting each dimension of the comparison vector to give an overall score indicative of the probability that each POLR is a true match.

Where the weight assigned to a dimension of the comparison vector is set to zero, the corresponding feature is disregarded for record linkage comparison. This thesis will refer to setting the weight to zero as '*disregard'.* Consider disregarding one of the features: the effect of not using a feature is a useful measure of the effectiveness of that feature. This thesis presents the Negational Positive Predictive Value (*NPPV*); the Positive Predictive Value (*PPV*) calculated using all the features except one feature or set of features that is disregarded. This measure of the quality of classification is useful because if a feature has a low *NPPV* then the PEOHH can avoid unnecessary computation by excluding it from the ensemble of features in the PEOHH.

3.3 Measures of the Richness of Health History

The claim that the health history has been enriched must be tested. This thesis is interested in whether the health history has been made more useful for the applications detailed in section 1.2. This thesis defines the richness of a health history data set as how much useful information it contains on each fault including the time that it occurred, the failure mode and the severity. The severity of a fault is determined by how much production was lost and the cost of repair. This section introduces the following RQ:

RQ3 How can the richness of historical data on wind turbine health be measured?

By joining work orders to outages, the PEOHH develops a new, integrated database, the EHH, that contains both information from the WOs such as material consumption and work hours and information from the database of outages such as lost production. It enriches the health history by joining these records together, enabling applications of the health history that improve maintenance such as the applications detailed in section 1.2.

For the applications detailed in section 1.2, the EHH that is required only describes corrective work, excluding preventive, retrofit, inspections and surveys and condition-based work. The PEOHH filters the WO data by type to include only corrective work.

Chapter 4 will frame RQ1 and chapter 7 will answer it. Sections 2.2 and 3.1 addressed RQ2. This section has presented RQ3 which will be addressed in Chapter 8, which will also discuss what is meant by enrichment and will quantify the uncertainty of the measures of enrichment.

3.4 Conclusion to Evaluating Health History Enrichment

This chapter has presented the theory that underlies the validation of the PEOHH. It has presented measures of the quality of record linkage that will be used in the following chapters to optimise and to validate the PEOHH. It has identified the research questions that will be addressed by this thesis. Chapter 8 will present measures of the richness of health history and will use them to validate the claim that the health history has been enriched.

4 Techniques for Health History Enrichment

Chapter 3 presented the techniques that will be used to evaluate the Process for the Enrichment of OWT Health History (PEOHH) developed in this research and identified the following research question (RQ):

RQ1 To enable improvements in WT CBM and troubleshooting, how can WT health history be enriched?

This chapter introduces the PEOHH. This process joins together four existing records; a database of alarms, a database of outages, a database of Work Orders (WO) and a database of material consumption[12]. Section 4.1 will present the PEOHH but will not detail the techniques that the PEOHH uses. Section 4.2 will discuss how this research developed the process. The techniques used for the validation of the PEOHH will be presented in section 4.3. The techniques that the PEOHH uses, structured by the type of feature that the technique uses, will be presented in section 4.4. This research selected a single wind farm for the development of the techniques that this thesis presents. The identity of the farm is commercially privileged information and to protect its anonymity the reasons for selecting it are not discussed in this thesis, but it is a relatively mature farm and therefore has a useful amount of maintenance history available, which would not be the case with a more recent wind farm. Table 4-1 presents the size of these existing records of OWT health history. The number of material consumption line items in Table 4-1 includes parts used for all types of maintenance. This should not be confused with the number of parts used as each material consumption line item details a quantity of material, for example a number of parts or a volume of paint or oil.

[12] The existing records of wind turbine health history were described in section 2.1.

Table 4-1 gives the number of corrective[13] WOs rather than the total number of all types of WO because, for reasons that will be presented in Section 4.4.2, the PEOHH filters the WO data by order type to include only corrective work orders. The population size of 9.82×10^3 WOs and a larger number of alarms, outages and Material Consumption Line Items (MLI) is smaller than that required by some Natural Language Processing (NLP) techniques[14] but chapter 8 will show that this dataset is large enough for this research to derive statistically robust results.

	Count
Alarms	6.70×10^6
Outages	5.32×10^4
Corrective WOs	9.82×10^3
MLIs	1.35×10^4

Table 4-1, Size of the Existing Records of OWT Health History

[13] Section 2.1.3.3 presented the WO 'type' feature which classifies each WO as either preventive, retrofit, inspections and surveys, condition-based or corrective.

[14] NLP techniques were described in section 2.2.3.

4.1 Process for the Enrichment of OWT Health History

Section 2.2 presented existing record linkage techniques and introduced blocking:

> "Probabilistic Record Linkage (PRL) techniques join two databases together to create a new database in which each row represents one Pair Of Linked Records (POLR). PRL techniques could compare each record in one database with each of the records in the other database but this would be computationally expensive and can lead to inaccuracy (Sadinle and Fienberg, 2013) so they instead split the data into smaller blocks to disregard very unlikely pairs. They compare features in the data sets being linked together to generate a comparison vector containing comparison features"

Any record linkage approach that did not include blocking would be computationally intractable.

The first step of the PEOHH is to join WOs to outages to create POLRs. It uses the WO start date for blocking after this research elicited the opinion from wind turbine experts that the technique that uses this feature would be the most effective. The PEOHH could have only used WOs that have their start date on the same day as or during the outage. This would mean that, if there was an error in the record of the start date, no POLR would be created and so the other features would not be compared. The PEOHH instead adds a 40-day margin to the outage start and finish times to create an extended duration. It joins each WO to each of the outages of which the WOs start date lies within the outages extended duration.[15] the PEOHH creates POLRs using a Python algorithm posted by Stack Overflow internet forum user "Josh Friedlander" (Friedlander, 2019).

The blocking process creates a Set of Pairs Of Linked Records (SPOLR). Some of the POLRs in the SPOLR link a WO to the outage that it refers to (true links) but others link a WO to a different outage (false links). The aim of this research is that the PEOHH should identify the true links in the SPOLR.

In record linkage, an agreement pattern is a matrix in which each row represents one POLR, each column represents one feature and each value registers agreement, disagreement or neither agreement nor disagreement for the corresponding feature and POLR. The PEOHH identifies the agreement pattern using techniques that will be detailed in section 4.3.[16]

The PEOHH calculates a Weight for each Feature for each POLR ($W_{Fe_{POLR}}$) using the Agreement Weight for that Feature (AW_{Fe}), the Disagreement Weight for that Feature (DW_{Fe}) and equation 4.1.

---

[15] This research investigated the effect of varying the 40-day blocking threshold and the results are presented in section 7.1.

[16] Chapter 5 will illustrate the agreement pattern by linking two example WOs to outages.

The following pages will present the process used in this research for estimating optimal values of AW_{Fe} and DW_{Fe} for each feature. For any feature for which AW_{Fe} and DW_{Fe} for a feature can be set to zero without reducing, this research cannot conclude that there is a benefit of calculating that feature and so operators might consider disregarding that feature so as to avoid computation with no proven drop in the quality of record linkage.[17]

$$W_{Fe_{POLR}} = \begin{cases} AW_{Fe}, & Feature\ registers\ agreement_{\ POLR} \\ DW_{Fe}, & Feature\ registers\ disagreement_{\ POLR} \\ 0, & Feature\ registers\ neither_{POLR} \end{cases} \quad (4.1)$$

The PEOHH calculates a Score for each POLR (S_{POLR}), the sum across all the features of the $W_{Fe_{POLR}}$ using equation 4.2.

$$S_{POLR} = \sum_{Fe} W_{Fe_{POLR}} \quad (4.2)$$

Each POLR in the SPOLR is assigned a unique identifier, 'match ID', in order of when the outage started. The PEOHH sorts the POLRs by match ID, by Δt_{St}[18] and then by S_{POLR} and then it filters the POLRs to retain only the first POLR for each WO. This process identifies a version of the EHH with the highest S_{POLR} for each WO. Where the highest S_{POLR} scores for a WO are equal, it selects the POLR with the smallest Δt_{St}; that is the POLR in which the WO start time is recorded closest to or during the outage. Where the smallest Δt_{St} scores for a WO are equal, for example where they are both zero (that is that the WOs is recorded as occurring during both of the outages) it selects the POLR with the smallest match ID; that is the POLR with the outage that started first. Selecting the POLR with the smallest match ID is an arbitrary decision taken to keep the results consistent. Two power outages cannot happen at the same time on the same OWT and so a WO cannot in reality start during more than one outage; however selecting the POLR with the smallest match ID makes the EHH consistent even when the PEOHH is working with unrealistic dummy data.

[17] Section 7.6.4 will consider the pros and cons of using features where their benefit is uncertain

[18] Section 4.4.1 will define Δt_{St}.

The PEOHH uses a set of thresholds[19] and a set of weightings.[20] Each of these will be optimised to maximise the Positive Predictive Value (*PPV*)[21, 22].

For any feature for which AW_{Fe} and DW_{Fe} can be set to zero without reducing *PPV*, this research cannot conclude that there is a benefit of calculating that feature and so operators might consider disregarding that feature so as to avoid computation with no proven drop in the quality of record linkage.[23]

This section introduced the PEOHH. It described how the PEOHH was implemented and how it was validated. Section 4.4 will present the hypotheses that different features of the health history data can be used for record linkage comparison. Chapter 7 will present the use of the GSSLR to optimize the weights and thresholds used by the PEOHH and chapter 8 will use it to validate the EHH. If the optimized agreement and disagreement weights for a feature are zero, then the hypothesis that this feature makes a useful contribution to the process of record linkage will have been disproved. If whether they are zero or not is uncertain then the hypothesis will have been neither proved nor disproved. If they are not zero, then the hypothesis will have been proved. With a larger GSSLR, the confidence of such a conclusion would be higher.

---

[19] For example, the Threshold for the Time difference between Outages and Alarms (*TTOA*) that will be defined in section 4.4.4.

[20] For example, the Agreement Weight for the Start time feature (AW_{St}). Agreement weight was defined in this section and the start time feature will be defined in section 4.4.1.

[21] *PPV* was described in section 3.2.2.

[22] This optimisation process will be described in chapter 7, after all the features have been presented. A larger Gold Standard Set of Linked Records (GSSLR) would improve the quality of this optimisation.

[23] Section 7.6.4 will consider the pros and cons of using features where their benefit is uncertain

4.2 Development of the Process

This section will describe how this research decided which record linkage features to investigate.

This researcher was trained within Ørsted's Advanced Analytics Lab in how they manually link work orders to outages. This training inspired the techniques using timestamps, particularly the Basic start Date. It also inspired the technique that uses the WO 'description' feature.

The WO 'description' feature is indicative of the failure mode and this inspired the investigation of other features indicative of the failure more.

The feature that uses material usage data is currently implemented as an algorithm that takes some time to run but the optimisation of these algorithms was outside of the scope of this thesis. Experiments to optimise the PEOHH required these algorithms to be run repeatedly and these made up time-consuming tests.[24]

Further discussions with data scientists at Ørsted inspired the feature using the order type and the feature visits to the wind turbine.

It was an important breakthrough in this research when it identified that linking records together, such as linking WOs to outages, is referred to in scientific literature as record linkage. This realisation occurred during the second year of research. This research had already developed some of the features of the PEOHH but this literature showed how to combine multiple features.[25]

A colleague at Durham University suggested that maintenance vessel tracking logs could be used to validate the outage timestamps. These logs show when the vessel visited each turbine and are independent of the maintenance records. This research did not investigate this feature because it judged that validating the outage timestamps would probably not contribute towards linking WOs to outages. Outage timestamps are more accurate than WO timestamps and it was expected that the vessel visits would coincide with the outage timestamps but not with the WO timestamps.

[24] Section 4.5 details and discusses the computation times.

[25] Section 2.2 reviews the literature on record linkage.

4.3 Validation

4.3.1 Aim

Chapter 3 identified the following research question (RQ) and described various validation processes that will be used in this thesis.

RQ2 How can the quality of the EHH be validated?

This section will focus on the operational validation of the Enriched Health History (EHH) and will present the Process for the Validation of the EHH (PVEHH). This includes the process used for selecting a sample from the database of WOs and for matching each WOs in this sample to its corresponding outage to identify a sample "Gold Standard" Set of Linked Records (GSSLR) and the process for comparing the GSSLR to the EHH.

4.3.2 Method

This research selected WOs for the sample using a process that was semi random and semi structured.

This thesis considers significant faults to be those that represent a significant cost to the operator due to any combination of lost production, worker time or material cost. It is these faults that analysts applying the EHH are interested in identifying in the historical record. The record also includes insignificant faults and the repair of these insignificant faults can also be associated with a WO recorded as 'corrective'. The PEOHH filters the WO records to only include those WOs that are recorded as 'corrective' but to identify the sample the PVEHH filters the corrective WOs again to include only those that are significant.

It is an essential requirement that the validation processes are as independent as possible from the process that they validate. It is therefore necessary that the PVEHH is independent of the PEOHH. That means that the process for selecting which records are significant could use either the WO records or the outage records. The PVEHH should not use the EHH which is those records joined together.

Lost production information is contained in the database of outages whereas worker time, material cost and order type information is contained in the database of WOs so to ensure that the PVEHH is independent of the PEOHH, the PVEHH could either use lost production information or worker time, material cost and order type information to decide which records are significant. This research elicited the opinion of Ørsted's wind energy experts about which features would be required to identify important health history events. They all agreed that worker time, the number of person hours recorded against the WO, is the most appropriate measure of the importance of faults. This is because

faults with significant lost production tend to require significant work to resolve and because work time tends to be accurately recorded as it is closely tied to the worker's remuneration.

This research identified more important WOs by including only those WOs with worker time booked against them. This research could alternatively have selected WOs at random from the full population of WOs but this would have resulted in a sample that included unimportant WOs. To identify the GSSLR, this research randomly selected a set of WOs from the set of important corrective WOs. Random selection was used with the intention of getting a sample representative of the population.

The researcher and a wind turbine expert with experience of the wind farm in question attended the validation meeting. The meeting was designed to elicit the correct outage for each WO in the sample, to identify the GSSLR. The researcher read out the WO features that are referred to as wind turbine ID, start date and description. At this point, if the wind turbine expert identified a single candidate outage to match to the WO then the researcher recorded the outage ID. Alternatively, if there were multiple candidate outages then the researcher read out the material consumption and long text for the WO. The researcher recorded the outages identified by the wind turbine expert in a table; the GSSLR. The GSSLR represents expert experience and this thesis will assume that it is a set of true matches although of course it could contain errors.

As discussed in section 2.2.1, the Positive Predictive Value (*PPV*), defined by equation 2.2, measures the proportion of classified matches that are correctly identified as such and indicates the quality of record linkage.

$$PPV = \frac{TP}{TP + FP} \tag{2.2}$$

This research will identify the *PPV* of the GSSLR and will use this sample to estimate the *PPV* of the EHH.

The first operational validation meeting checked 9 WOs in 2 hours. The second meeting checked 5 WOs in 1.5 hours. These were compared to the automatically generated results. Feedback from this validation exercise informed improvements to the PEOHH.

To derive a more accurate estimate of *p*, this research needed a larger GSSLR, however the wind turbine experts' time was constrained and so it needed a faster method of extracting the health history data used for validation. The third operational validation meeting checked 29 WOs in 2 hours. This meeting used data from a different farm from which, for reasons that are commercially privileged, data was quicker to extract.

Differences between farms and any consequent differences in the practice of record linkage are outside of the scope of this research. This thesis only uses the results from the third validation meeting

because it relates to a different farm from the results from the other meetings. The PEOHH only works with one farm at a time because the differences between different farms would be difficult to account for but further work might overcome this constraint. Consequently, this research did not combine the results of the validation meetings for the two farms. The PEOHH was tested using data from the same farm from which the GSSLR was derived.

To summarise, the size of the GSSLR, 29 WOs, was constrained by the amount of expert time that was available.

4.3.3 Result

For one of the WOs in the validation set, the wind turbine expert could not tell which outage matched it. The WO description field reads “Lightning card bracket missing in blade” and its long text reads “The Lightning card bracket is missing from blade A The Lightning card holder is missing from blade B”. This fault does not require immediate remedy and its repair would have been bundled with other jobs. It would not cause the turbine to stop working and so there is no outage associated with this fault. This research kept this WO in the GSSLR because it is a feature of the EHH that it contains WOs like this one that are not associated with an identifiable power outage. This limits the maximum value of the *PPV* of the GSSLR to (n-1) / n = (29-1) / 29 = 96.6%.

4.4 Features

This section presents the hypothesis that an ensemble of features from the WO data, from the outage data, from the alarm data and from the material consumption data can be used for record linkage comparison. It presents each record linkage feature as a technique for Health History Enrichment (HHE).

The techniques that will be presented in this section were developed by a process that started with conversations with Ørsted's wind turbine experts and with researchers from Durham University's Department of Engineering and Department of Mathematical Sciences. The starting point was observation of how Ørsted's wind turbine experts manually link WOs to outages. This research automated each of these techniques and further developed them by experiment. Chapter 7 will discuss which of these techniques should be applied in practice and how they can be optimised.

For each technique, this research developed code in the Python programming language. For some of these techniques, this research benefitted from support from volunteer contributors to the online forum Stack Overflow (StackExchange, accessed 2021).[26]

This section presents twelve record linkage techniques organised into four sub sections:

- Four timestamp-based techniques
- A technique that considers the recorded type of maintenance
- A technique that uses the records of visits to the wind turbine
- Six failure mode-based techniques

For each record linkage technique, the chapter only presents its method and the hypothesis that it might be effective as part of an ensemble of techniques.

Section 4.4.1 will present the hypothesis that that timestamps from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features. Section 4.4.2 will present the hypothesis that the 'type' feature from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features. Section 2.1.3 will present the hypothesis that the 'number visits' and 'duration' features from the outage data can be used for record linkage comparison as part of an ensemble of features. Section 4.4.4 will present two hypotheses: (1) that features indicative of the failure mode from the WO data, from the outage data, from the alarm data and from the material consumption data can be used as part of an ensemble of features for record linkage comparison and (2) that a selection of these techniques could be used to the same

[26] These contributions are individually referenced in this thesis.

effect. All these hypotheses will be tested in chapter 7. Techniques that will be reviewed in chapter 6 will identify uncertainty that will render the results of these hypothesis tests indeterminate. Chapter 9 will recommend an innovation in maintenance record keeping that would drastically reduce this uncertainty.

4.4.1 HHE Techniques Using Timestamps

This section will test the hypothesis that that timestamps from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features. It will present a novel technique for this purpose.

Section 2.1.2 described the outage start time and the outage finish time. Section 2.1.3.1 described the WO start date, the WO finish date and the WO notification date. Section 2.1.4 described the material 'posting date'. While the timestamps in the outage records are generated automatically from the alarm log[27], the timestamps in the WO records are human generated[28].

Papatzimos et al., 2017, linked OWT WOs to alarms using a single feature; a timestamp. To achieve a better quality of record linkage, this research builds on the state of the art with three innovations:

(1) Rather than using a single timestamp, it uses up to four timestamp features together as part of an ensemble of features.
(2) Rather than using only timestamps, it uses other features as well as part of an ensemble of features.
(3) Rather than considering the outage as a single point in time, it instead considers the full outage duration, such that a WO timestamp occurring during a long outage is recognised as agreeing even if it is not close in time to the outage start time.

This section presents techniques using four timestamps:

- WO start date
- WO finish date
- WO notification date
- Part posting date

[27] The database of outages was described in section 2.1.2.

[28] The database of WOs was described in section 2.1.3.

This chapter will compare the effectiveness of four techniques that each use one of these four features. The PEOHH compares the POLRs that it generated using the blocking process[29].

This section presents two techniques that both use any one of the four timestamps for record linkage comparison as part of an ensemble of features. Technique A returns the difference between the WO or part timestamp to the outage start or finish time. Technique B assesses whether the timestamp is during the outage, if so then it returns zero and if not then it returns how far outside. It returns a positive time difference whether the timestamp is before or after the outage. Technique B has the benefit of registering positive agreement for timestamps that occur during long outages, where they may not be close to the start or finish times and so the PEOHH uses technique B.

The PEOHH compares each WO timestamp to the outage start and finish times to give a positive time difference (Δt_{Fe}). If the timestamp is during the outage then it records a Δt_{Fe} of zero. This section presents time differences for each of four Features:

- The time difference between the WO Start date and the outage (Δt_{St})
- The time difference between the WO Finish date and the outage (Δt_{Fi})
- The time difference between the WO Notification date and the outage (Δt_{No})
- The time difference between the Part posting date and the outage (Δt_{Pa})

These time differences will be referred to as *time features*. This research presents the agreement Threshold for each time Feature (Th_{Fe}). If $\Delta t_{Fe} < Th_{Fe}$ then the PEOHH records positive agreement. Where the timestamp is missing, the PEOHH records neither positive nor negative agreement. The PEOHH otherwise records negative agreement. This research uses a nominal Th_{Fe} value of 2 days for each feature.[30]

The PEOHH could alternatively compare the WO start date or notification date to the outage start date, or it could compare the WO finish date to the outage finish date. These alternative methods are less attractive than the method used by the PEOHH because it would count WOs that occur in the middle of long outages as non-matches.

The PEOHH calculates the Δt_{Pa} for each part associated with the WO. It compares the minimum Δt_{Pa} in the WO to Th_{Pa}. If $\Delta t_{Pa} < Th_{Pa}$ then the PEOHH records positive agreement. Where the part

[29] The blocking process was described in section 4.1.

[30] 7.6 will present the optimisation of Th_{Fe} for each of the time features alongside other interrelated weights and thresholds.

posting date timestamp is missing or if no parts are associated with the WO then the PEOHH records neither positive nor negative agreement. The PEOHH otherwise records negative agreement.

Rather than comparing the minimum Δt_{Pa} in the WO to Th_{Pa}, the PEOHH could alternatively use the mean or the median Δt_{Pa} in the WO, or it could use a combination of these averages with the minimum. Such techniques could be worthy of further work. They are not considered in this thesis because comparing too many techniques makes the optimisation of weights and thresholds[31] intractable.

Table 4-2 presents the outage timestamps and the WO timestamps for two POLRs. Both POLRs relate to the same WO. Table 4-2 is included in this section as an example of the existing records of wind turbine health history. The notification date is a month prior to the start date which implies that the issue in this example did not require urgent resolution.

POLR ID	Work Order			Outage	
	Start Date	Finish Date	Notification Date	Date Time On	Date Time Off
1	06/06/2018	10/06/2018	29/04/2018	2018-06-02 06:48:41	2018-06-02 10:49:16
2	06/06/2018	10/06/2018	29/04/2018	2018-06-06 07:32:56	2018-06-06 13:02:09

Table 4-2, Outage and WO Timestamps for Two POLRs

[31] Chapter seven will present the optimisation of the weights and thresholds.

Table 4-3 presents the comparison of the outage timestamps to the material timestamps for the two POLRs detailed in Table 4-3. This WO has four material line items recorded against it, meaning that four types of spare part were used to do this work. One of the MLIs (Part 2) has its part posting date missing. The PEOHH calculates Δt_{Pa} for each part. As illustrated in Table 4-3, it identifies the minimum Δt_{Pa} for each POLR.

Material		$\Delta t_{Pa}\Delta t_{Pa}$ (days)	
MLI ID	Posting Date	POLR ID	
		1	2
1	29/05/2018	4.28	8.31
2			
3	07/06/2018	4.55	0.46
4	06/06/2018	3.55	0.31
minimum		3.55	0.31

Table 4-3, Comparison of Outage Timestamps to Material Timestamps for Two POLRs

Table 4-4 presents the comparison of the outage timestamps to the WO and material timestamps of the two POLRs detailed in Table 4-2. If $\Delta t_{Fe} < Th_{Fe}$ then the PEOHH records positive agreement. Where the timestamp is missing, the PEOHH records neither positive nor negative agreement. The PEOHH otherwise records negative agreement.

POLR ID	Δt (days)				Agreement			
	Start	Finish	Notification	Part Posting	Start	Finish	Notification	Part Posting
1	3.55	7.55	34.28	3.55	False	False	False	False
2	0.31	3.46	38.31	0.31	True	False	False	True

Table 4-4, Comparison of Outage Timestamps to WO and Material Timestamps for Two POLRs

Each of the four timestamp features will be evaluated as part of an ensemble of features for record linkage comparison. Chapter 7 will present the optimisation of the weights and thresholds used in the PEOHH, that were defined in section 4.1, including the agreement and disagreement weights. If the optimised agreement and disagreement weights for some or all of the four timestamp features are not zero then this research will have found that the hypothesis that timestamps from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features is

true. This section has presented the hypothesis that timestamps from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features.

4.4.2 HHE Techniques Using the Order Type

This section will present the hypothesis that the 'type' feature from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features. It will present a very simple technique for this purpose. The use of a simple technique for the purpose of record linkage is analogous to the gender feature in the linkage of records of people. For example, if two records are recorded as both being male then a process for linking records of people would record positive agreement for the gender feature. Features such as the gender feature of people or the type feature of maintenance records are not used by themselves for the purpose of record linkage because they are too general but can be used as part of an ensemble of features. The use of simple features such as these is typical in record linkage (Dunn, 1946, Churches et al., 2002).

Section 2.1.2 described the outage types:

> "The 'type' feature classifies each outage as either corrective, predetermined, 'Balance of Plant / OFfshore Transmission Owner' (BoP/OFTO), condition-based, environmental or unknown."

Section 2.1.3.3 described the WO types:

> "The 'type' feature classifies each WO as either preventive, retrofit, inspections and surveys, condition-based or corrective. The feature is human generated by the planner who selects a type from a list of options."

This section presents two techniques that use the WO and Outage 'type' features for record linkage comparison as part of an ensemble of features. Technique A is to filter the outage data by type to include only corrective work. Technique B is that when the outage is recorded as corrective then positive agreement is recorded and that when the outage is recorded as other than corrective then negative agreement is recorded. Technique A assumes that all outages that match corrective WOs are classified as corrective in the outage data. Preliminary validation meetings elicited the expert advice that such an assumption would not be valid and so the PEOHH uses Technique B.

Backing up this expert opinion, this section can report that the type matches in only seven of the twenty-nine records in the GSSLR. This does not imply that the records are incorrect; different record systems may use different terminology to define the same features. This result means that the *PPV* of the GSSLR would be limited to a maximum value of 25% under technique A, which disproves the hypothesis that technique A would be useful for record linkage. Technique B, on the other hand, does not exclude those outages not classified as corrective from the EHH; it merely investigates the

possibility that there could be a statistical likelihood that if the type matches then it is more likely that the WO and the outage are a true match. For equivalent data, technique B can yield a maximum *PPV* of 100%.

The PEOHH compares the POLR that it generated using the blocking process[32]. Table 4-5 presents the comparison of the outage type to the WO type for two POLRs. The WO type is corrective for all POLRs because only corrective WOs are required for the applications detailed in section 1.2 and so the WO data has been filtered to only include corrective WOs. The outage type for POLR 1 is predetermined, that is that the database of outages has identified this outage as for scheduled maintenance rather than for a corrective repair. The PEOHH registers disagreement for POLR 1 and agreement for POLR 2. Each WO and each outage has a type so there are no instances of POLRs that register neither agreement nor disagreement for this feature.

POLR ID	WO Type	Outage Type	Agreement
1	Corrective	Predetermined	False
2	Corrective	Corrective	True

Table 4-5, Comparison of the Outage Type to the WO Type for two POLRs

If the feature agrees then the PEOHH records positive agreement. If the feature disagrees then it records negative agreement. If the feature were not recorded then it would record neither negative nor positive agreement, but the data does not include missing values for this feature and so this does not occur in practice.

The ‘type’ feature will be evaluated as part of an ensemble of features. This section has presented the hypothesis that this feature from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features.

[32] The blocking process was described in section 4.1.

4.4.3 HHE Techniques Using Visits

This section will present the hypothesis that the ‘number visits’ and ‘duration’ features from the outage data can be used for record linkage comparison as part of an ensemble of features.

Section 2.1.2 described the outage ‘number visits’ and ‘duration’ features:

> It is a safety requirement that when an OWT is visited to carry out maintenance, it is brought under local operation; a setting in which it cannot produce electricity. Whether or not there is a visit to the OWT, operators refer to intervals of non-production as outages. The ‘reset’ feature classifies each outage that has been classified as corrective as either a visit, a remote reset or an automatic reset. The ‘number visits’ feature is either zero or a positive integer. It is Ørsted’s estimate of how many times the OWT was visited during the outage. It is calculated using the ‘wind turbine in local operation’ alarm (Papatzimos et al., 2019).

The database of outages records the duration of each outage. This section presents two techniques that use the ‘number visits’ and ‘duration’ features from the outage data for record linkage comparison as part of an ensemble of features. Technique A filters the outage data by the ‘reset’ feature so that only records of outages labelled as a visit are included. Three experts agreed that, in practice, the correct match for one of the WOs tested in the validation exercise is not correctly identified as a visit in the outage data set. This outage was not identified as of the ‘corrective’ ‘type’ in the outage data set and so was not labelled as a visit. This identified that technique A is not an appropriate technique for the PEOHH.

Technique B filters the outage data by duration. 20 minutes is the minimum duration of those outages associated with a visit to the WT so all outages under 20 minutes duration, even if they are not recorded as involving a visit, are excluded. The PEOHH records agreement if the number of visits is not null. This is a probabilistic approach that will test the hypothesis that outages associated with a visit are more likely to be true matches to WOs.

It was described in section 4.1 that the PEOHH uses the POLRs that it generated using the blocking process. Table 4-6 gives three examples of the three features of the outage data called '*Reset*', '*Duration*' and '*Number Visits*' and of their use by the PEOHH:

- POLR 1 is shorter than 20 minutes and so, prior to blocking, the PEOHH filters out the outage that would generate this POLR.
- POLR 2 records the 'reset' feature as 'visit' but the PEOHH does not use this feature. POLR 2 records the 'number visits' feature as '1' so the PEOHH records positive agreement for this feature on this POLR.
- POLR 3 does not record a positive integer in the 'number visits' feature so the PEOHH records negative agreement for this feature on this POLR.

POLR ID	Reset	Duration (mins)	Number Visits	Agreement
1	Remote	19		
2	Visit	213	1	True
3	Remote	27		False

Table 4-6, Examples of the Technique Using Visits

The 'number visits' and 'duration' features will be evaluated as part of an ensemble of features. This section has presented the hypothesis that this feature from the outage data can be used for record linkage comparison as part of an ensemble of features.

4.4.4 HHE Techniques Using Features Indicative of the Failure Mode

Records of wind turbine health history contain records of the identification and repair of faults and these records contain features indicative of the failure mode. This section will present the hypotheses (1) that features indicative of the failure mode from the WO data, from the outage data, from the alarm data and from the material consumption data can be used as part of an ensemble of features for record linkage comparison. Following on from that, another hypothesis (2) is that a selection of these techniques could be used to the same effect.

The practice of manually linking WOs to outages starts by identifying the wind turbine and matching the WO basic start date to a time interval around the outage start time. This yields a list of candidate outages to match to each WO. Wind turbine experts then use their experience of possible failure modes to compare the outage alarm code 'description' to the WO 'description' field. The usefulness of features indicative of the failure mode in manual record linkage suggests that they may also be applicable to automatic record linkage, as in hypothesis (1).

Most of the detailed data within maintenance records are indicative of the failure mode. These are important features that differentiate one outage from another. Section 2.2 presented existing record linkage techniques. It was described in that section that record linkage techniques using various features of the data have proven effective in the fields of linking medical records (Sayers et al., 2015, Nasseh and Stausberg, 2016, Oliveira et al., 2016), address data (Churches et al., 2002, Comber et al., 2019, Lin et al., 2019), census data (Jaro, 1989, Smith et al., 2016) and genealogical records (Wilson, 2011) and that they have been used to detect duplicate internet search results (Hajishirzi et al., 2010).

The PEOHH uses three features of the data that are each indicative of the failure mode: the description, the alarm code and the list of parts used. The first three sub-sections of this section present record linkage techniques that each use one of these three features; section 4.4.4.4 is the conclusion to the HHE techniques using the failure mode.

For each of these three features this section compares two methods for linking WOs to outages. The first of these two methods, like the methods presented in sections 4.4.1 to 0, for each POLR in the SPOLR, compares the specific feature of the WO to the equivalent feature of the outage.

Section 2.1.2 presented the database of outages:

> "Ørsted label each outage with an alarm code indicative of the failure mode using a combination of automatic and manual methods. Their automatic system selects an alarm code from the database of outages using a confidential algorithm. Ørsted's data scientists sometimes later manually adjust these labels to better reflect the failure mode by discussing what happened with the technical team involved in the repair."

The second method recognises the possibility that an outage may be labelled with an alternative failure mode label to that in the database of outages. For each POLR, the second method first links the outage to a set of alarms. It then compares the specific feature of the WO to the equivalent feature of each alarm. If any of the alarms agrees then the method records agreement for the POLR. There may be alarms indicative of the failure mode recorded before an outage occurs. An OWT is typically not rotating during an outage and so it is these alarms that occur prior to an outage that are of particular interest when identifying the failure mode. Therefore, the PEOHH adds a time difference to the outage start and finish times to yield a time interval longer than that of the outage. The PEOHH identifies all the alarms that occur during this extended interval. The Threshold for the Time difference between Outages and Alarms (*TTOA*) denotes the extra time added. The PEOHH extracts all the

alarms from the alarm log from +/- *TTOA* of the outage start and finish times. This research used an initial value of *TTOA* of 10 minutes, the value initially recommended by a wind turbine expert[33].

This section will consider three features indicative of the failure mode: description, alarm code and parts. In applying these two methods to these three features, this section will present six techniques.

4.4.4.1 HHE Techniques Using the Description

This section will present two hypotheses. The first is that that the 'description' feature from the WO data and the description of the alarm code from the outage data can be used as part of an ensemble of features for record linkage comparison. The second is similar to the first except that it uses the database of alarms from which the database of outages was derived. It is that the 'description' feature from the WO data and the description of the alarm code from a set of alarms associated with the outage can be used as part of an ensemble of features for record linkage comparison.

The WT control unit generates alarms and some of these are indicative of a failure mode. (Qiu et al., 2012). These are logged by the condition monitoring unit. The outage data is labelled with one alarm code per outage and each alarm code corresponds to a text field called Description such as "cabinet too hot".

Section 2.1.3.4 described the WO description field:

> "The 'description' feature is a short, free text description of the WO. For corrective maintenance it often refers to an alarm code. It may contain the alarm code but it more often contains an abbreviated reference to the standard text description of the alarm code."

The methods presented in this section use the comparison of text strings involving some simple Natural Language Processing (NLP) techniques.

In NLP, text pre-processing is the practice of cleaning and preparing text data; transforming it into a more digestible form to improve the effectiveness of the analytic techniques that follow it. One of the simplest problems in NLP is to compare two text strings to determine how similar they are to each other. This section presents techniques which compare text strings that are indicative of the failure mode to identify whether they are similar to each other; if they are similar then they are more likely to refer to the same failure mode than if they are different. If they do refer to the same failure mode then they are more likely to refer to the same maintenance activity.

---

[33] Section 7.2 will use the GSSLR to investigate the effect of varying *TTOA* but will recommend keeping it at 10 minutes.

Prior to comparison, both strings are pre-processed to remove everything but the words. This section presents a simple, scalar technique for comparing what letters these text strings contain. An alternative would be to use a vector technique such as word2vec to apply semantic record linkage techniques such as those described in section **Error! Reference source not found.** that might find d escriptions that refer to the same activity using different words. Word2vec or other relatively complicated NLP techniques are not proportionate for this project which uses a simpler approach instead.

The PEOHH compares the POLRs that it generated using the blocking process.[34]

The PEOHH uses text pre-processing techniques that are standard in natural language processing. This section will describe these processes using for example the original text string:

‘404 Pitch A tracking during stops’:

- The PEOHH converts each string to upper case letters:

‘404 PITCH A TRACKING DURING STOPS’

- Tokenisation is to separate each string into tokens, that is shorter strings without spaces. Tokens can be words, numbers or alpha-numeric strings.

‘404, PITCH, A, TRACKING, DURING, STOPS’

- Lemmatisation is the removal of inflectional endings from each token (Porter, 1980):

‘404, PITCH, A, TRACKING, DURING, STOP’

- The PEOHH removes any numbers. (These numbers are typically alarm codes and these are used in a separate technique[35]). It does this using a Python algorithm posted by Stack Overflow internet forum user “Silenced Temporarily”, (Silenced Temporarily, 2019).

‘PITCH, A, TRACKING, DURING, STOP’

- The PEOHH joins the tokens together to generate a hyphenated string:

‘PITCH-A-TRACKING-DURING-STOP’

---

[34] The blocking process was described in section 4.1.

[35] A technique using alarm codes will be described in section 4.4.4.2.

By pre-processing each text string, the PEOHH avoids some errors that would otherwise be caused by differences between strings that have similar contents but different formatting.

Table 4-7 demonstrates the techniques used by the PEOHH for the pre-processing and for the comparison of text strings. It uses the example of two pairs of text strings (POLR 1 and POLR 2).

The *Levenshtein distance* between two text strings is the minimum number of single-character edits required to change one string into the other (Levenshtein, 1966). This research chose to use it because it provides a simple comparison of text strings. While it does not replicate the use of expert knowledge in manual record linkage that different words may refer to the same repair activity, it does identify different spellings that may refer to the same word.

The PEOHH calculates Levenshtein distance using a Python algorithm posted by Stack Overflow internet forum user "Adam Smith" (Smith, 2017).

The Similarity Ratio (*SR*) is defined by equation 4.3 (Sayers et al., 2015).

$$SR = 1 - \frac{\text{Levenshtein distance}}{Min\ string\ length} \tag{4.3}$$

The PEOHH registers agreement when *SR* is above the Description Threshold (Th_{De}). Th_{De} is nominally set to 0.75 but section 7.3 will investigate the effect of varying it.

POLR ID		WO Description	Outage or Alarm Alarm Text	Levenshtein distance	Minimum String Length	*SR*	Agreement
1	Raw	Inv. Cool. W. temp hiqh	Inv. cool. w. temp high				
	Pre-processed	LNV-COOL-TEMP-HIQH	INV-COOL-TEMP-HIGH	2	18	89%	Yes
2	Raw	UPS battery failure	404 Pitch A tracking during stops				
	Pre-processed	UPS-BATTERY-FAILURE	PITCH-A-TRACKING-DURING-STOP	24	19	21%	No

Table 4-7, Pre-processing and Comparison of Text Strings from two POLRs

Some of the outages in the database of outages are labelled with an alarm code that is not indicative of the failure mode, such as 'manual stop'. In these cases, the PEOHH sets SR to zero and

consequently registers negative agreement. This is done to avoid the risk that descriptions not indicative of the failure mode might falsely register agreement.

The ‘description’ feature will be evaluated as part of an ensemble of features. This section has presented the hypotheses that:

1) The ‘description’ feature from the WO data and the description of the alarm code from the outage data can be used as part of an ensemble of features for record linkage comparison.
2) The ‘description’ feature from the WO data and the descriptions of the alarm codes from a set of alarms associated with the outage can be used as part of an ensemble of features for record linkage comparison.

4.4.4.2 HHE Techniques Using Alarm Codes

This section will present two hypotheses. The first is that that the ‘long text’ feature from the WO data and the alarm code from the outage data can be used as part of an ensemble of features for record linkage comparison. The second is similar to the first except that it uses the database of alarms from which the database of outages was derived. It is that the ‘long text’ feature from the WO data and the alarm code from a set of alarms associated with the outage can be used as part of an ensemble of features for record linkage comparison.

Section 2.1.3.5 described the WO ‘long text’ field:

> “The ‘long text' feature is a free text description of the notification and of the WO of unlimited length. For corrective maintenance it often contains semi structured recent entries from the alarm log that include alarm codes. These alarm log entries are automatically copied in when the notification is created and are typically error free. It can also contain unstructured notes made by the maintenance team relating to faults or to maintenance activities, particularly if these are considered unusual.”

The methods presented in this section identify alarm codes in the WO ‘long text’ field. If the WOs ‘long text’ feature contains the alarm code then the PEOHH records agreement. If it does not, including if the long text field is empty, then it registers disagreement. Table 4-8 presents two examples.

POLR ID	WO Long Text	Outage or Alarm Alarm Code	Agreement
1	11.01.2017 09:32:25 CET name (email) transformer room temperature sensor error, fault code 12114, ----------[36] 01.03.2017 14:06:53 CET name (email) We found the pt100 sensor had detached from the cable in the transformer = room. So we replaced the the pt100. All ok after power up.	1001	No
2		12114	Yes

Table 4-8, Examples of the Technique Using Alarm Codes

An additional technique could identify alarm codes in the WO ‘description’ field. That additional technique was not developed in this research because this research identified by familiarisation with the data that the WO ‘long text’ field is a richer data set than the WO ‘description’ field. The WO ‘long text’ field for corrective maintenance includes the text of the notification of a fault that initiated the generation of a WO. While there often is an alarm code in the WO ‘description’ field, this alarm code is always copied from the WO ‘long text’ field. In the 3 years that this research spent investigating the WO data, it found no examples of an alarm code in the WO ‘description’ field that was not also in the WO ‘long text’ field. Wind turbine experts agreed that, where WO ‘long text’ data is available, there is no value in also looking for alarm codes in the WO ‘description’ field.

The use of the WO ‘long text’ field, rather than its ‘description’ field, for alarm code identification contrasts with the techniques presented in section 4.4.4.1 that use the WO ‘description’ field for the comparison of text strings. The WO ‘description’ field, after pre-processing, is of a comparable length to the alarm code description, allowing for a straightforward comparison that would be less computationally expensive than searching for text strings in the WO ‘long text’ field would be. The decision to use the WO ‘description’ field in the techniques presented in section 4.4.4.1 but the WO ‘long text’ field in this section is therefore appropriate.

The PEOHH calculates the ‘alarm code’ feature using a Python algorithm posted by Stack Overflow internet forum user “BENY” (BENY, 2019).

[36] ---- Confidential information redacted

The ‘alarm code’ feature will be evaluated as part of an ensemble of features. This section has presented the hypotheses that:

1) The ‘long text’ feature from the WO data and the alarm code from the outage data can be used as part of an ensemble of features for record linkage comparison.
2) The ‘long text’ feature from the WO data and the alarm codes from a set of alarms associated with the outage can be used as part of an ensemble of features for record linkage comparison.

4.4.4.3 HHE Techniques Using the Parts

This section will present two hypotheses. The first is that the material consumption data and the alarm code from the outage data can be used as part of an ensemble of features for record linkage comparison. The second is similar to the first except that it uses the database of alarms from which the database of outages was derived. It is that the material consumption data and the alarm code from a set of alarms associated with the outage can be used as part of an ensemble of features for record linkage comparison.

Section 2.1.4 described the database of material consumption:

> “The material consumption database lists what parts were used in the maintenance of the OWTs. Each Material consumption Line Item (MLI) refers to a single part number and is assigned to an order number. Some WOs have no material consumption line items assigned to them while others have many. Materials include replacement parts as well as consumables such as oil, grease or paint. The ‘material’ feature is the part number: the identifier of the design of the part. The ‘description’ feature describes the part number.”

The methods presented in this section use statistical techniques that seek to identify whether the parts assigned to a WO are typical of the failure mode of the outage. When the maintenance team decide which parts to use to repair a fault they provide an expert validation of the fault diagnosis. The methods presented in this section study patterns in the material consumption data that embody this expert diagnosis.

This section presents two techniques: BNB Classification and a frequency-based technique. It shows that BNB classification does not support useful interpretation with the unbalanced health history data but that the frequency-based method does. Both techniques require training data that they use with the intention of recognising which parts are typical of each failure mode.

4.4.4.3.1 Training Data

This section will present the Process for the Identification of the Training data (PIT). The PIT provides the training data to the PEOHH, which uses it to identify which parts are indicative of each alarm code.

Both the BNB classification technique presented in section 4.4.4.3.2 and the frequency-based technique presented in section 4.4.4.3.3 use the same training data selected for their relatively good quality. This quality is predicted using the overall score from the other features as a metric for the confidence of the correctness of the failure mode. The use of other features than the parts to identify the training data means that the parts-based techniques are not independent of the other techniques. The parts-based techniques are consequently also not independent of the database of alarms. While the material consumption data embody an expert validation of the fault diagnosis, these dependencies on the other features hinder the technique's ability to capture that independence from the other features. Unfortunately, such dependencies are logically necessary because any model would need to be trained to identify which parts are typical of each failure mode.

Section 4.3 presented the technique that the PEOHH uses to select POLRs for the EHH. It was described in that section that:

> "The PEOHH sorts the POLRs by match ID, by Δt_{St} and then by S_{POLR} and then it filters the POLRs to retain only the first POLR for each WO. This process identifies a version of the EHH with the highest S_{POLR} for each WO."

The PEOHH selects POLRs for inclusion in the training data using the same criteria that it uses to select POLRs for the EHH. It then filters the training data to retain those POLRs that have S_{POLR} above a threshold (Th_{SP}) of 1.7. Section 7.4 will present the effect of varying this threshold. A higher Th_{SP} would mean that the training data were of higher quality, tending to contain fewer misleading POLRs where the WO is matched to an outage that does not relate to it, but of lower quantity, containing fewer POLRs. At the optimum value of Th_{SP} there will be enough data to train on but it will be of high enough quality that it is not too misleading.

Chapter 7 will recommend the optimum weights and thresholds for the PEOHH. This section uses these the optimised weights and thresholds detailed in table 7.10 to identify the training data.

Both techniques presented in this chapter consider which MLIs are assigned to the WO; whether a given part was used rather than how many of the part were used.

The PEOHH identifies the training data as a two-dimensional binary dataset of which parts are assigned to each WO. Where a part is assigned to a WO, the PEOHH ascribes the value TRUE to the data element for that part number and that order number. Where a given part number is not assigned

to a given WO, the PEOHH ascribes the value FALSE to the data element for that part number and that order number.

Unbalanced data are data in classification problems where there are unequal instances for different classes. The training data is unbalanced. For the farm used, the training data contains 560 parts of which only 192 are used more than once. It contains 156 alarm codes of which only 21 occur more than 5 times. Unbalanced data tends to bias machine learning models to predict the more common class.

Sections 4.4.4.3.2 and 4.4.4.3.3 present two techniques that use the material usage data for record linkage comparison. Section 4.4.4.3.3 presents a simpler technique that works more robustly despite the unbalanced data.

4.4.4.3.2 HHE Techniques Using Bernoulli Naïve Bayes Classification

This research investigated a technique in which the PEOHH used a Bernoulli Naïve Bayes (BNB) classifier[37] to predict the probability for each POLR that the parts used in the WO correspond to the alarm code of the outage. This section shows that the parts data can't be used to calculate a useful metric using BNB Classification and investigates why it doesn't work.

Methodology

A BNB classifier was used to predict the probability that the parts used in the WO correspond to the alarm code of the outage. Each feature of a multi-variate Bernoulli model is a binary variable. A probabilistic model is trained using binary data; in this case which parts were used. Each row; in this case a Pair of Linked Records; is assigned a class, in this case the outage alarm code.

BNB identifies the Log Odds that each POLR belongs to a given class (LO_{Class}). LO_{Class} is defined by equation 4.4:

$$LO_{Class} = log\left(\frac{P_{Class}}{1 - P_{Class}}\right) \tag{4.4}$$

Where P_{Class} is the probability predicted by the BNB algorithm that each POLR belongs to a given class.

This section presents the results of testing the BNB method on a known failure mode. A specific set of fans are replaced when one of them fails and this failure mode is indicated by a specific "cabinet too

[37] Section 2.3 reviewed classification techniques including BNB.

hot" alarm. The classifier should predict a higher probability of this alarm when this part is used than when it is not used.

This report presents a measure called 'classification score' that measures whether clear classification has been achieved. It is defined by equation 4.5 where *A, B, C, D* are integrals of the number of linked records with respect to predicted probability of the "cabinet too hot" alarm. *A, B, C, D* are defined below but equation 4.5 only uses *C* and *D*. This section will use this measure to identify whether successful classification has been achieved. Successful classification has been achieved if the range of *LO* scores for one class does not overlap with the range of *LO* scores for the other class.

$$Classification\ Score = \frac{D - C}{C + D} \tag{4.5}$$

Consider for example the LO_{Class} of the "cabinet too hot" alarm (LO_{Alarm}). In this example:

- *A* is the number of POLRs where both the fan was replaced and the LO_{Alarm} is less than the minimum LO_{Alarm} for all the POLRs where the fan was not replaced.
- *B* is the number of POLRs where both the fan was replaced and the LO_{Alarm} is more than the minimum LO_{Alarm} for all the POLRs where the fan was not replaced.
- *C* is the number of POLRs where the fan was not replaced and the LO_{Alarm} is less than the maximum LO_{Alarm} for all the POLRs where the fan was replaced.
- *D* is the number of POLRs where the fan was not replaced and the LO_{Alarm} is more than the maximum LO_{Alarm} for all the POLRs where the fan was replaced.

A and *D* are the number of POLRs in the regions where the classes do not overlap. *B* and *C* are the overlap between the two classes.

A, *B*, *C* and *D* are illustrated in Figure 4-1, a frequency diagram of the Count of POLRs (*CO*) against LO_{Alarm}. The figure shows those POLR where the fan was replaced (*A* and *B*) and where it was not replaced (*C* and *D*).

Figure 4-1(a) presents an example where perfect classification has not been achieved. The classes do overlap so *B* and *C* are not zero. Imperfect classification means that the distributions *A* and *B* have an overlap with distributions *C* and *D* and is indicated by a classification score less than 1. Negative classification scores mean that *C* > *D*; more than half of the POLRs are not clearly classified. The figure features vertical lines that illustrate aspects of the definitions of *A*, *B*, *C* and *D* (listed above). The line between *A* and *B* is the minimum LO_{Alarm} for all the POLRs where the fan was not replaced and the line between *C* and *D* is the maximum LO_{Alarm} for all the POLRs where the fan was replaced.

Figure 4-1(b) presents an example where perfect classification has been achieved. The classes do not overlap so *B* and *C* are zero. Perfect classification means that the distributions *A* and *B* have no overlap with the distributions *C* and *D* and is indicated by a classification score of 1.

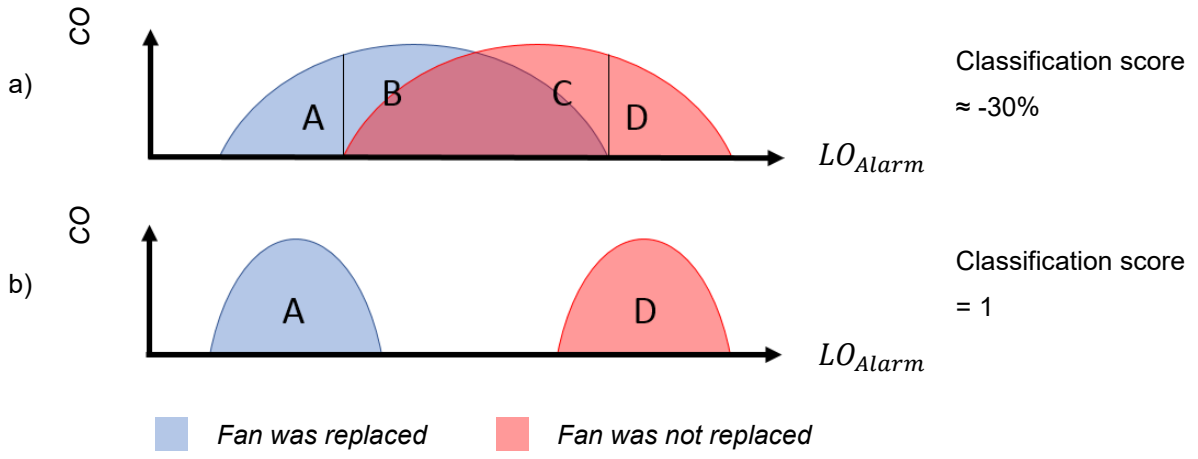


Figure 4-1 Demonstration of the Classification Score: Count (POLRs) (CO) against Log Odds (LO_{Alarm})

When the only part considered is the fan, clear classification is achieved, which confirms the hypothesis that a part can be used to identify the failure mode. A useful record linkage technique could be developed if parts could be used to identify the failure mode when the relationship between the part and the failure mode was not already known.

This research implemented the model using Python library Scikit-learn (Pedregosa et al., 2011).

Results

This research investigated the effect of considering additional parts as well as the fan. Sets of additional parts of various number of parts were considered as well as the fan. The parts to include in these sets were selected at random. The result of using each set of parts is shown in Figure 4-2. It shows that clear classification is achieved when fewer than around 20 to 30 parts are included but that it is not achieved when 60 or more are. For this application, all the 156 failure modes and all the 560 parts would need to be included, which exceeds the number of parts from this unbalanced dataset for which BNB classification would be an appropriate method.

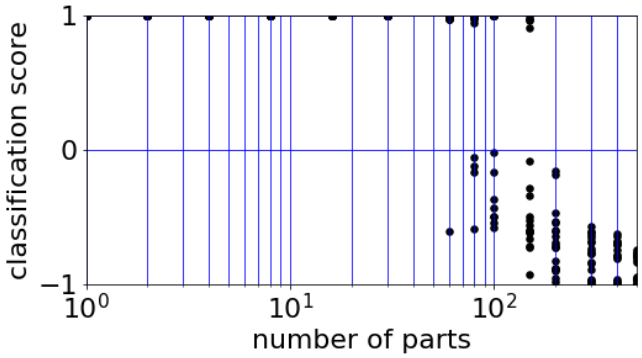


Figure 4-2, BNB Classification Score against Number of Parts

The following section will present an alternative technique that cannot suffer from the problem with unbalanced data because it only looks at the relevant event code.

4.4.4.3.3 HHE Techniques Using the Parts Frequency

This section will present a simpler technique to that presented in the previous section that also uses the records of material consumption for record linkage comparison, the Parts Frequency Technique (PFT).

The PFT can be used with any test data. It identifies an individualised set of filtered training data for each Pair of Linked Records in the test data (test POLR). The failure mode labels from the training data are used to train the PFT. For each test POLR, the training data[38] are filtered to only include those POLRs with the same outage alarm code as the test POLR. To avoid testing and training on the same data, the test POLRs WO is excluded from the filtered training data. The test POLR is given a

[38] The training data were described in section 4.4.4.3.1.

“Parts Score” (*PS*) with an initial value of 0. If the WO contains a part that is in the filtered training data, then *PS* is increased by 1. If the WO contains a part that is not in the filtered training data then *PS* is reduced by 1. The PFT repeats this process of identifying an individualised set of filtered training data and calculating *PS* for each POLR in the test data. Figure 4-3 presents the algorithm that calculates *PS*.

As an example of the technique, consider a WO with 4 parts. 2 of these parts are, according to the training data, typical of the alarm code but 2 of the parts are atypical of it. *PS* = 2 - 2 = 0. Further work could look at applying unequal weightings to the typical and to the atypical parts.

```
1:  C_F_H          = Identify the set of the best POLRs for training
2:  Tr_D           = Join C_F_H to the parts data
3:  Di_Q_B         = Create a dictionary of the alarm code for each POLR in Tr_D
4:  Loop: for each POLR in the test data:
5:                 WO          = Identify the order number of the POLR
6:                 Te_C        = Identify the parts for the WO
7:                 Tr_C        = Filter the training data to remove the WO
8:                 Tr_E        = Filter Tr_C to only include the alarm code of the POLR,
                                 identified using Di_Q_B
9:                 Q_Tr        = Create a table of whether each part is in each WO in Tr_E
10:                Z_Tr        = Create a table of whether any WOs in Q_Tr contain each part
11:                train       = Create list of which parts are in Z_Tr
12:                Te_C['T']   = Identify whether the parts in Te_C are in train
13:                typical     = sum of Te_C['T']
14:                atypical    = (length of Te_C) - typical
15:                PS          = typical - atypical
```

Figure 4-3, Pseudo Code for the PFT

Line 9 of the code in Figure 4-3 considers all the parts in the database of material consumption. This is a longer list of parts than that in the training data. The PEOHH uses this longer list of parts because there are cases where a POLR to be tested contains a part that is not in the training data. In these cases, the part is considered to be atypical of the alarm code. The PEOHH does this using a Python algorithm posted by Stack Overflow internet forum user “piRSquared” (piRSquared, 2018).

While this technique considers whether the parts assigned to a WO are typical of the POLR alarm code, further work could consider whether they are also typical of another alarm code. If they are then this might imply that the POLR is less likely to be a true match than if the parts are exclusively associated with the alarm code.

4.4.4.3.4 Conclusion to the HHE Techniques Using the Parts

This section used a database of material consumption in which each material line item is labelled with an order number. It developed two techniques that use these data as part of an ensemble of features for record linkage comparison.

The first technique used a BNB classifier to predict the probability for each POLR that the parts assigned to the WO correspond to the alarm code of the outage. Historical data on machinery failures tends to be unbalanced such that some failure modes feature more than others. It showed that, because of the unbalanced data, BNB classification is not an appropriate method.

The second technique is simpler, checking whether or not records for the outage failure mode contain each part assigned to the WO. It does not suffer from the problem with unbalanced data because it only looks at the relevant event code.

The 'parts' feature will be evaluated as part of an ensemble of features. This section has presented the hypotheses that:

1) The material consumption data and the alarm code from the outage data can be used as part of an ensemble of features for record linkage comparison.
2) The material consumption data and the alarm codes from a set of alarms associated with the outage can be used as part of an ensemble of features for record linkage comparison.

4.4.4.4 Conclusion to the HHE Techniques Using the Failure Mode

The hypothesis presented in section 4.4.4 is that that features indicative of the failure mode from the WO data and from the outage data can be used for record linkage comparison as part of an ensemble of features. It considered three of these; description, alarm code and parts. For each of these three features this section compares two methods for linking WOs to outages. The first of these two methods uses the alarm labels in the outage data while the second uses the alarm codes of a set of alarms that occurred around the start and end of the outage. In applying these two methods to these three features, this section presented six techniques.

Chapter 7 will show that this thesis cannot offer conclusive advice as to whether or not to use any of the features indicative of the failure mode for record linkage comparison as part of an ensemble of features. This research predicts that this uncertainty will be drastically reduced by an upcoming

innovation in maintenance record keeping that will be discussed in chapter 9; the automatic linking of new WOs to outages will result in a larger GSSLR and consequently in more certainty for this estimate.

4.5 Computation Times

Table 4-9 presents the computation time for each technique. The technique using the description and the technique using the parts are both relatively time consuming. These computations were performed on a standard laptop[39]. These computation times do not represent a problem for running them once per day which would be sufficient for everyday record linkage implementation. Such algorithms are routinely run automatically and are the responsibility of Ørsted's IT department, they are not run by the technicians themselves. These run times did represent a problem for the optimisation of the weights and thresholds in this research, where they needed to be repeated hundreds of times[40].

Process	Duration seconds	Technique Described in Section
Import the data	19.71	4.1
Link WOs to outages to create the POLRs	3.34	4.1
Type	4.27	4.4.2
Visits	0.06	0
Link alarms to outages	109.39	4.4.4
Pre-processing	428.11	4.4.4.1
Description	520.61	4.4.4.1
Alarms	12.67	4.4.4.2
Identify the training data	0.30	4.4.4.3.1
Parts	1200.94	4.4.4.3.3
Calculate the weights using all features	0.32	4.1

Table 4-9, Computation Time for Each Technique

Further work optimising these computationally expensive algorithms or running them on a faster computer (such as a parallel cluster) would improve these computation times but is outside of the

[39] Intel Core i7-8665U CPU @ 1.90GHz 2.11GHz

[40] Computational run time was an issue for calculations of the coverage of techniques that estimate the uncertainty of a difference between two binomial proportions (section 6.2.6) and for optimising the weights and thresholds (chapter 7).

scope of this thesis. Ørsted's confidential data may not be copied onto Durham University parallel clusters, but Ørsted might consider using their own in-house parallel clusters for this purpose.[41]

4.6 Conclusions to the Techniques for Health History Enrichment

This chapter presented the PEOHH and how this thesis will validate it. The PEOHH can use an ensemble of up to twelve record linkage features for joining WOs to outages. That ensemble of features was broken down into four sub sections:

- Four timestamp-based features
- A feature that considers the recorded type of maintenance
- A feature that uses the records of visits to the wind turbine
- Six failure mode-based features

The following results chapters will test that ensemble of features.

[41] Chapter 9 discusses the effect on this research of not using more computing power for the optimisation described in chapter 7.

5 Results: Examples of Health History Enrichment

Chapter four introduced the Process for the Enrichment of OWT Health History (PEOHH). This chapter offers two example Work Orders (WO) and uses them to illustrate the PEOHH by linking them to outages. Agreement Weights (*AW*) and the agreement Threshold for each time Feature (Th_{Fe}) were defined in chapter four and chapter seven will illustrate the effect that varying them has on the Enriched Health History (EHH). This chapter will illustrate the effect that varying them has on the individual Pairs of Linked Records (POLR) that make up the EHH.

5.1 Example 1

Table 5-1 presents the first example WO, the replacement of a detached temperature sensor, PT100, a resistance thermometer.

Some information about the WO is not published here for reasons of commercial sensitivity. The first five characters of the functional location that identify the wind farm are redacted, as are the following three characters that identify the row and the specific wind turbine. In the example in Table 5-1, the WO does not record any further details of the functional location. The name (name) and email address (email) of the two technicians who recorded their work in the WO long text field are also redacted.

The order ID is a unique identifier for the WO[42]. The table shows that the start date is recorded as two months later than the notification date. The long text corroborates that the notification was generated on the recorded notification date and that the corrective work was done on the recorded basic start date. The long text also details the times that these notes were generated at, in Central European Time (CET). There is no finish date recorded for this WO.

Order ID	Notification Date	Start Date	Finish Date	Maintenance Activity Type	Functional Location	Description
80116285	11/01/2017	01/03/2017		Corrective	********	transformer room temp sensor error
Long Text	11.01.2017 09:32:25 CET name (email) transformer room temperature sensor error, fault code 12114, ----------[43] 01.03.2017 14:06:53 CET name (email) We found the pt100 sensor had detached from the cable in the transformer = room. So we replaced the the pt100. All ok after power up.					

Table 5-1, First Example WO

[42] The database of WOs was described in section 2.1.3.

[43] ---- Confidential information redacted

Table 5-2 presents the material consumption data for the first example WO. The 'posting date', which refers to when the material line item was used, corroborates the date on which the temperature sensor was replaced. The 'reserved' field details the number of pieces (PC) that were reserved to be available so that this maintenance activity could be actioned: one sensor. The material description field details that the sensor is fitted with a 6m long cable that is shielded against electromagnetic noise.

Material	Material Description	Reserved	Unit	Posting Date
A9B00017802	SENSOR PT100 /6M SHIELDED WIRE	1	PC	01/03/2017

Table 5-2, Material Consumption Data for Order 80116285

The PEOHH links the WO to all the outages that start or finish within forty days of the start date[44]. Table 5-3 presents two of the outages linked by the PEOHH to the first example WO, referred to here as outage A and outage B. Ørsted label each outage with an alarm code indicative of the failure mode using a combination of automatic and manual methods[45]. The table details the description of each alarm code.

The database of outages classifies outage A as 'predetermined' and outage B as 'corrective'. Section 2.1.2 explained that the 'reset' feature classifies each outage that has been classified as corrective as either a visit, a remote reset or an automatic reset and that "the 'number visits' feature is either zero or a positive integer. It is Ørsted's estimate of how many times the OWT was visited during the outage. It is calculated using the 'wind turbine in local operation' alarm."

Outage ID	Date Time On	Date Time Off	Alarm Code	Description	Outage Type	Reset	Number Visits
A	01/03/2017 08:36:33	01/03/2017 11:20:41	1001	Manual stop	Predetermined		1
B	30/03/2017 13:08:12	01/04/2017 10:39:12	63025	Smoke in the A3 box	Corrective	Visit	1

Table 5-3, Two Outages Linked to Order 80116285

[44] This blocking threshold will be optimised in section 7.1.

[45] The database of outages was described in section 2.1.2.

It will be clear to the reader, by comparison of the WO 'long text' field with the outage timestamps, that outage A is the correct match for the WO. Outage B happened a month later and in this case, correct record linkage could be achieved using one feature, the WO start date. This section will use this example to illustrate the PEOHH.

Table 5-4 shows the duration of outage A and outage B. Both outages have more than the minimum duration of 20 minutes required by section 4.4.3. It also shows the time difference (Δt_{Fe}) for each of the four time features[46].

The POLR made up of the WO and outage A has Δt_{St} and Δt_{Pa} of 0.36 days (8.6 hours), which is less than the 2-day agreement threshold (Th_{Fe}) that was proposed in section 4.4.1 and that will be optimised in section 7.6. These time differences are shown in green. The other time differences are above this threshold and are shown in pink. There is no time difference for the finish time feature because this item is missing from the WO and these missing time differences are shown in orange.

Outage ID	Duration (days)	Δt_{Fe} (days)			
		Start	Finish	Notification	Part Posting
A	0.11	0.36		49.36	0.36
B	1.9	29.55		78.55	29.55

Table 5-4, Duration and Time Difference (Δt_{Fe}) for Two Outages Linked to Order 80116285

Section 4.1 presented the PEOHH:

> "The PEOHH calculates a Weight for each Feature for each POLR ($W_{Fe_{POLR}}$) using the Agreement Weight for that Feature (AW_{Fe}), the Disagreement Weight for that Feature (DW_{Fe}) and equation 4.1. The following pages of this section will present the process used in this research for estimating optimal values of AW_{Fe} and DW_{Fe} for each feature. Where AW_{Fe} and DW_{Fe} for a feature can be set to zero, this research has not identified evidence of a benefit of calculating that feature and so operators might consider

[46] The time features were described in section 2.1.3.1.

disregarding that feature so as to avoid computation with no proven drop in the quality of record linkage.

$$W_{Fe_{POLR}} = \begin{cases} AW_{Fe}, & Feature\ registers\ agreement\ _{POLR} \\ DW_{Fe}, & Feature\ registers\ disagreement\ _{POLR} \\ 0, & Feature\ registers\ neither_{POLR} \end{cases} \quad (4.1)$$

The PEOHH calculates a score for each POLR (S_{POLR}), the sum across all the features of the Weight for that Feature and for that POLR ($W_{Fe_{POLR}}$) using equation 4.2."

$$S_{POLR} = \sum_{Fe} W_{Fe_{POLR}} \quad (4.2)$$

Table 5-5 shows the calculation of S_{POLR} for the POLRs made up of the WO and the two outages. In this example, for each feature, AW_{Fe} is set to one (shown in green) and DW_{Fe} is set to negative one (shown in red). These weights will be optimised in section 7.6. Because the finish time is missing from the WO, this feature registers neither agreement nor disagreement and is set to zero (shown in orange). The table shows each of the ensemble of twelve features presented in chapter 4. For both POLRs in this case, most of the features register disagreement.

For the six features indicative of the failure mode that were presented in section 4.4.4, all register disagreement between the WO and the outage. The alarm code for outage A, 1001, “manual stop”, is not indicative of any specific failure mode and the alarm code for outage B, 63025, ‘smoke in the A3 box’, does not match either the WO description or long text. The disagreement for the parts feature[47] indicates that this alarm code is not linked to this temperature sensor in the parts feature training data. The alarm code feature registers agreement between the WO and an alarm that the PEOHH has linked to outage B[48].

S_{POLR} is the sum of the row and in this case both POLRs yield the same value, negative five.

Outage ID	Timestamp				Failure Mode						Type	Visits	S_{POLR}
					Outage			Alarm					
	Start	Finish	Notification	Part Posting	Alarm Code	Description	Parts	Alarm Code	Description	Parts			
A	1	0	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-5
B	-1	0	-1	-1	-1	-1	-1	1	-1	-1	1	1	-5

Table 5-5, Calculation of S_{POLR} for Two Outages Linked to Order 80116285

As both POLRs yield the same value of S_{POLR}, the PEOHH selects the POLR with the smallest Δt_{St}, in this case outage A. For this WO, the PEOHH has selected the correct outage without optimisation. For other WOs, optimisation of the weights and thresholds used in the PEOHH improves the quality of record linkage, as will be demonstrated in Chapter 7.

[47] The parts feature was presented in section 4.4.4.3.3.

[48] The linkage of outages to alarms was presented in section 4.4.4.

Table 5-6 investigates the effect of changing the Agreement Weight for the Start time feature (AW_{St}) from positive one to negative one (underlined). The POLR with the highest value of S_{POLR} is now the POLR made up of the WO and outage B and has an S_{POLR} of negative five while the POLR made up of the WO and outage A now has an S_{POLR} of negative seven. Due to the change to AW_{St}, the PEOHH has selected a different outage. This change would reduce the quality of record linkage and it would reduce its measure, Positive Predictive Value (*PPV*).

Outage ID	Timestamp				Failure Mode								
					Outage			Alarm					
	Start	Finish	Notification	Part Posting	Alarm Code	Description	Parts	Alarm Code	Description	Parts	Type	Visits	S_{POLR}
A	<u>-1</u>	0	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-7
B	-1	0	-1	-1	-1	-1	-1	1	-1	-1	1	1	-5

Table 5-6, Calculation of S_{POLR} for Two Outages Linked to Order 80116285 after changing the Agreement Weight for the Start time feature (AW_{St})

This section has illustrated the effect that varying AW_{St} has on an individual POLR. Chapter seven will illustrate the effect that varying it has on the Enriched Health History (EHH).

5.2 Example 2

Table 5-7 presents the second example WO, this time triggered by an alarm in the wind turbine's drive train subsystem which is referred to in the Reference Designation System for Power Plants standard for wind turbines (V.G.B. PowerTech, 2014) as 'MDK'. The Low speed Monitoring Unit (LMU) is located on the low speed side of the power train; between the rotor and the gearbox. It is an emergency system that triggers a hydraulic brake in the event of an 'overspeed'; an otherwise dangerous event where the rotary speed of the wind turbine exceeds a set limit. (European Patent Application EP 2 339 174 A1, 2011). This time, there is no entry in the 'long text' field.

Order ID	Notification Date	Start Date	Finish Date	Maintenance Activity Type	Functional Location	Description
80166733	22/11/2018	24/11/2018	25/11/2018	Corrective	******** MDK	LMU alarm overspeed
Long Text						

Table 5-7, Second Example WO

The material consumption data in Table 5-8 show that as well as the LMU, the maintenance team also replaced an inductive sensor, a cable, an optical isolator (optocoupler) and a converter for linearising electronic measurement (PR module 4222). The replacement of such auxiliary components is standard operating practice because re-using a component, even one that is not faulty, is typically associated with a shorter time to failure than replacement.

Material	Material Description	Reserved	Unit	Posting Date
A9B00030596	Sensor cable M12 8POL 5M	1	PC	
A9B00030713	Sensor Inductiv M12 5Mtr.	1	PC	
A9B00300151	OPTO COUPLER 24-60VDC	1	PC	
A9B10124433	LMU UNIT 3.6MW	1	PC	24/11/2018
A9B10162244	PR modul 4222 V/F 0-5V/0-10.7Hz	1	PC	

Table 5-8, Material Consumption Data for Order 80166733

Table 5-9 presents three of the outages linked by the PEOHH to the second example WO, referred to here as outage C, D and E. Outage C involved a visit to the wind turbine but not the resolution of the fault as it recurs three hours later and is remotely reset. Work to replace the LMU was carried out on a second visit, two and a half hours after the remote reset, outage E.

Outage ID	Date Time On	Date Time Off	Alarm Code	Description	Outage Type	Reset	Number Visits
C	22/11/2018 21:30:51	23/11/2018 10:14:34	6101	LMU alarm overspeed	Corrective	Visit	1
D	23/11/2018 13:29:29	23/11/2018 17:06:27	6101	LMU alarm overspeed	Corrective	Remote	
E	23/11/2018 19:24:22	24/11/2018 11:38:14	6101	LMU alarm overspeed	Corrective	Visit	1

Table 5-9, Two Outages Linked to Order 80166733

Table 5-10 shows the duration of outage C, D and E. All three outages have more than the minimum duration of 20 minutes required by section 4.4.3. It also shows Δt_{Fe}. For each POLR in the table, the PEOHH registers agreement for each of the four time features.

Outage ID	Duration (days)	Δt_{Fe} (days)			
		Start	Finish	Notification	Part Posting
C	0.53	0.57	1.57	0.90	0.57
D	0.15	0.29	1.29	1.56	0.29
E	0.68	0.00	0.52	1.81	0.00

Table 5-10, Duration and Time Difference (Δt_{Fe}) for Three Outages Linked to Order 80166733

Table 5-11 shows the calculation of S_{POLR} for the POLRs made up of the WO and the three outages. Outage D registers disagreement for the 'visits' feature because this outage was resolved by a remote reset and so did not involve a visit to the wind turbine. Apart from this feature, all the other features register the same agreement pattern for each outage.

Outage ID	Timestamp				Failure Mode								
					Outage			Alarm					
	Start	Finish	Notification	Part Posting	Alarm Code	Description	Parts	Alarm Code	Description	Parts	Type	Visits	S_{POLR}
C	1	1	1	1	0	1	-1	0	1	-1	1	1	6
D	1	1	1	1	0	1	-1	0	1	-1	1	-1	4
E	1	1	1	1	0	1	-1	0	1	-1	1	1	6

Table 5-11, Calculation of S_{POLR} for Three Outages Linked to Order 80166733

As both outage C and outage E yield the same value of S_{POLR}, the PEOHH selects the POLR from these with the smallest Δt_{St}; outage C.

Table 5-12 investigates the effect of changing the agreement Threshold for each time Feature (Th_{Fe})[49] from two days to zero (underlined). Setting these four thresholds to zero has the effect of only registering agreement when the WO timestamp occurs during the outage. The POLR with the highest value of S_{POLR} is now the POLR made up of the WO and outage E. Changing the thresholds has caused the PEOHH to link the WO to the outage where the work was carried out.

Outage ID	Timestamp				Failure Mode						Type	Visits	S_{POLR}
					Outage			Alarm					
	Start	Finish	Notification	Part Posting	Alarm Code	Description	Parts	Alarm Code	Description	Parts			
C	-1	-1	-1	-1	0	1	-1	0	1	-1	1	1	-2
D	-1	-1	-1	-1	0	1	-1	0	1	-1	1	-1	-4
E	1	-1	-1	1	0	1	-1	0	1	-1	1	1	2

Table 5-12, Calculation of S_{POLR} for Three Outages Linked to Order 80166733 after changing the agreement Threshold for each time Feature (Th_{Fe})

This section has illustrated the effect that varying Th_{Fe} has on an individual POLR and the necessity of optimising these four thresholds[50].

5.3 Conclusions to the Examples of Health History Enrichment

This chapter has illustrated the PEOHH by linking two WOs to outages creating POLRs and has shown the effect on individual POLRs of varying weights and thresholds. Chapter seven will illustrate the effect that varying the weights and thresholds has on the EHH.

[49] The agreement Threshold for a time Feature (Th_{Fe}) was presented in section 4.4.1.
[50] Th_{Fe} will be optimised in section 7.6.

6 Quantifying Uncertainty

Any statistic is only meaningful if its uncertainty is understood. The small size of the sample "Gold Standard" Set of Linked Records (GSSLR)[51] creates uncertainty. Understanding this uncertainty informs maintenance decision making which can improve productivity. Investigating uncertainty has enabled this research to present the uncertainty of its conclusions.

As was described in section 4.3, the GSSLR was created by manually matching a random sample from the corrective Work Order (WO) data to outages, to avoid systematic sampling errors. Random variations in a population can introduce differences between its statistical characteristics and those of a small sample from it. This chapter will review techniques for quantifying such uncertainty.

Section 6.1 will review techniques for estimating the uncertainty of the extent to which a binomial proportion[52] should be representative of the population from which it has been sampled. These techniques will be used in chapters seven and eight for estimating the uncertainty of the extent to which the *PPV* of the GSSLR should be representative of the *PPV* of the EHH and in chapter eight for estimating the uncertainty of the extent to which a specific measure of the richness of the EHH should be representative of the that property of the GSSLR.

Section 6.2 will consider the estimation of the uncertainty of a difference between two such uncertain estimates. These techniques will be used in chapter seven to estimate the uncertainty of a difference between two estimates of the *PPV* of the EHH and in chapter eight to estimate the uncertainty of a difference between two estimates of a measure of the richness of the EHH; of how much the health history has been enriched by.

Section 6.3 will investigate the probability that one such an uncertain estimate is greater than another. These techniques will be used in chapter seven to estimate the probability that a change to a parameter would increase the *PPV* of the EHH and in chapter eight to estimate the probability that the Process for the Enrichment of wind turbine Health History (PEOHH) developed in this research does enrich the health history. Section 6.4 will summarise this chapter.

---

[51] Section 4.3.2 presented the method for the validation of the techniques for health history enrichment used in this thesis. The size of the GSSLR, 29 WOs, was constrained by the amount of expert time that was available.

[52] The binomial proportion was defined in section 3.2.2. To recap, the binomial distribution describes the behaviour of a count variable for a fixed number of observations where each observation is independent, where each observation has one of two possible outcomes and where the probability of each outcome is the same for each observation. The binomial proportion is the number of successes divided by the number of trials.

6.1 Interval Estimation for a Binomial Proportion

Section 3.2.2 discussed the Positive Predictive Value (*PPV*):

$$PPV = \frac{TP}{TP + FP} \tag{2.8}$$

> "This thesis will compare a sample from the EHH to the GSSLR. It will calculate the *PPV*, referred to as the *PPV* of the GSSLR or just as the *PPV*. A point estimate is a single value estimate of an unknown population parameter. The *PPV* of the GSSLR is a point estimate of the *PPV* of the EHH.
>
> The binomial distribution describes the behaviour of a count variable for a fixed number of observations where each observation is independent, where each observation has one of two possible outcomes and where the probability of each outcome is the same for each observation. The binomial proportion is the number of successes divided by the number of trials. The *PPV* of the GSSLR will be modelled as a binomial proportion."

This section will compare five techniques for estimation of the uncertainty of a binomial proportion, referred to as intervals.

Section 4.3.2 presented the method for the validation of the techniques for health history enrichment used in this thesis. Please recall from the chapter that the size of the GSSLR, 29 WOs, was constrained by the amount of expert time that was available.

A standard technique to get a point estimate of any measure (p) of a property of a population is to apply the same measure ($\hat{p}$) to a representative sample from that population. $\hat{p}$ corresponds to the *PPV* of the GSSLR and p corresponds to the *PPV* of the EHH.

A Confidence Interval (CI) is a range of values for an unknown parameter with a specified, nominal probability that it contains the feature of interest. This probability is known as the Confidence Level (*CL*). For example, a 95% CI for p predicts with 95% *CL* that p lies within the CI. This section will construct CIs for p.[53]

The coverage of a CI is a measure of the quality of a technique for constructing CIs. It will be defined in section 6.1.2, after the standard technique for constructing CIs for a binomial proportion has been presented. That sequence for the introduction of concepts in this section enables it to illustrate

[53] To identify the confidence of the correctness of the EHH, section 8.2 will use a method that will be selected in this section to construct CIs for the *PPV* of the EHH.

coverage with examples from a CI. The standard technique for a binomial proportion is referred to in this thesis as the Wald interval (Wald, 1943). It has surprisingly poor coverage and so sections 6.1.3 to 6.1.6 will go on to present a variety of alternative techniques: Wilson (1927), Clopper and Pearson (1934), Jeffreys (1973) and Agresti and Coull (1998). All these techniques are also presented in Brown et al., 2001, who recommend that either the Wilson interval or the Jeffreys interval should be used for sample size (n) ≤ 40. This will be followed by a review of the literature on the comparison of these techniques and the presentation, for the first time, of a novel and useful method of comparison to identify the technique with the best coverage in the region of interest. The technique that this section selects will be used thereafter throughout this thesis.

Section 6.1.8 will review the "bootstrapping" technique and explain that the statistical package that this research used does not include a bootstrapping method. While python libraries do include bootstrapping methods, they do not include a tool for calculating the coverage of confidence intervals calculated using bootstrapping. The coverage of the Wilson interval is acceptable, so this research did not find it necessary to investigate the coverage of a bootstrapping interval for constructing confidence intervals of a binomial proportion.

Section 6.1.9 is the conclusion to this section.

This research performed the statistical tests that generated Figure 6-3 to Figure 6-11 using the R software suite (R Core Team, 2021) with the binom package (Dorai-Raj, 2014) through the rpy2 Python module (Cock, 2005). It benefitted from support in how to implement these techniques posted by Stack Overflow internet forum users "SergioR" (SergioR, 2020) and "horseoftheyear" (horseoftheyear, 2020).

6.1.1 Wald Interval

Laplace, 1812 and Wald, 1943, both present what is referred to in this thesis as the Wald interval but is alternatively known as the normal approximation interval. It relies on the assumption that the binomial distribution can be approximated by the Gaussian without any modifications or corrections. The Gaussian ($\phi(z)$), otherwise known as the probability density function of the standard normal distribution, is defined by equation 6.1, where z is a normalised random variable.

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \tag{6.1}$$

Significance (*S*) is 1 – *CL*, for example for a 95% CI, *S* = 0.05. The probit function (Φ^{-1}) is the inverse[54] of the cumulative Gaussian distribution. Equation 6.2 gives the $1 - {}^{S}/_{2}$ quantile of the probit function (κ).

$$\kappa = \Phi^{-1}\left(1 - {}^{S}/_{2}\right) \tag{6.2}$$

That is:

$$\kappa = \Phi^{-1}\left(\frac{CL + 1}{2}\right) \tag{6.3}$$

The Wald interval is defined by equation 6.4 and is a function of the sample size (*n*) and the sample proportion ($\hat{p}$)[55].

$$CI_{Wald}(p) = \hat{p} \pm \kappa \sqrt{\frac{\hat{p}.(1 - \hat{p})}{n}} \tag{6.4}$$

Brown et al., 2001, report that the Wald interval was at the time of writing the standard technique found in textbooks and that its properties, described in the following pages, are 'erratically poor'. Consider for example the extreme case where a sample of one from a population is successful so that *n*=1 and the number of True Positives in the sample (*TP*)=1. Figure 6-1(a) presents upper and lower limits of the CI for this case predicted using the Wald interval against p. In this case the Wald interval predicts the 95% CI as (1, 1). The figure appears to be blank because all the estimates of both the upper and lower bounds of the CI are one which means that the interval predicts with all values of confidence that the proportion of classified matches that are correctly identified as such in the whole population lies between one and one. This is incorrect as one positive match does not prove that each item in the population would also be a positive match. It is not performing appropriately in this range.

As a less extreme example, consider the case where, of a sample of 20 from a population, 11 are successful so that *n*=20 and *TP*=11. Figure 6-1(b) presents upper and lower limits of the CI for this

---

[54] The inverse cumulative distribution function transform, also known as the quantile function or percent point function, specifies the value of the random variable such that the probability of the variable being less than or equal to that value equals the probability *u,* where ($0 \leq u \leq 1$).

[55] The sample proportion ($\hat{p}$) was discussed in the introduction to section 6.1.

case predicted using the Wald interval against *p*. In this case the interval predicts the 95% CI as (0.332, 0.768) and zero confidence that p is exactly 11/20. *p* could be 11/20 but with a large population it is unlikely to be exactly that and this assumption is not problematic in practice. Problematically, the interval predicts the 100% CI as (-inf, inf), whereas p must, by definition, be between zero and 1.

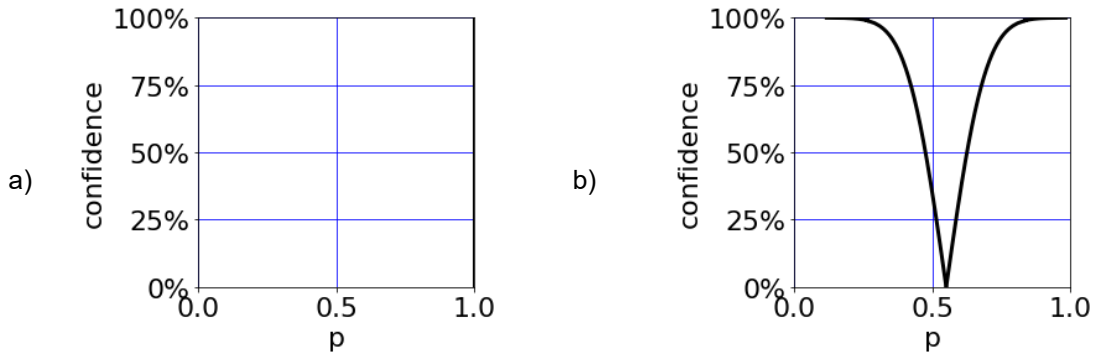


Figure 6-1, Confidence Level against Upper and Lower Limits of CI(p), Predicted using the Wald Interval, for (a) n=1, TP=1 and (b) n=20, TP=11

Figure 6-2 presents upper and lower limits of the 95% CI (*p*) predicted using the Wald interval against the number of samples (*n*) for cases where (a) half of the samples are correct ($\hat{p}$ = 1/2) and (b) all but one of the samples are correct ($\hat{p}$ = (*n*-1) / *n*). It also shows $\hat{p}$, the point estimate of *p*. Both examples show the CI narrowing down on a smaller range of *p* as *n* increases, which is intuitively credible.

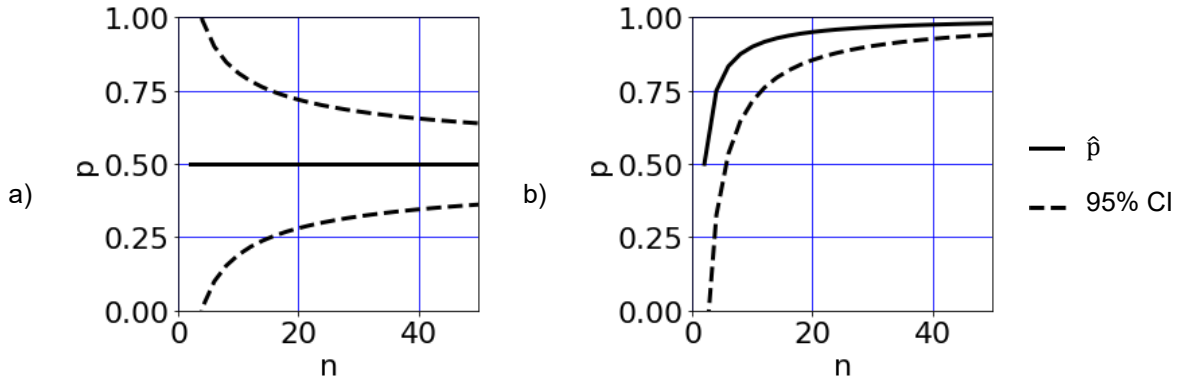


Figure 6-2, $\hat{p}$ and 95% CI of p, Predicted using the Wald Interval, against n for (a) $\hat{p}$ = 1/2 and (b) $\hat{p}$ = (n-1) / n

6.1.2 Length and Coverage

The length of an interval ($Length_{Interval}$) is defined by equation 6.5 and is the difference between the Upper limit of the interval ($U_{Interval}$) and its Lower limit ($L_{Interval}$).

$$Length_{Interval}(TP, n) = U_{Interval} - L_{Interval} \tag{6.5}$$

For example, in the case where, of a sample of 20 from a population, 11 are a positive match; *n*=20, *TP*=11. The Wald interval predicts the 95% CI as (0.332, 0.768) and has a length of 0.436. Where other considerations are equivalent, a shorter interval is considered more informative or useful because it expresses more certainty about *p*.

To recap, a Confidence Interval (CI) is a region that contains a feature of interest with a specified, nominal probability, the Confidence Level (*CL*). The probability that the CI contains this feature of interest is called the Coverage Probability (*CP*). *CP* would ideally equal CL, so the difference between them will be used as a measure of the quality of techniques used for the construction of CIs.

Figure 6-3 to Figure 6-12 show features of the binomial distribution that are not features of any data and that are not intuitive. Brown et al., 2001, describe such features: "An interesting phenomenon for the standard interval is that the actual coverage probability of the confidence interval contains nonnegligible oscillation as both *p* and *n* vary."

Some examples of this phenomenon will now be presented. Figure 6-3 shows *CP* against *CL* predicted using the Wald interval for *p*=0.5, (a) *n*=100 and (b) *n*=10. The coverage would ideally equal the confidence level but both examples show errors where the result deviates from the x=y line in a step pattern. This error is caused by the underlying structure of the binomial distribution (Brown et al., 2001). The figures show that increasing *n* or *CL* tends to reduce the size of the steps. Other CIs, presented in the following pages, have better coverage properties.

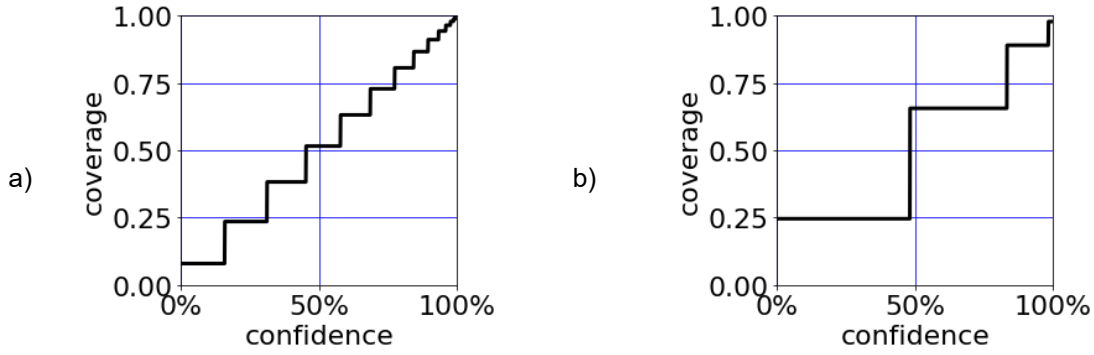


Figure 6-3, Coverage against Confidence Level, Predicted using the Wald Interval, for p=0.5, (a) n=100, and (b) n=10

Figure 6-4 shows coverage of the 95% CI predicted using the Wald interval (a) against *p* for *n* = 25, (b) against *n* for *p* = 0.99. While coverage would ideally equal the *CL*, 95%, it instead shows oscillation in coverage both with *p* and with *n* and extremely bad coverage for low *n* or small or large *p*. Figure 6-4 is a reproduction of a figure in Brown et al., 2001.

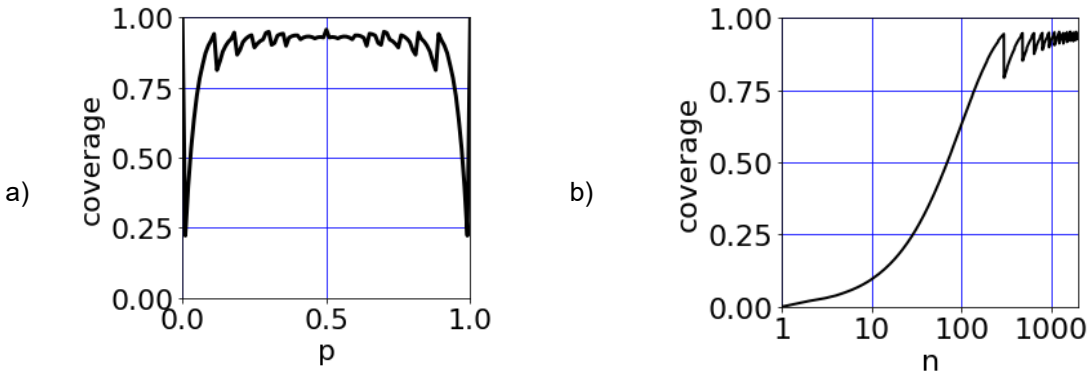


Figure 6-4, Coverage of the 95% CI, Predicted using the Wald Interval, (a) against p for n = 25, (b) against n for p = 0.99

The following subsections will use the same example cases as in Figure 6-4 to illustrate that other techniques for constructing CIs for a binomial proportion have better coverage properties. Section 6.1.7 will compare the coverage and length of the intervals in the region that is of interest to this thesis.

6.1.3 Wilson Interval

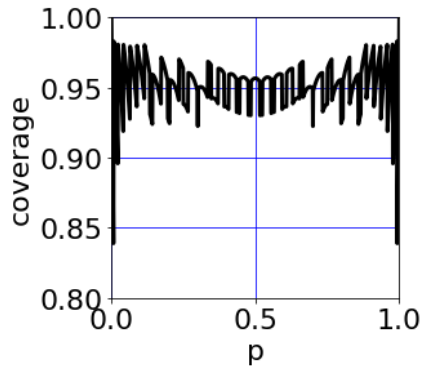
The Wilson Interval (Wilson, 1927) is a modification of the Wald interval, adjusted to improve coverage. It is a function of the sample size (*n*), the sample proportion ($\hat{p}$) and of [56]. The derivation of the interval is not included in this thesis, which is only concerned with its coverage and length.

$$CI_{Wilson}(p) = \frac{TP + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa . n^{1/2}}{n + \kappa^2}\left(\frac{\hat{p}(1-\hat{p}) + \kappa^2}{4n}\right)^{1/2} \quad (6.6)$$

Figure 6-5 shows coverage of the 95% CI predicted using the Wilson interval (a) against p for *n* = 25, (b) against *n* for *p* = 0.99. While coverage would ideally equal the *CL*, 95%, it instead shows oscillation in coverage both with *p* and with *n*. The ordinate is expanded to show the oscillation. In comparison with Figure 6-4 it shows better coverage. Section 6.1.7 will compare the coverage and length of the intervals in the region that is of interest to this thesis.

[56] κ was defined by equation 6.2.

a)



b)

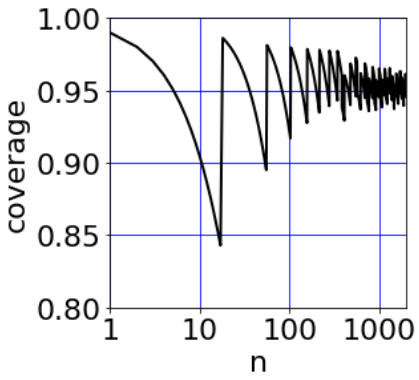


Figure 6-5, Coverage of the 95% CI, Predicted using the Wilson Interval, a) against p for n = 25, (b) against n for p = 0.99

6.1.4 Clopper-Pearson Interval

The Clopper-Pearson interval (Clopper and Pearson, 1934) is alternatively known as the exact interval. For all values of probability and sample size, it has coverage at a minimum of *CL*, that is it never under covers, but this causes it to over cover. It is based on the exact binomial distribution whereas the Wald and Wilson intervals are based on the Gaussian.

Equation 6.7 defines the Beta distribution ($B(x;\alpha,\beta)$) where α and β are positive shape parameters, c is a normalisation constant to ensure that the total probability is 1 and $0 \leq x \leq 1$.

$$B(x;\alpha,\beta) = c.x^{\alpha-1}(1-x)^{\beta-1} \tag{6.7}$$

The Clopper-Pearson interval is a function of the sample size (*n*), the number of True Positives in the sample (*TP*) and Significance (*S*), which is 1 – *CL*. For a 95% CI, *S* = 0.05. The derivation of the interval is not included in this thesis, which is only concerned with its coverage and length.

$$CI_{Clopper-Pearson}(p) = \left[L_{Clopper-Pearson}, U_{Clopper-Pearson}\right] \tag{6.8}$$

Equation 6.9 defines its Lower limit ($L_{Clopper-Pearson}$).

$$L_{Clopper-Pearson} = B\left({}^{S}/_{2}\,;\, TP, n-TP+1\right) \tag{6.9}$$

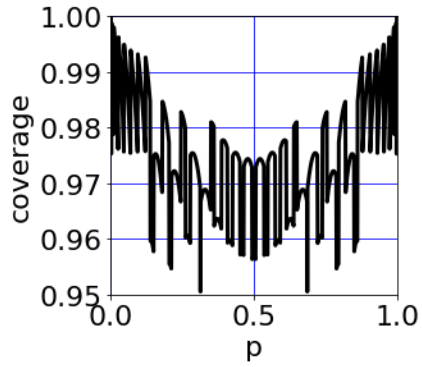
Equation 6.10 defines its Upper limit ($U_{Clopper-Pearson}$).

$$U_{Clopper-Pearson} = B\left(1-{}^{S}/_{2}\,;\, TP+1, n-TP\right) \tag{6.10}$$

Figure 6-6 shows coverage of the 95% CI predicted using the Clopper-Pearson interval (a) against *p* for *n* = 25, (b) against *n* for p = 0.99. The ordinate is expanded to show the oscillation. While coverage

would ideally equal the *CL*, 95%, it instead shows oscillation in coverage both with *p* and with *n*. In comparison with Figure 6-4 it shows better coverage. Section 6.1.7 will compare the coverage and length of the intervals in the region that is of interest to this thesis.

a)



b)

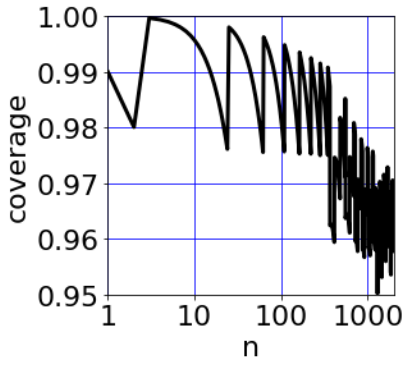


Figure 6-6, Coverage of the 95% CI, Predicted using the Clopper-Pearson Interval, a) against p for n = 25, (b) against n for p = 0.99

6.1.5 Jeffreys Interval

The Jeffreys interval (Jeffreys, 1973) is alternatively known as the Bayesian highest posterior density credible interval. It is based on Bayesian statistical inference which assumes a prior distribution of expected outcomes (Bayes, 1763). The Jeffreys interval uses a Beta[57] prior with shape parameters α = β = 0.5. The derivation of the interval is not included in this thesis, which is only concerned with its coverage and length.

$$CI_{Jeffreys}(p) = \left[L_{Jeffreys}, U_{Jeffreys}\right] \tag{6.11}$$

Equation 6.12 defines its Lower limit ($L_{Jeffreys}$). It is a function of the sample size (*n*) and the sample proportion ($\hat{p}$). It makes an exception for the case where $\hat{p}$= zero where the lower limit of *p* is known to be zero as at least one item in the population is incorrect.

$$L_{Jeffreys} = \begin{cases} 0, & \hat{p} = 0 \\ B\left(S/2\,;\, n.\hat{p} + 1/2, n.(1-\hat{p}) + 1/2\right), & \hat{p} \neq 0 \end{cases} \tag{6.12}$$

[57] The Beta distribution was defined in equation 6.7.

Equation 6.13 defines its Upper limit ($U_{Jeffreys}$). It makes an exception for the case where $\hat{p}$= 1 where the upper limit of p is known to be 1 as at least one item in the population is correct.

$$U_{Jeffreys} = \begin{cases} 1, & \hat{p} = 1 \\ B\left(1 - S/2\,;\; n.\hat{p} + 1/2, n.(1-\hat{p}) + 1/2\right), & \hat{p} \neq 1 \end{cases} \quad (6.13)$$

Brown et al., 2001, include a table of 95% limits of the Jeffreys interval. The research presented in this thesis validated its implementation of this interval against their table.

Figure 6-7 shows coverage of the 95% CI predicted using the Jeffreys interval (a) against *p* for *n* = 25, (b) against *n* for *p* = 0.99. The ordinate is expanded to show the oscillation. While coverage would ideally equal the *CL*, 95%, it instead shows oscillation in coverage both with *p* and with *n*. In comparison with Figure 6-4 it shows better coverage. Section 6.1.7 will compare the coverage and length of the intervals in the region that is of interest to this thesis.

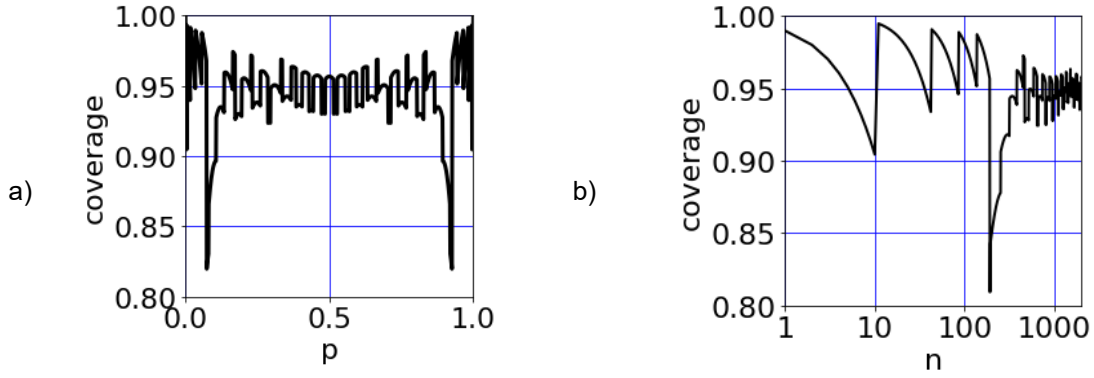


Figure 6-7, Coverage of the 95% CI, Predicted using the Jeffreys Interval, a) against p for n = 25, (b) against n for p = 0.99

6.1.6 Agresti-Coull Interval

The Agresti-Coull interval (Agresti and Coull, 1998) is a simple modification to the Wald interval. It adds two to the number of True Positives in the sample (*TP*) and four to the sample size (*n*). Adding these artificial observations has the effect of pulling the distribution of *p* towards 0.5, where the coverage of the Wald interval is closer to *CL*.

Figure 6-8 shows coverage of the 95% CI predicted using the Agresti-Coull interval (a) against *p* for *n* = 25, (b) against *n* for *p* = 0.99. The ordinate is expanded to show the oscillation. While coverage would ideally equal the *CL*, 95%, it instead shows oscillation in coverage both with *p* and with *n*. In comparison with Figure 6-4 it shows better coverage. Section 6.1.7 will compare the coverage and length of the intervals in the region that is of interest to this thesis.

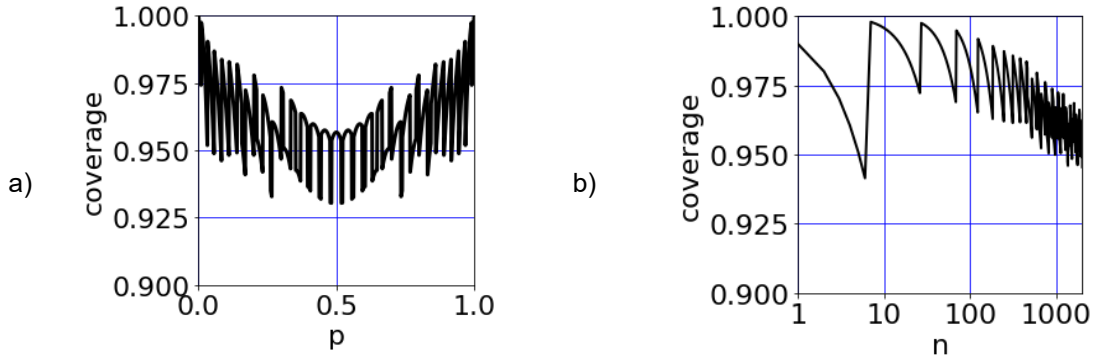


Figure 6-8, Coverage of the 95% CI, Predicted using the Agresti-Coull Interval, a) against p for n = 25, (b) against n for p = 0.99

6.1.7 Interval Selection Process

This section will compare a set of standard methods for CI estimation. It will use results from chapter 7, the sample size (*n*) and the number of true positives (*TP*), to select a technique for constructing CIs for a binomial proportion to be used throughout this thesis. Table 6-1 shows that the length of the confidence interval for *n*=29, *TP*=24 for each of the techniques presented in the preceding chapters are similar.

Technique	Upper Limit	Lower Limit	Length
Wald	0.690	0.965	0.275
Clopper-Pearson	0.642	0.942	0.299
Agresti-Coull	0.650	0.929	0.279
Wilson	0.655	0.924	0.270
Jeffreys	0.663	0.931	0.268

Table 6-1, Estimates of the 95% CI for n=29, TP=24

Several researchers have compared techniques for estimating CIs on binomial population proportions.

Cameron, 2011, reviews three techniques for estimating CIs on binomial population proportions. He finds that the Clopper-Pearson interval consistently provides a mean level of coverage close to the nominal level, even for small sample sizes. Dean and Pagano, 2015, compare techniques by calculating the mean coverage in three ranges of *p*. The mean coverage is not an appropriate indicator of the quality of coverage, as over-coverage in one part of the range would offset under-coverage in another part.

This thesis presents Coverage Error (*CE*), defined by equation 6.14. It is the difference between the nominal coverage of a CI or Confidence Level (*CL*), typically 95%, and the interval's true coverage. The CE will be used to assess the risk of poor coverage.

$$CE = coverage - CL \tag{6.14}$$

This section presents a Process for predicting the Distribution of the *CE* (PDCE) given the size of the sample (*n*) and number of True Positives in the sample (*TP*). While previous approaches have used the average coverage, the PDCE uses the distribution of coverage to compare techniques for constructing CIs for a binomial proportion in the relevant region. This is an improved approach because it reduces the risk of poor coverage. The PDCE comprises the steps listed in Figure 6-9.

PDCE step 1 uses the Jeffreys' interval. It could equally well use any interval that has acceptable coverage.

PDCE step 2 makes the simplifying assumption that *p* is equally likely to lie anywhere within the 95% CI of *p*. Further work could adjust the PDCE to put more weight on more likely values of *p*. This research tried increasing the number of values and found that it had no discernible effect on the result and so concluded that 30 is an appropriate number of values.

1:	Use the Jeffreys' interval to calculate the 95%CI of *p*
2:	Calculate an equally spaced range of 30 values of *p* that lie within this CI
3:	Calculate the coverage for the given value of *n* and each of these 30 values of *p*
4:	Use equation 6.14 to calculate the CE for each of these 30 values of *p*
5:	Present the distribution of this list of CE values

Figure 6-9, Process for Predicting the Distribution of the CE (PDCE)

Consider for example the case where the sample size is 10 and that they are all a positive match; n=10, TP=10. In this case, the PDCE uses the Jeffreys' interval to predict the 95% CI for p of [0.783, 1]. Figure 6-10(a) presents the *coverage* for this case where n=10 against p. PDCE *s*tep 4 uses equation 6.14 to calculate the *CE* for each value of p that is within the 95% CI (Figure 6-10(b).

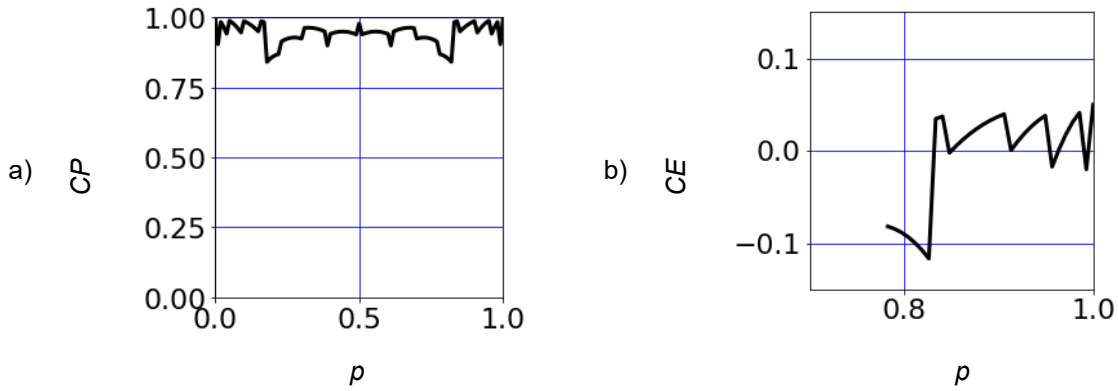
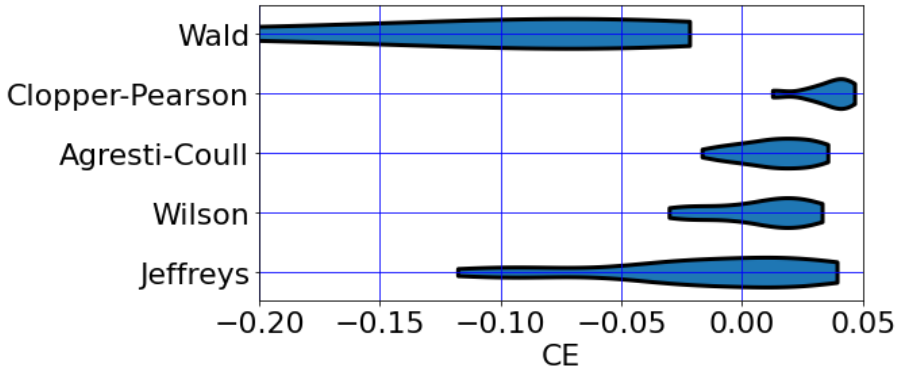


Figure 6-10 (a) Coverage Probability (CP) against p, (b) Coverage Error (CE) against p, of the 95% CI for n=10, TP=10

Finally, the PDCE compares the distribution of this list of CE values. This thesis uses the PDCE to compare techniques for constructing CIs for a binomial proportion in the relevant region. A kernel density plot is a smooth curve estimating the probability density function of a continuous variable. Figure 6-11 presents kernel density against Coverage Error (*CE*) for n=29, TP=24 of a selection of techniques for constructing CIs for a binomial proportion of the (a) 95% CI (b) 99% CI.

Figure 6-11 shows the results for the PDCE. Please recall that figures 6.3 to 6.12 show features of the binomial distribution that are not features of any data and that are not intuitive. If a technique has a low risk of poor coverage (that is that the technique is relatively appropriate) then in figure 6.11 *CE* will be clustered around zero. Figure 6-11 shows that the Wald interval has very bad coverage in this region, the Clopper-Pearson interval has too high coverage in this region and the other intervals have acceptable coverage in this region. Figure 6-11(b) shows that the Wilson interval has *CE* clustered the most closely around zero, giving it the best coverage on the 99% CI. This indicates that it has even tails on the 95% CI; that is that the (small) probability of over-coverage is close to the (small) probability of under-coverage for this interval in this region. Figure 6-11 shows that the Wald and Clopper-Pearson intervals are both less appropriate than the Agresti-Coull, Wilson or Jeffreys intervals. Figure 6-11(a) and (b) show that the Wilson interval has *CE* clustered the most closely about zero which means that it has the least risk of poor coverage in the region of interest and so this thesis will use that technique for constructing confidence intervals.

a)



b)

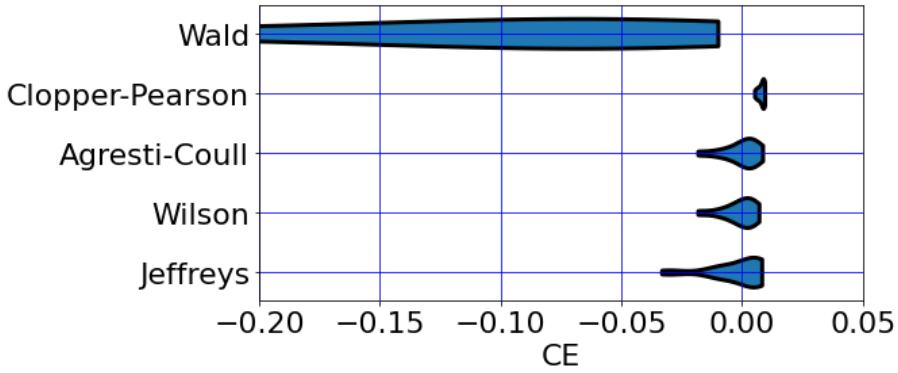


Figure 6-11, Kernel Density against Coverage Error (CE) for n=29, TP=24 of a Selection of Techniques for Constructing Confidence Intervals (CI) of a Binomial Proportion for (a) Confidence Level (CL) = 95%, (b) CL=99%

6.1.8 Bootstrapping

Bootstrapping (Efron, 1979) is a statistical procedure that identifies subsamples from the original sample from a population. Each element of the sample can be included repeatedly in the same subsample. To average out random sampling errors, a large set of subsamples is used. This is known as random sampling with replacement. Bootstrap methods can be used for constructing confidence intervals of a binomial proportion (Mantalos and Zografos, 2008).

This research performed the statistical tests that generated Figure 6-3 to Figure 6-11 using the R software suite (R Core Team, 2021) with the binom package (Dorai-Raj, 2014) through the rpy2 Python module (Cock, 2005). Binom does not include a bootstrapping method, and, while python libraries including scikit-learn, scipy-stats and pandas do include bootstrapping methods, they do not include a tool for calculating the coverage of confidence intervals calculated using bootstrapping.[58] Consequently, this thesis does not investigate the use of bootstrapping for the estimation of the uncertainty of a binomial proportion, for which it considers that the coverage from the Wilson interval is acceptable. It will use bootstrapping for other applications which will be presented in sections 6.2 and 6.3; the comparison binomial proportions by estimating the uncertainty of, firstly, the ratio between them and, secondly, the probability that one proportion is greater than another.[59]

[58] Section 6.2.6 will compare the computational expense of assesing the coverage of an application that uses bootstrapping against using other, less computationally expensive, techniques.

[59] The start of this chapter explained why this research required these methods.

.

6.1.9 Conclusion to Interval Estimation for a Binomial Proportion

This section has reviewed techniques for constructing confidence intervals for a binomial proportion. Brown et al., 2001, demonstrated that there is difficulty understanding a class of uncertainties that are caused by the underlying structure of the binomial distribution and this chapter has reproduced their result. It has demonstrated how to assess the risk of poor coverage and shown that this risk can be high when using some popular techniques. It has investigated how to minimise this risk by selecting an appropriate technique.

This chapter has presented a new and useful approach for deciding which technique to select for constructing confidence intervals for a binomial proportion. Previous approaches selected a technique by its mean coverage, but this new approach instead selects a technique by the distribution of its coverage. This research reviewed the literature and did not find this approach, so it can be concluded that the approach is novel. This approach is useful in that it minimises the risk of poor coverage more effectively than using the mean coverage.

To minimises the risk of poor coverage, chapters 8 and 9 will use the Wilson interval for constructing confidence intervals for binomial proportions. This will enable those chapters to quantify their uncertainty as to the extent to which measures of a random sample from a population can be used to draw inferences about that population.

6.2 Interval Estimation for the Ratio of Two Proportions

The previous section reviewed techniques for estimating the uncertainty as to the extent to which a binomial proportion is representative of the population from which it has been sampled. This section will compare two such uncertain estimates. They will be considered as two cases, referred to in this chapter as case A and case B, applied to the same sample.

This section will consider the comparison of the enriched health history to the unenriched health history using the GSSLR. This statistical situation occurs repeatedly in this thesis and one of its occurrences will be used as an example, presented in section 6.2.1. Sections 6.2.2 to 6.2.5 present three techniques for interval estimation for a change in a proportion. These techniques make different assumptions about how these proportions are related to each other. Section 6.2.6 compares them and section 6.2.7 concludes this section.

6.2.1 Example of Comparing Proportions

This section uses the following example to illustrate techniques for estimating the uncertainty of a change in a proportion and to select a technique. This thesis will use the selected technique in its analysis of other results as well.

To investigate whether the Process for the Enrichment of wind turbine Health History (PEOHH) developed in this research does meet its object, that is to enrich health history, chapter 8 will address RQ3:

RQ3 How can the richness of historical data on wind turbine health be measured?

To quantify how much enrichment is achieved by the PEOHH, chapter 8 will compare the Enriched Health History (EHH) to the unenriched health history. It will define richness and then quantify richness using a variety of measures.

Section 1.2.4 reviewed the literature on troubleshooting as an aspect of the maintenance of offshore wind turbines. Please recall from that section that troubleshooting is the activity of repairing a faulty OWT by the replacement of components. Section 8.5 will consider the application of wind turbine health history to troubleshooting. This chapter considers the statistical theory that will be required to understand uncertainty in section 8.5 and in other applications as well, such as to understand the uncertainty of a change in the quality of record linkage. As a measure of the richness of historical data on wind turbine health, it will present the concept of the Proportion of POLRs in a set where the health history would recommend all the Material needed, (*PM*), a measure of the quality of simulated troubleshooting guides. *PM* is defined by equation 6.15 using the number of POLRs in the EHH where the health history Would Recommend all the required parts (WR_{EHH}) and the number of POLRs in the

EHH where the Health History would Not Recommend all the required parts (NR_{EHH}), making various assumptions that will be detailed in section 8.5.

$$PM = \frac{WR_{EHH}}{WR_{EHH} + NR_{EHH}} \tag{6.15}$$

PM is estimated using the *PM* of the GSSLR ($\widehat{PM}$) and is consequently expressed as a binomial proportion. The small size of the GSSLR[60] creates uncertainty about the true value of *PM* prior to and after enrichment. Rather than considering the uncertainty of these two values of *PM* separately, as in section 6.1, this section considers the uncertainty of the increase.

Section 8.5 will define those WOs already labelled with an alarm code as the unenriched health history. These are the existing records of health history prior to enrichment. This section refers to the case prior to enrichment as 'A' and to that post enrichment as 'B'. Since knowledge of both cases is based on the GSSLR they are not independent of each other. Most statistical literature describes inferences from independent datasets and so this section will address the unusual situation where independence cannot be assumed.

This example considers the Increase in *PM* between cases A and B (I_{AB}). I_{AB} is important to this research because, in this example, it estimates the extent to which the health history has been enriched.[61] Equation 6.16 defines I_{AB} using the Proportion of POLRs where the health history would recommend all the Material needed for case A (PM_A) and the same Proportion for case B (PM_B).

$$I_{AB} = \frac{PM_B}{PM_A} - 1 \tag{6.16}$$

PM_A and PM_B are unknown and so this chapter will use estimates of them.

This chapter will consider the uncertainty of the estimate of I_{AB} caused by the small number of samples. Equation 6.17 identifies a point estimate of I_{AB}, ($\widehat{I_{AB}}$) using the number of POLRs in the GSSLR where the health history Would Recommend all the required parts for case A (WR_{GSSLR_A}) and the same quantity for case B (WR_{GSSLR_B}), again making various assumptions that will be detailed in

[60] Section 4.3.2 presented the method for the validation of the techniques for health history enrichment used in this thesis. The size of the GSSLR, 29 WOs, was constrained by the amount of expert time that was available.

[61] Chapter 8 will further discuss and interpret I_{AB}.

section 8.5. In this example, case A has a WR_{GSSLR} of 11 but case B has a WR_{GSSLR} of 18. *PM* has increased by 64%.

$$\widehat{I_{AB}} = \frac{WR_{GSSLR_B}}{WR_{GSSLR_A}} - 1 = \frac{18}{11} - 1 = 64\% \tag{6.17}$$

This example will also use the following experimental results from Chapter 8:

- For 11 Pairs Of Linked Records (POLR) in the GSSLR, for both case A and case B, the health history would recommend all the required parts.
- For no POLRs in the GSSLR, for case A the health history would recommend all the required parts but for case B it would not.
- For 7 POLRs in the GSSLR, for case A the health history would not recommend all the required parts but for case B it would.
- For 10 POLRs in the GSSLR, for neither case A nor case B, the health history would recommend all the required parts.

Table 6-2 presents this example as a table.

		Case B	
		Would Recommend	Would Not Recommend
Case A	Would Recommend	11	0
	Would Not Recommend	7	10

Table 6-2, Example of a Change in a Proportion

The following sections will present techniques for constructing Confidence Intervals (CI) for I_{AB}.

6.2.2 Assuming Independence

A statistical theory based on independent samples is not applicable to a case where the samples are far from independent. The vast majority of statistical research has been on systems that can usefully be modelled as independent, such as gambling and medicine. In medical research for example, comparative trials are not carried out on the same patients. If method A is tested on sample A and method B is tested on sample B, sample A is made up of different individuals from sample B and the results can often be assumed to be independent.

Section 6.2.6 will demonstrate that the assumption of independence presented in this section is inappropriate for this application of comparing record linkage techniques on the same GSSLR.

Newcombe, 1998 compares methods for the estimation of CIs for the difference between independent proportions, using examples from medicine. In appendix I, Newcombe uses the Wilson interval to construct CIs for the ratio of two independent binomial proportions. If PM_A were independent of PM_B then it would be appropriate to use the approach described by Newcombe, 1998, to construct a CI for I_{AB}.

This section presents the Technique Assuming Independence between PM_A and PM_B (TAI). It hypothesises that PM_A is not independent of PM_B and that therefore TAI will have poor coverage.

Equation 6.18 is derived from Newcombe, 1998, appendix I. For a specific Confidence Level (*CL*), L_W is the Lower limit of the Wilson CI[62] and U_W is its Upper limit.

$$CI_{TAI}\left(\frac{PM_B}{PM_A}\right) = \left[\frac{L_W}{1-L_W}, \frac{U_W}{1-U_W}\right] \tag{6.18}$$

Substituting equation 6.18 into equation 6.16, equation 6.19 estimates the uncertainty of the Increase (I_{AB}):

$$CI_{TAI}(I_{AB}) = \left[\frac{2L_W - 1}{1-L_W}, \frac{2U_W - 1}{1-U_W}\right] \tag{6.19}$$

For the case in the example described in section 6.2.1, the TAI estimates the 95% CI of I_{AB} as (-21%, 241%). Section 6.2.6 will show that this CI grossly over-covers, indicating that TAI is not an appropriate technique for this application because the results of tests conducted on the same sample are not independent of each other.

Unlike TAI, the techniques described in the following sections will both require significant computation.

[62] The Wilson interval was defined in section 6.1.3.

6.2.3 Assuming Equally Representative Samples

The previous section presented the Technique Assuming Independence between PM_A and PM_B (TAI). The opposite assumption is that the sample is just as representative of the population for case A as it is for case B. This section presents the Technique Assuming Equally representative samples (TAE). TAE approximates I_{AB} by evaluating probability distribution functions PM_A and PM_B.

Equation 6.20 defines a vector of Probabilities (*Prob*) where:

- P_{11} is the Probability that a randomly sampled POLR would recommend all the required parts for both case A and case B
- P_{01} is the Probability that it would not recommend all the required parts for case A but would recommend all the required parts for case B
- P_{10} is the Probability that it would recommend all the required parts for case A but would not recommend all the required parts for case B
- P_{00} is the Probability that it would not recommend all the required parts for both case A and case B

$$Prob = [P_{11}, P_{01}, P_{10}, P_{00}] \tag{6.20}$$

Where:

$$\sum_i \sum_j P_{ij} = 1 \tag{6.21}$$

TAE estimates P_A and P_B for a full range of points in probability space that reflect how representative the sample is of the population. This embeds the assumption that the sample is just as representative of the population for case A as it is for case B in TAEs estimation of I_{AB}.

A categorical distribution is a discrete probability distribution that describes the possible results of a random variable that can take on one of a set of possible categories, with the probability of each category separately specified. The multinomial distribution models the outcome of a set of tests where the outcome of each test is modelled by a categorical distribution. The binomial distribution is the case of the multinomial distribution where the number of categories is 2. Section 6.1 reviewed the application of the Wilson interval to binomial proportions, but it can also be used to estimate CIs for multinomial proportions (Wilson, 1927). This section will model *Prob* as a multinomial distribution.

TAE finds a CI for each of the four elements of *Prob* for any *CL* where ($0 \leq CL \leq 1$). It assumes a multinomial distribution and estimates this using the Wilson interval. The natural multidimensional extension of the Wilson interval is the same as that interval for a binomial proportion.

The Wilson interval is defined for the multinomial case by equation 6.22 and is a function of the sample size (n), the sample proportion ($\hat{p}$) and [63].

$$CI_{Wilson}(P_{ij}) = \frac{x_{ij} + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa n^{1/2}}{n + \kappa^2}\left(\frac{\widehat{p_{ij}}\left(1 - \widehat{p_{ij}}\right) + \kappa^2}{4n}\right)^{1/2} \tag{6.22}$$

Where:

$$n = \sum_i \sum_j x_{ij} \tag{6.23}$$

And:

$$\widehat{p_{ij}} = \frac{x_{ij}}{n} \tag{6.24}$$

This research investigated the effect on each element of *Prob* of varying *CL* for the case in the example described in section 6.2.1. The results, taking P_{00} as an example element, are shown in Figure 6-12. The median of P_{00} is where *CL* is zero and the figure shows that in this example it is at approximately 0.35.

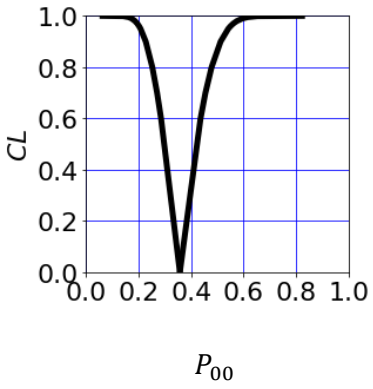


Figure 6-12, Confidence Level (CL) against the Probability that a Random Sample would Not Recommend all the Required Parts for either Case A or Case B (P_{00}) for the Example in Section 6.2.1

TAE approximates I_{AB} by evaluating probability distributions P_A and P_B. It uses the inverse Cumulative Distribution Function (CDF) transform, (also known as the quantile function or percent point function), which specifies the value of the random variable such that the probability of the variable being less than or equal to that value equals the probability u, where ($0 \leq u \leq 1$). The median is at $u = 0.5$. For the

[63] κ was described in section 6.1.1.

case in the example described in section 6.2.1, for *u* = 0.5, the inverse CDF transform returns a median of P_{00} = 0.35.

To approximate the inverse CDF transform of the Wilson interval, this section assumes that the Wilson interval is symmetric about the median such that for any *CL,* where (0 ≤ *CL* ≤ 1), the median minus the lower bound of the CI approximates the upper bound of the CI minus the median. This assumption is conceivable because Figure 6-12 does appear to be approximately symmetric about the median. This section will sensitivity test the assumption of symmetry by replacing it with an alternative assumption. This sensitivity test will show that the assumption is appropriate because replacing it does not make a significant difference. The assumption of symmetry implies equation 6.25.

$$CL \approx |2u - 1| \tag{6.25}$$

TAE takes a set of equally distributed samples from *u* with Step Size (*SS*), as defined by equation 6.26.

$$u = (SS, 2SS, 3SS \ldots 1 - SS) \tag{6.26}$$

For each value of *u*, (u_i), TAE uses equation 6.25 to identify the corresponding value of *CL*. TAE uses the Wilson interval to find the multinomial CIs for *Prob*, yielding a CI for each of $(P_{11}|u_i)$, $(P_{01}|u_i)$, $(P_{10}|u_i)$ and $(P_{00}|u_i)$. Equations 6.27 and 6.28 are implied by the definitions of PM_A and PM_B.

$$PM_A = P_{11} + P_{10} \tag{6.27}$$

$$PM_B = P_{11} + P_{01} \tag{6.28}$$

Substituting equations 6.27 and 6.28 into the definition of a CI yields equations 6.29 and 6.30. TAE uses equations 6.31 and 6.32 to construct CIs (CI_{TAE}) for $(PM_A|u_i)$ and for $(PM_B|u_i)$.

$$CI_{TAE}(PM_A|u_i) = \left[\left((L_{11}|u_i) + (L_{10}|u_i)\right), \left((H_{11}|u_i) + (H_{10}|u_i)\right)\right] \tag{6.29}$$

$$CI_{TAE}(PM_B|u_i) = \left[\left((L_{11}|u_i) + (L_{01}|u_i)\right), \left((H_{11}|u_i) + (H_{01}|u_i)\right)\right] \tag{6.30}$$

Substituting $(PM_A|u_i)$ and $(PM_B|u_i)$ into equation 6.16 yields equation 6.31, which TAE uses to construct CI_{TAE} for $(I_{AB}|u_i)$.

$$CI_{TAE}(I_{AB}|u_i) = \frac{(PM_B|u_i)}{(PM_A|u_i)} - 1 \tag{6.31}$$

For each item in the list of values of probability *(u)*, TAE now has an estimate of I_{AB}. TAE estimates the 95% CI of I_{AB} as the 2.5th percentile and the 97.5th percentile of this list of estimates of I_{AB}.

This research investigated the effect of varying *SS*[64] and he results are shown in Figure 6-13. It shows that the estimate of I_{AB} levels out below an *SS* of 10^{-4} which indicates that 10^{-4} is an appropriate value for *SS*.

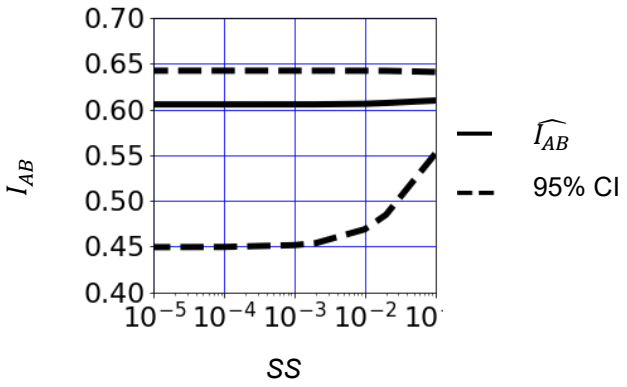


Figure 6-13, Increase (I_{AB}) and 95% Confidence Interval (CI) against Step Size (SS)

For the example described in section 6.2.1, the TAE estimates the 95% CI of I_{AB} as (45%, 64%). Section 6.2.6 will show that this CI grossly under-covers, indicating that TAE is not an appropriate technique for this application because the results of tests conducted on the same sample are not just as representative of the population for case A as they are for case B.

6.2.4 Alternative Assuming Equally Representative Samples

The previous section presented the Technique Assuming Equally representative samples (TAE). It explained that the TAE assumes that the CI of each element of *Prob* is symmetric. To sensitivity test this assumption, this section will replace it with an alternative assumption. This section presents the Alternative Technique Assuming Equally representative samples (ATAE).

Just as does TAE, ATAE also approximates I_{AB} by evaluating probability distributions PM_A and PM_B. To estimate I_{AB} for each of a list of values of probability *(u)*, ATAE models PM_A and PM_B using the skew normal distribution which is a skewed version of the normal distribution. ATAE fits this distribution to P_A and P_B using Differential Evolution (DE), which was reviewed in section 2.4.1.2.

The Cumulative Distribution Function (CDF) of a random variable *X*, evaluated at *x*, is the probability that *X* will take a value less than or equal to *x*. ATAE uses the inverse CDF transform of the skew

[64] The Step Size (*SS*) was defined by equation 6.26.

normal distribution to yield an estimate of A and of B for each of the list of values of probability *(u)*. Figure 6-14(a) shows the results for the example in section 6.2.1.

A Probability Density Function (PDF) is a function whose value at any given sample can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample. It is the differential of the CDF with respect to *x*. To visualise the distributions, this research used ATAE to construct the PDF for case A and for case B. Figure 6-14(b) shows the results for the example in section 6.2.1.

Figure 6-14 shows that most estimates of PM_B, the probability that a truly matching POLR for case B, the health history after enrichment, would be yielded from a record selected at random from the general population, are higher than most estimates of PM_A, the probability that a truly matching POLR for case A, the health history prior to enrichment, would be yielded from a record selected at random from the general population.

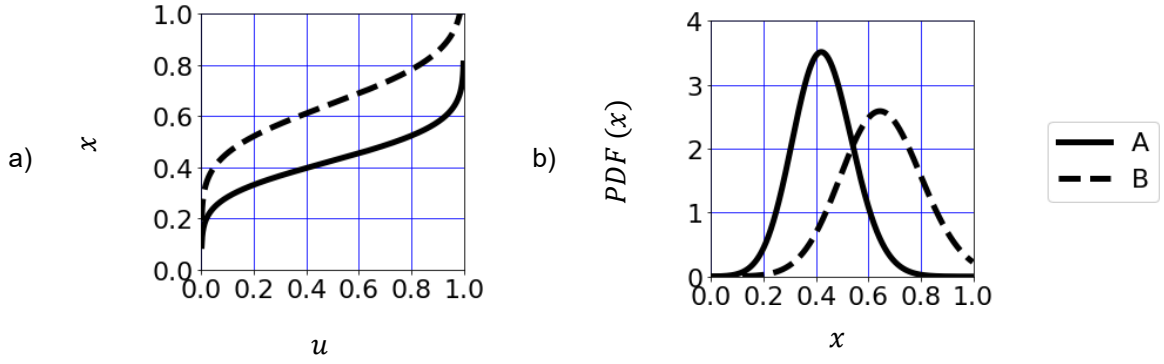


Figure 6-14, (a) Inverse CDF Transform, (b) PDF, for Case A and for Case B

ATAE identifies a list of estimates of I_{AB} using equation 6.16 to calculate I_{AB} for each pair of estimates of A and B. It estimates the 95% CI of I_{AB} as the 2.5th percentile and the 97.5th percentile of this list of estimates.

Running ATAE repeatedly yielded various estimates of the 95% CI of I_{AB}. The results are shown in Figure 6-15. This inconsistency in ATAE estimates is caused by the difficulty of fitting skew normal distributions to P_A and to P_B. The estimate from TAE is in the same region as the estimates from ATAE which indicates that the assumption of symmetry made by TAE is appropriate.

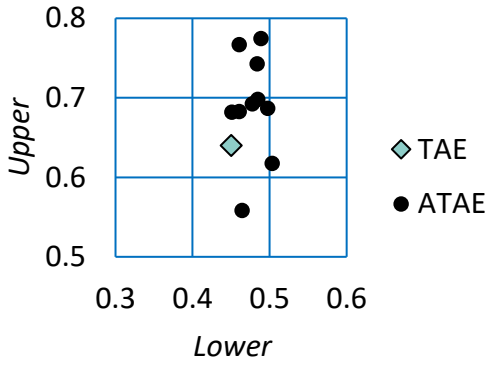


Figure 6-15, Upper and Lower Bounds of the 95% CI of I_{AB}, Estimated by the Technique Assuming Equally Representative Samples (TAE) and by the Alternative Technique Assuming Equally Representative Samples (ATAE)

The previous section explained that the TAE assumes that the CI of each element of *Prob* is symmetric. To sensitivity test this assumption, this section replaced it with an alternative assumption and showed that this yields a similar estimate of the 95% CI of I_{AB}. This showed that the assumption of symmetry made by TAE is appropriate.

6.2.5 Using Bootstrapping

Bootstrapping[65] (Efron, 1979) is a statistical procedure that identifies subsamples from the original sample from a population. Each element of the sample can be included repeatedly in the same subsample. To average out random sampling errors, a large set of subsamples is used. This is known as random sampling with replacement. The size of the subsample is known as the Bootstrap Size (*BS*). It is recommended that *BS* should be equal to the size of the original sample (*n*).

---

[65] Section 6.1.8 introduced bootstrapping and explained why this research thesis does not use it for constructing confidence intervals of a binomial proportion. This section does use bootstrapping to estimate the uncertainty of a difference between two binomial proportions and section 6.3 will use it to estimate the probability that one binomial proportion is greater than another.

Bootstrap methods can be used for constructing Confidence Intervals (CI) of a binomial proportion (Mantalos and Zografos, 2008). This section investigates the use of bootstrapping for constructing a CI for Increase (I_{AB}), defined in section 6.2.1, and presents it as the Technique Using Bootstrapping (TUB).

TUB generates a list of estimates of I_{AB}. It estimates the 95% CI of I_{AB} as the 2.5th percentile and the 97.5th percentile of this list of estimates of I_{AB}.

This research investigated the effect of varying *BS* between 1 and 100 for the case in the example described in section 6.2.1 which has sample size *n* = 28. (This is the number of WOs in the GSSLR that have materials assigned to them). Figure 6-16(a) presents I_{AB} against *BS*. It shows that TUB would estimate a wider CI if *BS* were less than 28 and a narrower CI if *BS* were more than 28. This research will use *BS* = *n*, as recommended by Efron, 1979.

This research investigated the effect of varying the number of subsamples. The results are shown in Figure 6-16(b), which shows that in this case 1000 subsamples are sufficient to average out random sampling errors.

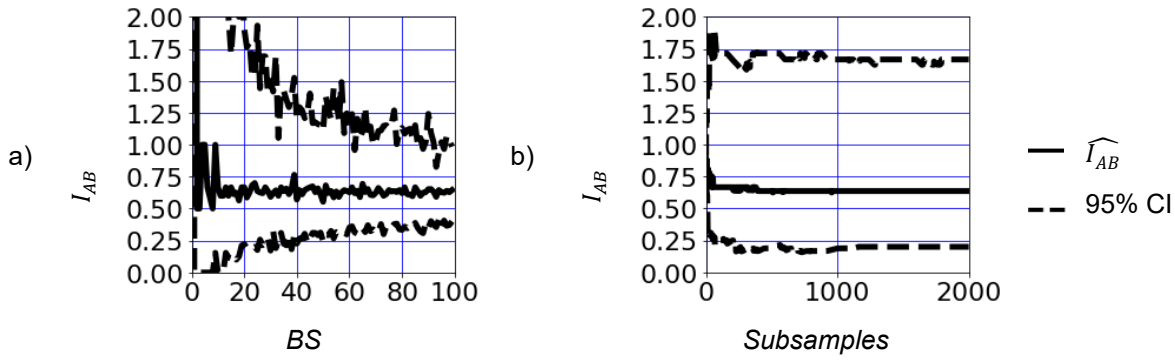


Figure 6-16, Increase (I_{AB}), (a) against Bootstrap Size (BS) for Number of Subsamples = 100, (b) against Number of Subsamples for BS = n = 28

For the case in the example described in section 6.2.1, TUB estimates the 95% CI of I_{AB} as (20%, 167%). Section 6.2.6 will show that this CI slightly under-covers, indicating that TUB is an appropriate technique for this application because the its underlying assumptions are more appropriate than those of TAI or TAE.

Bootstrapping assumes that each element of the sample is independent of the other elements. The next section will show that it has much better coverage than the techniques presented in the previous sections.

6.2.6 Coverage

Section 6.1.2 reviewed the coverage of a CI. Please recall from the section that:

> The probability that the CI contains this feature of interest is called the Coverage Probability (*CP*). *CP* would ideally equal CL and the *CP* will be used as a measure of the quality of methods used for the construction of CIs.

Sections 6.2.2 to 6.2.4 presented three techniques for interval estimation for a change in a proportion:

- The Technique Assuming Independence (TAI)
- The Technique Assuming Equally representative samples (TAE)
- The Technique Using Bootstrapping (TUB)

This section presents a Process for Comparing Techniques for interval estimation for a change in a proportion (PCT) and uses it to compare the *CP* of these three techniques. The *CP* of a CI can be estimated by simulation; however, *CP* varies with *n* and with the probability of each outcome (Brown et al., 2001).

The PCT calculates *CP* using a Python algorithm that repeats an experiment and stores the results, posted by Stack Overflow internet forum user “anky” (anky, 2019).

TAE and TUB are both computationally expensive, as detailed in Table 6-3, which makes it intractable to estimate the *CP* for many values of the probability of each outcome. As an alternative, and using a similar approach to that was used in the interval selection process that was presented in section 6.1.7, PCT concentrates on the region of interest.

Just as in the TAE, the PCT also uses the Wilson interval to find the multinomial CIs for $P_{11}, P_{01}, P_{10},$ and P_{00}. For the case in the example described in section 6.2.1, these are detailed in Table 6-3.

	Estimate	Low	High
P_{11}	0.393	0.236	0.576
P_{01}	0.000	0.000	0.121
P_{10}	0.250	0.127	0.434
P_{00}	0.357	0.207	0.542

Table 6-3, Estimate and 95% CI of the Probability of Each Outcome

Section 6.2.3 presented the TAE. Please recall equations 6.22 and 6.23 from that section:

$$Prob = [P_{11}, P_{01}, P_{10}, P_{00}] \tag{6.22}$$

Where:

$$\sum Prob = 1 \tag{6.23}$$

The PCT estimates *CP* by simulating sets of values of *Prob* that are in the region of interest (as detailed in Table 6-3 for the case in the example described in section 6.2.1). It identifies the range of likely values of P_{01} (the Probability that a random sample from the population NR for case A but WR for case B) and selects four values across this range, ranging from zero to 0.12. Keeping P_{11} and P_{00} constant, it varies P_{10} to maintain equation 6.23. For the case in the example described in section 6.2.1, this process yields four sets of values of *Prob*:

- *Prob* = [0.3, 0.0, 0.3, 0.4]
- *Prob* = [0.3, 0.04, 0.26, 0.4]
- *Prob* = [0.3, 0.08, 0.22, 0.4]
- *Prob* = [0.3, 0.12, 0.18, 0.4]

These four *Prob* vectors will be used to identify the coverage of each of the three techniques for constructing CIs. Such a sensitivity study of the effect of variation in P_{01} in the region of interest would ideally be combined with similar studies varying P_{11}, P_{10}, and P_{00} but this would be intractable as TAE and TUB are both computationally expensive when repeated 10,000 times to yield a useful estimate of coverage for each set of values of *Prob*, as detailed in Table 6-4. The table shows that TAI is computationally inexpensive and so could be repeated 100,000 times for each set of values of *Prob*, yielding a more accurate estimate of the coverage.

The computational expense of TAE and of TUB is only a concern when repeating them to estimate coverage, not when using them a single time. The computation times detailed in Table 6-4 are for a standard laptop PC processor[66]. It is important that the computation times detailed are for 10,000 runs, 100,000 in the case of TAI, and this is the reason for the lengthy computation time. Computation time is a difficulty for the problem of measuring the coverage of a statistical technique, not for the real world application the health history enrichment techniques presented in this thesis[67]. Further work optimising the bootstrapping algorithm or running it on a faster computer would improve its computation time.

	TAI	TAE	Bootstrap
BS			28
Subsamples			1,000
Reps	100,000	10,000	10,000
run time	41 seconds	32 hours	41 hours

Table 6-4, Comparing the Computation Time of Techniques for Interval Estimation

Substituting equations 6.27 and 6.28 into equation 6.16 yields equation 6.24. For each of the sets of values of *Prob,* the PCT uses equation 6.34 to identify the 'true' value of I_{AB} that would be yielded from an infinite sample from a population with the given set of values of *Prob.*

$$I_{AB} = \frac{P_{11} + P_{01}}{P_{11} + P_{10}} - 1 \tag{6.24}$$

To generate sample data, the PCT takes random samples from a multinomial distribution generated from the given set of values of *Prob* with sample size *n.* It then applies the technique on test to estimate the 95% CI of I_{AB}. If the estimate of the 95% CI of I_{AB} includes the 'true' value of I_{AB} then it records a positive outcome and it otherwise records a negative outcome. It repeats this many times (*Reps*). The PCT estimates *CP* as the proportion of positive outcomes.

Brown et al., 2001, demonstrated that "For larger *n*, the Wilson, the Jeffreys and the Agresti-Coull intervals are all comparable", so, when *n*=40, any of these three intervals can be used with acceptable coverage. The PCT estimates the 95% CI for the coverage of the 95% CI, again using the Wilson interval.

[66] Section 4.5 detailed the processor.

[67] Section 4.5 detailed the computation times for the PEOHH.

Figure 6-17 shows CP against P_{10} for each of the three techniques. The coverage of TAI is around 99.88%, grossly over the target 95%. (4.88% may appear to be a small error but it is in fact a large error at this end of the probability range. The odds is the probability that an event will occur divided by the probability that the event will not occur and is an alternative format for describing probability. 95% has the equivalent odds of 19 while 99.88% has the much greater equivalent odds of 832.) The coverage of TAE is around 32%, grossly under the target 95%. The coverage of TUB is around 93%, slightly under the target 95%. It is clear that, of the techniques presented, only TUB gives acceptable performance and consequently this thesis will use TUB for interval estimation for a change in a proportion.

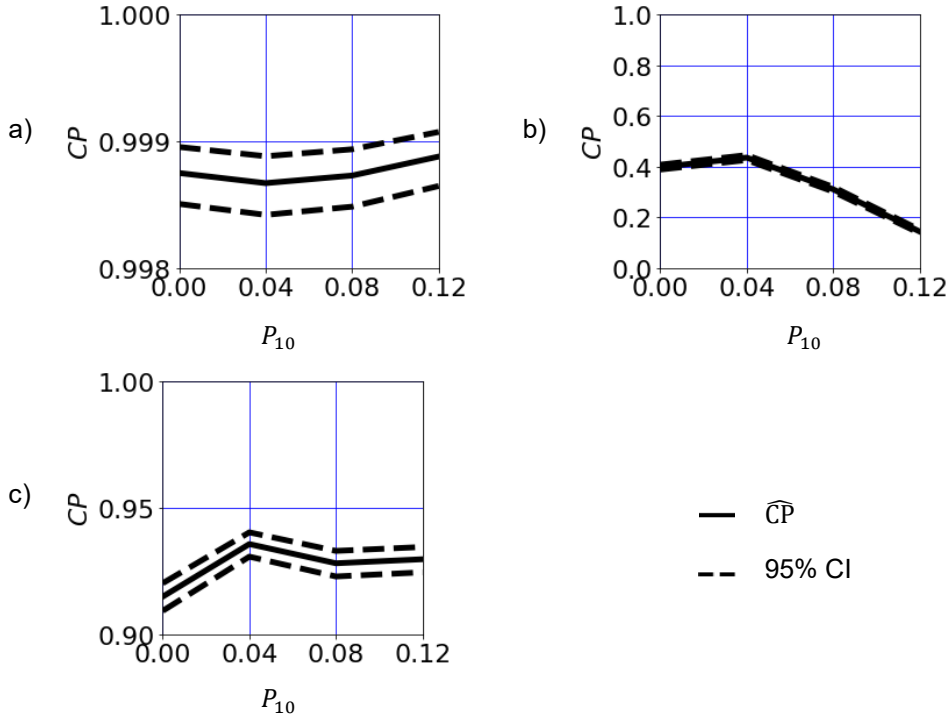


Figure 6-17, Coverage Probability (CP) against the Probability that a Random Sample from the Population WR for case A but NR case B (P_{10}) for the:

(a) Technique Assuming Independence (TAI)

(b) Technique Assuming Equally Representative Samples (TAE)

(c) Technique Using Bootstrapping (TUB)

6.2.7 Conclusion to Interval Estimation for a Change in a Proportion

This section presented three techniques for interval estimation for a change in a proportion:

- The Technique Assuming Independence (TAI)
- The Technique Assuming Equally representative samples (TAE)
- The Technique Using Bootstrapping (TUB)

It showed that TAI and TAE yield very poor coverage. This is an important result because one might intuitively make the assumptions from TAI when assessing the uncertainty of a difference in an uncertain estimate.

Chapters 7 and 8 will use TUB for constructing CIs for changes in proportions. This will enable the chapters to quantify their uncertainty as to the extent to which changes in measures of a random sample from a population can be used to draw inferences about that population.

6.3 The Comparison of Proportions

Section 6.2.5 presented the Technique Using Bootstrapping (TUB) for interval estimation for comparing proportions. This section will use TUB to investigate the probability that one proportion is greater than another.

Table 6-5 presents two examples of changes in a proportion. Making various assumptions that will be detailed in section 8.5, these are example results of the number of POLRs in the GSSLR where a simulated troubleshooting guide would recommend all the required parts and the number of POLRs in the EHH where that same guide would not recommend all the required parts. As in the previous section, this section also refers to the case prior to enrichment as 'A' and to that post enrichment as 'B'.

Example I is taken from the results that will be presented in section 8.5 while example II is included to illustrate the inferences that would be drawn from different results.

Example I				Example II			
		Case B				Case B	
		Would Recommend	Would Not Recommend			Would Recommend	Would Not Recommend
Case A	Would Recommend	11	0	Case A	Would Recommend	8	1
	Would Not Recommend	7	10		Would Not Recommend	8	11

Table 6-5, Two Examples of Changes in a Proportion

Please recall equation 6.16 which defines the Increase (I_{AB}) using the Proportion of POLRs where the health history would recommend all the Material needed for case A (PM_A) and the same Proportion for case B (PM_B).

$$I_{AB} = \frac{PM_B}{PM_A} - 1 \tag{6.16}$$

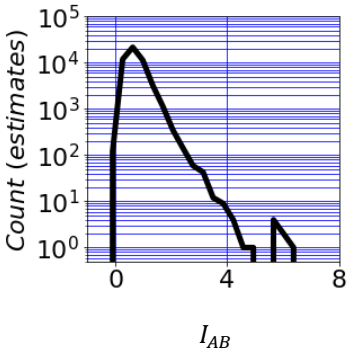
To estimate the confidence that the PEOHH enriches the health history, this section will illustrate techniques to estimate the probability that $PM_B > PM_A$, in other words that I_{AB} is positive. This section uses as an example the question of whether the health history has been enriched but this research will also use these techniques for other applications, such as to estimate the probability that the quality of record linkage has increased.

Please recall that TUB generates a list of estimates of I_{AB}. This research investigated the distribution of this list for both examples. The results are shown in Figure 6-18.

Bootstrapping is conditional on the original sample. It is limited by the sample data and this limitation of the technique is illustrated by Figure 6-18(a) which shows that the estimates of I_{AB} for example I are restricted to a narrow range of values that are all non-negative. This occurs because of the zero count of records in the sample where Case A would recommend all the required parts but Case B would not. In example I, TUB can never yield a negative estimate of I_{AB} but the probability that $PM_B < PM_A$ cannot really be zero because it is possible that the sample is not fully representative of the population. The proportion of bootstrap samples in example I for which $PM_B < PM_A$ is however zero and this strongly indicates that, in the parent population, $PM_B \geq PM_A$.

These results indicate that in both examples the probability that the health history had not been enriched would be low.

a)



b)

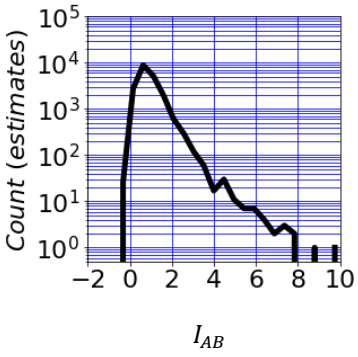


Figure 6-18, Frequency of Estimates of I_{AB}, with Subsamples = 4000, for (a) Example I, (b) Example II

The risk that the health history has not been enriched is the probability that $PM_B \leq PM_A$ and is estimated by the proportion of bootstrap subsamples for which $PM_B \leq PM_A$. The list of estimates is filtered to only include negative and zero values. The proportion of bootstrap subsamples for which $PM_B \leq PM_A$ is then simply the length of this filtered list divided by the number of subsamples.

The risk that the health history has been made less rich is the probability that $PM_B < PM_A$ and is estimated by the proportion of bootstrap samples for which $PM_B < PM_A$. This time, the list of

estimates is filtered to only include negative values. The proportion of bootstrap samples for which $PM_B < PM_A$ is then simply the length of this filtered list divided by the number of subsamples.

This research investigated the effect of varying the number of subsamples. The results are shown in Figure 6-19. Figure 6-19(a) uses data from example I. It shows no change in $P(PM_B < PM_A)$ with the number of subsamples because, in example I, as explained above, however many times TUB resamples, it can never yield a negative estimate of I_{AB}.

Figure 6-19(b) uses data from example II. It shows that in this example, $P(PM_B < PM_A)$ does change with the number of subsamples. The figure shows that the number of subsamples that are sufficient to adequately average out TUBs random sampling errors varies. For example I, $P(PM_B \leq PM_A)$ has just about stabilised after 50,000 subsamples. For example II, $P(P_B < P_A)$ has stabilised after 4000 subsamples but $P(PM_B \leq PM_A)$ has just about stabilised after 20,000 subsamples. For each application, this thesis will plot the proportions of interest against the number of subsamples to check that sufficient subsamples have been used to adequately average out TUBs random sampling errors.

The estimate of $P(PM_B < PM_A)$ having the value zero for example I, which represents real data, indicates that it is in fact unlikely that the health history has been made less rich and the estimate of $P(PM_B \leq PM_A)$ of 0.03% for the same example indicates that it is unlikely that it has not been enriched. The estimate of $P(PM_B < PM_A)$ of 0.2% for example II indicates that even with these alternative results it would still be unlikely that the health history had been made less rich and the estimate of $P(PM_B \leq PM_A)$ of 0.7% for the same example indicates that it would still be unlikely that the health history had not been enriched.

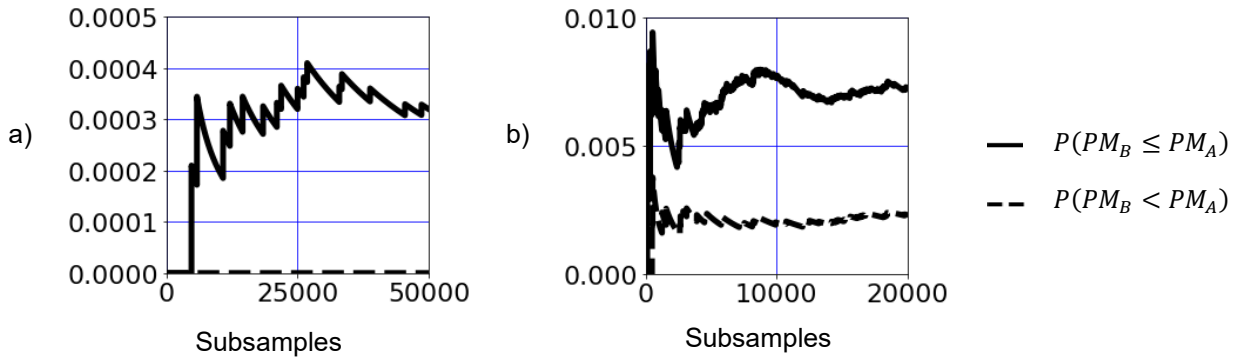


Figure 6-19, Proportion of Bootstrap Samples for which $P_B \leq P_A$ and that for which $P_B < P_A$ against Number of Subsamples for (a) Example I, (b) Example II

This section has shown that TUB can be used to identify the proportion of bootstrap samples for which $PM_B \leq PM_A$ and that for which $PM_B < PM_A$. This technique will be used to assess:

- The probability that the health history has been enriched
- The probability that the health history has been made less rich
- The probability that the quality of record linkage has increased
- The probability that the quality of record linkage has decreased

6.4 Conclusion to Quantifying Uncertainty

Section 6.1 proposed the use of statistical techniques for the interval estimation of a binomial proportion to understand the uncertainty of measures of the quality of record linkage. It reviewed literature (Wilson, 1927, Clopper and Pearson, 1934, Jeffreys, 1973, Agresti and Coull, 1998, Brown et al., 2001) demonstrates that, due to the underlying structure of the binomial distribution, there is no perfect technique for estimating the uncertainty as to the extent to which a binomial proportion should be representative of the population from which it has been sampled. It also demonstrates that some techniques such as that of Wilson, 1927, yield much better coverage than the standard technique of Wald, 1943. The section presented a novel and useful technique for comparing these methods.

Section 6.2 considered how to estimate the uncertainty of a difference in an uncertain estimate. It considered the case of comparing the enriched health history to the unenriched health history using the GSSLR. That case is analogous to comparing two medical techniques by testing them on the same patients at the same time. Such a medical experiment would of course not be feasible and statistical techniques for assessing this situation are consequently obscure. This chapter used a similar approach to that taken by Brown et al., 2001; estimating the Coverage Probability (*CP*) of the Confidence Interval (CI) at a specific Confidence Level (*CL*). It found that bootstrapping gave relatively good coverage and is an appropriate technique for this application.

Section 6.3 presented a test that will be applied to questions such as whether the PEOHH actually enriches the health history by finding the proportion of bootstrap samples for which it does.

Chapter 7 will use techniques from section 6.1 to understand the uncertainty of the Positive Predictive Value (*PPV*)[68], from section 6.2 to understand the uncertainty of how much changing a record linkage parameter changes the *PPV* and from section 6.3 to understand the uncertainty of whether changing a record linkage parameter increases the *PPV*. Section 7.6.2 will use bootstrapping to estimate the *PPV*.

Sections 8.3 to 8.5 will use techniques from section 6.1 to understand the uncertainty of various measures of the enrichment of the health history, from section 6.2 to understand the uncertainty of how much changing a health history filtering parameter changes these measures and from section 6.3 to understand the uncertainty of whether changing a health history filtering parameter increases these measures.

[68] The *PPV*, a measure of the quality of record linkage, was defined in section 2.2.1

7 Results: Optimisation of the Weights and Thresholds

This chapter will present the results from the validation of the Process for the Enrichment of OWT Health History (PEOHH) developed in this research. The PEOHH uses agreement and disagreement weights (defined in section 4.1) for each feature (defined in section 4.4) as well as various thresholds (each defined in chapter 4). It will present their optimisation. The literature on global optimisation will be reviewed and a global optimisation process applied.

Five of the thresholds that the PEOHH uses will be optimised outside of the global optimisation process and will instead be optimised individually. Optimising these five thresholds individually means that each of them is optimised only for a single value of each of the other weights and thresholds in the PEOHH. This has the disadvantage that it might not find the vector of values of the weights and thresholds that would yield the best quality of record linkage.

The five thresholds feature at the early stages of the PEOHH. This chapter optimises them by running the PEOHH for a range of values of the threshold. The optimisation of thresholds that feature early in the PEOHH is more computationally expensive than that of those weights and thresholds that feature later in the PEOHH because repeatedly running the last part of the PEOHH, the application of the weights, is less computationally expensive than repeatedly running more of the PEOHH. The inclusion of the five thresholds in the global optimisation process would make that process far too computationally expensive to be tractable.[69]

The review of global optimisation in section 2.4 described the so called ‘curse of dimensionality’ which, alongside the early stages of the PEOHH that the five thresholds feature at, is another reason why including these five thresholds in the global optimisation process would be so computationally expensive as to be unfeasible.

The computation times detailed in Section 4.5 are for a standard laptop PC processor. This research did not try a faster computer because it is Ørsted’s policy that their confidential data may not be copied on to non-Ørsted machines such as a Durham University parallel cluster. Further work could optimise the weights and thresholds using faster computers within Ørsted or another data owner’s permitted hardware.

This chapter will optimise the PEOHH by maximising the Positive Predictive Value (*PPV*), the proportion of classified matches that are correctly identified as such. Before this chapter presents the

[69] Section 4.5 details the computation times for the PEOHH.

optimisation of the five thresholds it will first explain how changes in the weights and thresholds affect changes in the *PPV*.

Figure 7-1 considers the effect of changing two of the weights that will be optimised in the global optimisation process. It shows the *PPV* against: (a) Agreement Weight for the Start time feature (AW_{St}) by Disagreement Weight for the Start time feature (DW_{St}) and (b) Disagreement Weight for the Start time feature (AW_{St}) by Agreement Weight for the Start time feature (DW_{St}) with all the other agreement weights set to 1 and all the other disagreement weights set to -1.

The figure shows that *PPV* changes with AW_{St} and with DW_{St}. The next figure will present an explanation of this phenomenon.

There are no missing work order start times and so, in the case of the start time feature, those POLRs in the GSSLR that do not agree all disagree. This causes symmetry between Figure 7-1(a) and (b).

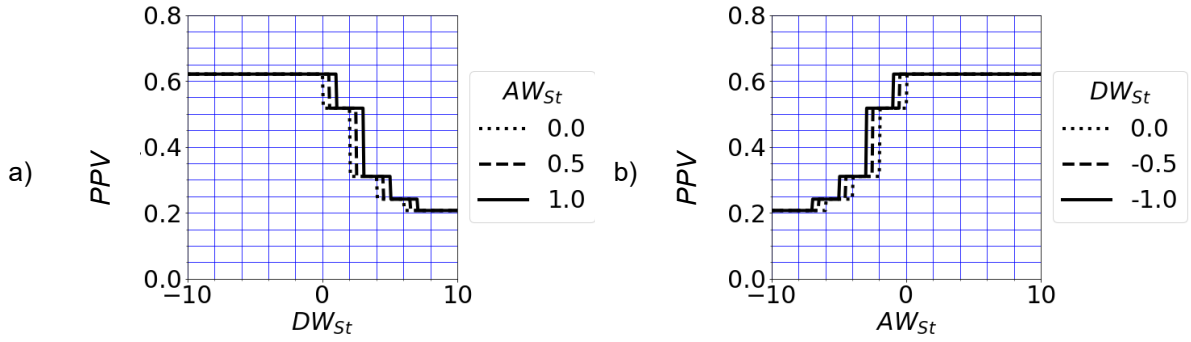


Figure 7-1, Positive Predictive Value (PPV) against:
(a) Agreement Weight for the Start time feature (AW_{St})
by Disagreement Weight for the Start time feature (DW_{St})
(b) Disagreement Weight for the Start time feature (DW_{St})
by Agreement Weight for the Start time feature (AW_{St})

To explain what causes the change in the *PPV* of the GSSLR, this chapter will now present in greater detail the effect of varying a weight. Figure 7-2 shows the effect of varying AW_{St} with all other weights and thresholds constant. Each row in Figure 7-2 represents one WO. As AW_{St} is varied, different versions of the EHH are created. Where the outage linked to the given WO in the version of the EHH generated at the specified value of AW_{St} is the same as that in the GSSLR, figure 7.2 shows blue. Where they are different, Figure 7-2 shows white. Values of AW_{St} with a higher proportion of blue in the figure yield a higher *PPV*. As AW_{St} changes, which WOs it is that are matched to their correct outage changes and so *PPV* fluctuates.

Consider a set of WOs, each one linked correctly to its corresponding outage. This will be referred to as the 'true EHH'. At a hypothetical weighting that yielded the true EHH, the column would be blue and the *PPV* of the GSSLR would be 1.

As shown in Figure 7-2, the start time feature registers either agreement or disagreement. All the features of the PEOHH register either agreement, disagreement or neither for each WO at a given value of the relevant thresholds. This three-state approach is what causes the stepped behaviour in the fluctuations of the PEOHH that were shown in Figure 7-1.

In Figure 7-2, some WOs in the GSSLR register agreement for all values of AW_{St} while others register disagreement for all values.

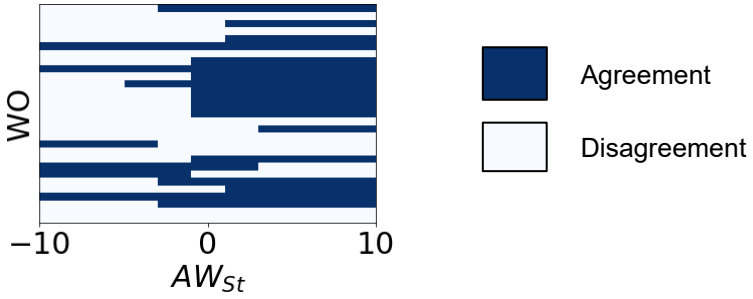


Figure 7-2, Effect of varying the Agreement Weight for the Start Time Feature (AW_{St}) on the Agreement or Disagreement of the EHH with GSSLR for each Work Order (WO) in the GSSLR

Using AW_{St} as an example, we have seen that changing the weights and thresholds changes the *PPV*. Each element of the vector of values of weights and thresholds can be conceived of as a dimension that the *PPV* can be plotted against.

The *PPV* of the GSSLR is an uncertain estimate of the *PPV* of the EHH. The previous chapter found an appropriate technique for quantifying that uncertainty. If we conceive of the plot of the *PPV* of the EHH against all the elements of the vector of values of weights and thresholds as a stepped, multi-dimensional surface then we should understand that we cannot measure the exact level of that surface but that we can only estimate it from the *PPV* of the GSSLR.

7.1 Optimisation of the Blocking Threshold

Section 4.1 presented the blocking threshold. It was described in that section that the first step of the PEOHH is to join WOs to outages to create POLR. The PEOHH adds a 40-day margin, known in this thesis as the blocking threshold, to the outage start and finish times to create an extended duration. It joins each WO to each of the outages of which the extended duration includes the WOs start date.

Section 4.3.2 presented the method for the validation of the techniques for health history enrichment used in this thesis. Please recall from that section that the size of the sample "Gold Standard" Set of Linked Records (GSSLR), 29 WOs, was constrained by the amount of expert time that was available.

Section 4.4.1 introduced the time difference between the WO start date and the outage (Δt_{St}). To optimise the blocking threshold, this research investigated the distribution of Δt_{St} in the GSSLR and the results are shown in Figure 7-3. Of the 29 POLRs in the GSSLR only 2 have a Δt_{St}, defined in chapter 4, of zero. This result shows that the blocking threshold is required. The figure shows that one POLR in the GSSLR is an outlier in terms of Δt_{St}, at 39.5 days. On the assumption that this extreme point is genuine, and therefore representative of the real-life situation, it should be within the blocking threshold and so this should be at least 40 days. It is not possible that the real date when work started was not during the correctly matched outage. A WO start date could be outside of an outage for two reasons: that the WO refers to a different outage or that there is an error in the start date record. The start date is a human generated data point and all the wind turbine experts that this research elicited the opinions of expressed the opinion that such errors are expected. In the design of the process for developing the PEOHH, this research assumed that 40 days was a conservatively high starting value and that results would show that it could be reduced. The result in Figure 7-3, however, shows that the full 40-day threshold is required and that further work might consider an even larger threshold.

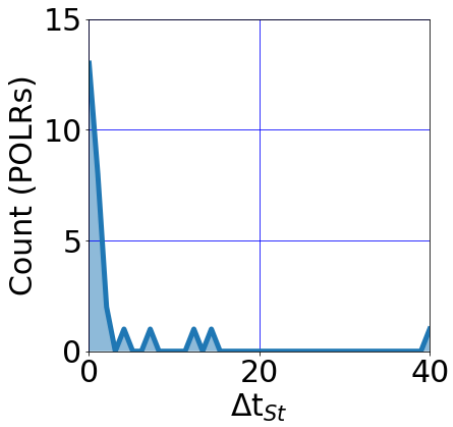


Figure 7-3, Count (POLRs) against Time Difference between the WO Start Time and the Outage (Δt_{St}) for the GSSRL

7.2 Optimisation of the Time Difference Between the Outage and the Alarm

Section 4.4.4 presented the Threshold for the Time difference between the Outage interval and Alarm interval (*TTOA*). It was described in that section that the PEOHH extracts all the alarms from the alarm log from +/- *TTOA* of the outage start and finish times.

This research investigated the effect of varying *TTOA* to maximise the *PPV*. The results are shown in Figure 7-4which shows *PPV* against *TTOA.* Other weights and thresholds are kept constant. The Agreement Weight for the Description and the Alarm Code features using Alarms and Outages ($AW_{De_{Ou}}$, $AW_{AC_{Ou}}$, $AW_{De_{Al}}$, $AW_{AC_{Al}}$) are set to 1 and their Disagreement Weights ($DW_{De_{Ou}}$, $DW_{LT_{Ou}}$, $DW_{De_{Al}}$, $DW_{LT_{Al}}$) are set to -1 while all other agreement and disagreement weights are set to 0.

The maximum *PPV* in Figure 7-4 (0.517) occurs at *TTOA* = 0; at this value, the effect of the features using alarms is null and so the failure mode features have the least effect, increasing the relative effect of the start date time difference which results in a higher *PPV*. The second highest *PPV* (0.483) occurs when *TTOA* is at values between 4 and 21 minutes. The initial value of *TTOA* was 10 minutes and this is shown to be within the range that achieves the highest value of *PPV* (disregarding zero as this value effectively turns the feature off). This result implies that further iteration is not required, and that this thesis can recommend a *TTOA* of 10 minutes, the value initially recommended by a wind turbine expert. This section will consider the uncertainty of this recommendation.

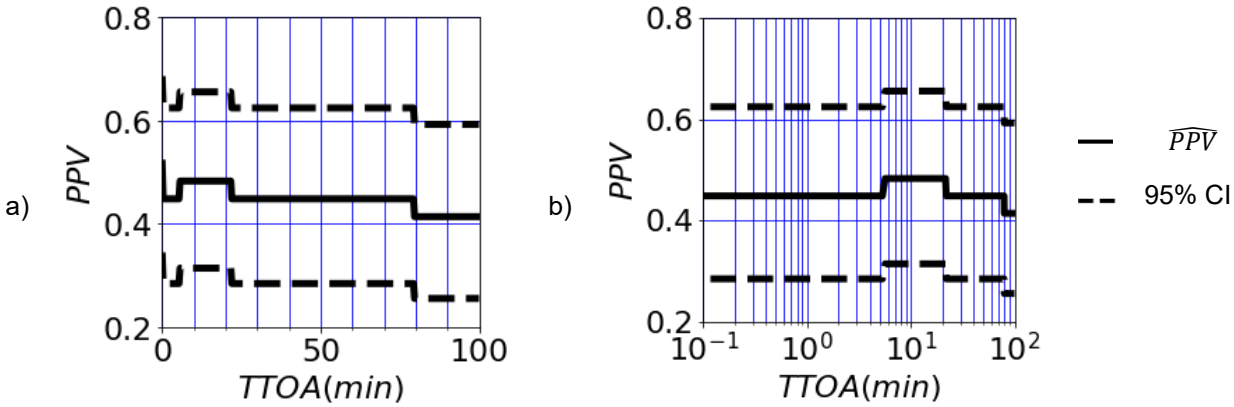


Figure 7-4, Positive Predictive Value (PPV) and 95% CI against Time Difference Threshold Between the Outage Interval and the Alarm Interval (TTOA), (a) Linear scale, (b) Log scale for the x axis

Figure 7-4 shows the 95% CI for the *PPV* of the EHH identified using the Wilson interval which was discussed in section 6.1.3. The small size of the GSSLR, only 29 POLRs, means that, for a *PPV* of the GSSLR of 0.483, the 95% CI of the *PPV* of the EHH is (0.314, 0.656).

Section 6.2 considered techniques for interval estimation for a change in a proportion. It showed that the Technique Assuming Independence (TAI) yielded very poor coverage and that the Technique Using Bootstrapping (TUB) yielded acceptable coverage. In Figure 7-4, the width of the CI is greater than the difference between the maximum and minimum values of the *PPV* of the GSSLR. Intuitively, one might infer from this that there is a significant risk that the optimum value has not been identified. Such an inference would however be based on the invalid assumption that the measurements of the *PPV* of the GSSLR are independent of each other. This section will instead use TUB for constructing a CI for the change in the *PPV* of the EHH.

Table 7-1 shows the effect on the quality of record linkage of varying *TTOA*. '*True'* is the number of records from the GSSLR in which the version of the health history linked the WO to the same outage as that in the GSSLR and '*False'* is the number in which it linked the WO to a different outage. TUB uses these results to estimate the uncertainty of the difference that it makes.

		TTOA = 10 mins	
		True	False
TTOA = 100 mins	True	12	0
	False	2	15

Table 7-1, Effect on the Quality of Record Linkage of Varying the Time Difference Threshold between the Outage Interval and the Alarm Interval (TTOA)

Equation 7.1 defines the Increase in the *PPV* of the EHH that would by yielded by a *TTOA* of 10 minutes over that yielded by a *TTOA* of 100 minutes (${}^{PPV}_{TTOA}I^{100}_{10}$). It uses the Probability distribution of estimates of the *PPV* of the EHH that would be yielded by a *TTOA* of 10 minutes ($P(PPV \mid TTOA = 10)$) and the same distribution that would be yielded by a *TTOA* of 100 minutes ($P(PPV \mid TTOA = 100)$)

$$ {}^{PPV}_{TTOA}I^{10}_{100} = \frac{P(PPV \mid TTOA = 10)}{P(PPV \mid TTOA = 100)} - 1 \qquad (7.1) $$

Equation 7.2 evaluates $\widehat{{}^{PPV}_{TTOA}I^{10}_{100}}$, a point estimate of ${}^{PPV}_{TTOA}I^{10}_{100}$, using the data from table 7.1.

$$\widehat{{}^{PPV}_{TTOA}I^{10}_{100}} = \frac{12+2}{12+0} - 1 = 17\% \tag{7.2}$$

This research used TUB to estimate the uncertainty of ${}^{PPV}_{TTOA}I^{10}_{100}$, which yielded a 95% CI of (0, 56%). This research predicts that this uncertainty will be drastically reduced by an upcoming innovation in maintenance record keeping that will be discussed in chapter 9; the automatic linking of new WOs to outages will result in a larger GSSLR and consequently in more certainty for this estimate.

To better understand this uncertainty, this research investigated the distribution of TUBs list of estimates of ${}^{PPV}_{TTOA}I^{10}_{100}$ using techniques discussed in section 6.3. The results are shown in Figure 7-5 (a). The most important features of this distribution are the proportion of negative estimates and the proportion of zero estimates because these indicate respectively that the quality of record linkage has reduced and that it has not increased, as detailed in Figure 7-5. The probability that a *TTOA* of 10 minutes would yield a lower *PPV* of the EHH than a *TTOA* of 100 minutes is estimated by the proportion of bootstrap samples for which ${}^{PPV}_{TTOA}I^{10}_{100} \leq 0$, which this research found to be 12%. This indicates that there is a significant risk that optimising *TTOA* using a GSSLR of the size available does not improve the quality of record linkage.

Figure 7-5 (b) also shows that the proportion of bootstrap samples for which ${}^{PPV}_{TTOA}I^{10}_{100} < 0$ is 0. As explained in section 6.3, bootstrapping is conditional on the original sample, but this indicates that there is little risk that optimising *TTOA* using a GSSLR of the size available reduces the quality of record linkage.

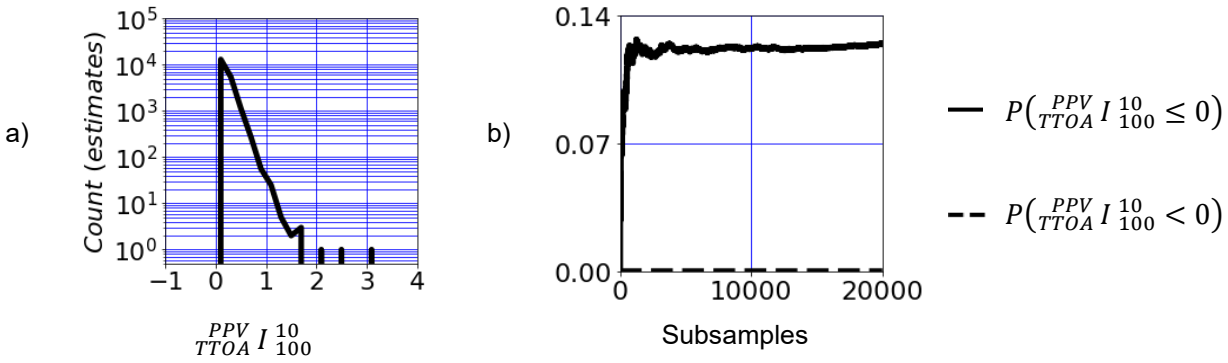


Figure 7-5, (a) Frequency of ${}^{PPV}_{TTOA}I^{10}_{100}$ for 20,000 Subsamples, (b) Proportion of Bootstrap Samples for which ${}^{PPV}_{TTOA}I^{10}_{100} \leq 0$ and for which ${}^{PPV}_{TTOA}I^{10}_{100} < 0$ against Number of Subsamples

This research kept *TTOA* as a single variable for all three of the failure mode features. It could instead have used a separate variable for each of these features. The decision to use a single variable was a simplifying assumption taken because the relationship between outages and alarms can be assumed to be independent of which feature indicative of the failure mode is used.

This research did not investigate the effect of varying *TTOA* on the feature that uses parts data to identify the failure mode (described in section 4.4.4.3.3) because this feature takes a long time to run and so running it multiple times would take far too long. The results from varying *TTOA* on the other two features that use the failure mode (described in section 4.4.4.1 and in section 4.4.4.2) are sufficient to optimise *TTOA* because all these techniques refer to the same WOs and to the same outages.

This section presented the effect of varying *TTOA* and recommended keeping it at 10 minutes.

7.3 Optimisation of the Description Threshold

This section will present the effect of varying the Description Threshold (Th_{De}) that was introduced in section 4.4.4.1. Please recall from that section that the PEOHH records agreement for the description feature when the Similarity Ratio (*SR*) is above a Th_{De} nominally set to 0.75. Section 7.6.3 will present results that indicate that the effect of this feature and of all of the set of features that use the failure mode as part of an ensemble of features for record linkage comparison are too small to measure with a GSSLR of the size available. For that reason, this section will consider this feature used alone.

This research investigated the effect of varying Th_{De} to maximise the PPV. The results are shown in Figure 7-6 which shows *PPV* against Th_{De}. Other weights and thresholds are kept constant. The Agreement Weight for the Description feature using Alarms and Outages ($AW_{De_{Ou}}$, $AW_{De_{Al}}$) are set to 1 and their Disagreement Weights ($DW_{De_{Ou}}$, $DW_{De_{Al}}$) are set to -1 while all other agreement and disagreement weights are set to 0.

Figure 7-6 shows that a wide range of values of Th_{De} (0.5 to 1) all yield the same value of *PPV* of 0.552. This thesis recommends that Th_{De} continues to be set to 0.75. This section will consider the uncertainty of this recommendation.

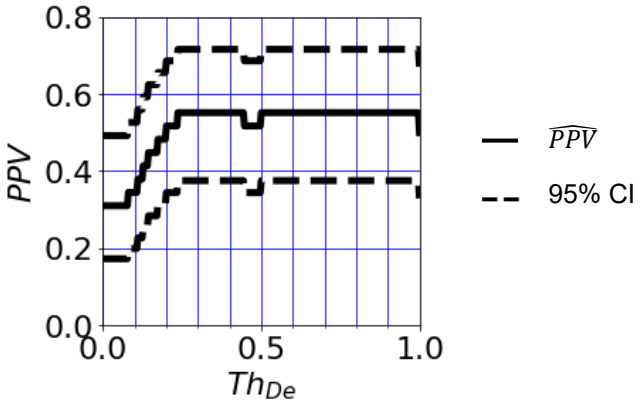


Figure 7-6, Positive Predictive Value (PPV) and 95% CI against Description Threshold between the Outage Alarm Description and the WO Description (Th_{De})

Figure 7-6 shows the 95% CI for the *PPV* of the EHH identified using the Wilson interval which was discussed in section 6.1.3. The small size of the GSSLR, only 29 POLRs, means that, for a *PPV* of the GSSLR of 0.552, the 95% CI of the *PPV* of the EHH is (0.375, 0.716).

Again, this section will use TUB, presented in section 6.2, for constructing a CI for the change in the *PPV* of the EHH. Table 7-2 shows the effect of varying Th_{De}. '*True'* is the number of records from the GSSLR in which the version of the health history linked the WO to the same outage as that in the GSSLR and '*False'* is the number in which it linked the WO to a different outage. TUB uses these results to estimate the uncertainty of the difference that it makes.

		Th_{De} = 0.75	
		True	False
Th_{De} = 0	True	8	1
	False	8	12

Table 7-2, Effect on the Quality of Record Linkage of Varying the Description Threshold (Th_{De})

Equation 7.3 defines the Increase in the *PPV* of the EHH that would by yielded by a Th_{De} of 0.75 over that yielded by a Th_{De} of zero (${}^{PPV}_{Th_{De}}I^{\,0}_{0.75}$). It uses the Probability distribution of estimates of the *PPV* of the EHH that would be yielded by a Th_{De} of 0.75 ($P(PPV \mid Th_{De} = 0.75)$) and the same distribution that would be yielded by a Th_{De} of zero ($P(PPV \mid Th_{De} = 0)$).

$$ {}^{PPV}_{Th_{De}}I^{\,0.75}_{0} = \frac{P(PPV \mid Th_{De} = 0.75)}{P(PPV \mid Th_{De} = 0)} - 1 \tag{7.3} $$

Equation 7.4 evaluates ${}^{PPV}_{Th_{De}}\widehat{I^{\,0.75}_{0}}$, a point estimate of ${}^{PPV}_{Th_{De}}I^{\,0.75}_{0}$, using the data from Table 7-2.

$$ {}^{PPV}_{Th_{De}}\widehat{I^{\,0.75}_{0}} = \frac{8+8}{8+1} - 1 = 78\% \tag{7.4} $$

This research used TUB to estimate the uncertainty of ${}^{PPV}_{Th_{De}}I^{\,0}_{0.75}$, which yielded a 95% CI of (13%, 225%). Again, this research predicts that this uncertainty will be drastically reduced by an upcoming innovation in maintenance record keeping that will be discussed in chapter 9.

To better understand this uncertainty, this research investigated the distribution of TUBs list of estimates of ${}^{PPV}_{Th_{De}}I^{\,0.75}_{0}$ using techniques discussed in section 6.3. The results are shown in Figure 7-7 (a). The most important features of this distribution are the proportion of negative estimates and the proportion of zero estimates because these indicate respectively that the quality of record linkage has reduced and that it has not increased, as detailed in Figure 7-7 (b). The probability that a Th_{De} of 0.75 would yield a lower *PPV* of the EHH than a Th_{De} of zero is estimated by the proportion of bootstrap

samples for which ${}^{PPV}_{Th_{De}} I^{0.75}_{0} \leq 0$, which this research found to be 0.6%. This indicates that there is little risk that optimising Th_{De} using a GSSLR of the size available does not improve the quality of record linkage.

Figure 7-7 (b) also shows that the proportion of bootstrap samples for which ${}^{PPV}_{Th_{De}} I^{0.75}_{0} < 0$ is 0.2%. This indicates that there is little risk that optimising Th_{De} using a GSSLR of the size available reduces the quality of record linkage.

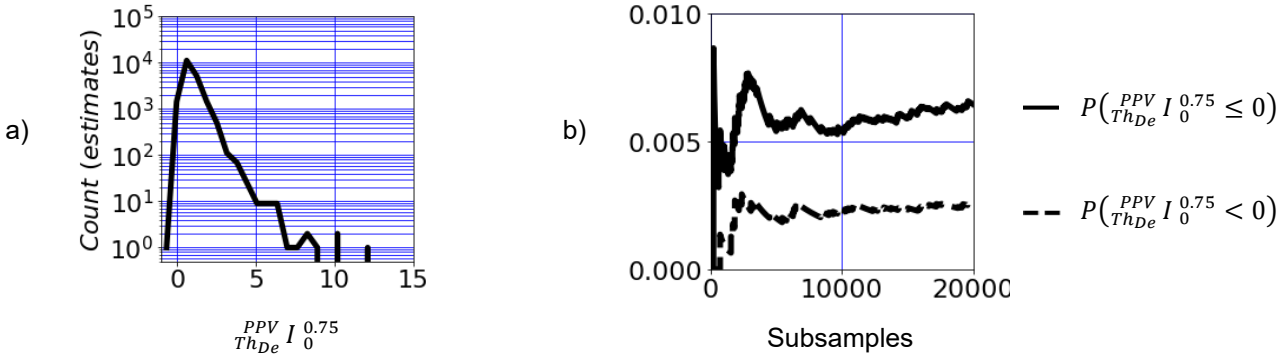


Figure 7-7, (a) Frequency of ${}^{PPV}_{Th_{De}} I^{0.75}_{0}$ for 20,000 Subsamples, (b) Proportion of Bootstrap Samples for which ${}^{PPV}_{Th_{De}} I^{0.75}_{0} \leq 0$ and for which ${}^{PPV}_{Th_{De}} I^{0.75}_{0} < 0$ against Number of Subsamples

The conclusion of this section is to recommend keeping Th_{De} at 0.75.

7.4 Optimisation of the Parts Training Data Score Threshold

Section 4.4.4.3.1 showed how the training data to be used by the HHE techniques using the parts are selected. Please recall from the section that:

> "The PEOHH selects POLRs for inclusion in the training data using the same criteria that it uses to select POLRs for the EHH. It then filters the training data to retain those POLRs that have S_{POLR} above a Threshold (Th_{SP})."

This section will present the Process for the Optimisation of Th_{SP} (POTS) and its result. It will show that a value of 1.7, as is used throughout the other sections of this thesis, is appropriate. This value was originally identified by inspection of the distribution of S_{POLR}.

The parts frequency technique presented in section 4.4.4.3.3 is trained for each WO separately against all the other WOs in the dataset. POTS simulates this process using a computationally less expensive process.

POTS trains its model using a version of the EHH generated using the optimised weights and thresholds that will be presented in section 7.6.3. These weights will disregard the parts data. If the optimised PEOHH did use the parts data then it would be necessary to generate an alternative version of the EHH disregarding the parts data in order to ensure that the inputs to the PEOHH are independent of its output. This would ensure that POTS was measuring features of the real material consumption data and that it was not distorted by circular logic.

POTS identifies an Edited version of the EHH (EEHH) by removing those WOs that feature in the GSSLR from the EHH. POTS is trained against the EEHH. Generating a single set of training data that can be used for each WO in the GSSLR is less computationally expensive than training for each WO against all the other WOs in the dataset.

POTS trains the parts frequency technique on health history datasets generated using a full range of values of Th_{SP}.

Table 7-8 will present the optimised weights and thresholds. The sum of the column of agreement weights will be 3.127 and the sum of the column of disagreement weights will be -2.643. These values are the limits of possible values of Score for each POLR (S_{POLR}). Varying Th_{SP} from -4 to 4 exceeds the range of possible values of S_{POLR} and so it is a full range of values.

This research investigated the effect of varying Th_{SP} to maximise the PPV. The results are shown in Figure 7-8. It shows that Th_{SP} values from 1.1 to 1.9 yield the maximum *PPV* of 0.552. This section

consequently recommends that Th_{SP} should remain at its initial value of 1.7. This section will consider the uncertainty of this recommendation.

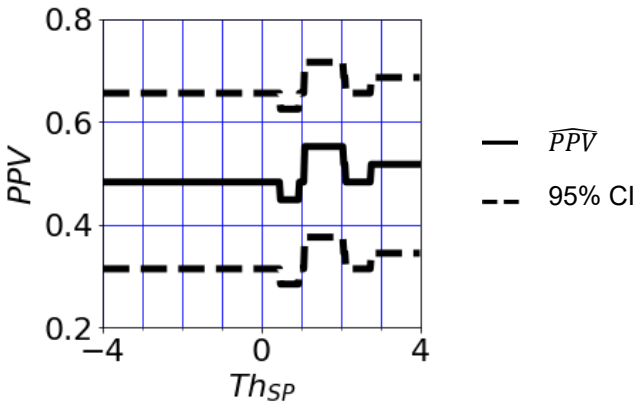


Figure 7-8, Positive Predictive Value (PPV) and 95% CI Against Parts Training Data Score Threshold (Th_{SP})

Figure 7-8 shows the 95% CI for the *PPV* of the EHH identified using the Wilson interval which was discussed in section 6.1.3. The small size of the GSSLR, only 29 POLRs, means that, for a *PPV* of the GSSLR of 0.552, the 95% CI of the *PPV* of the EHH is (0.375, 0.716).

Again, this section will use TUB, TUB, presented in section 6.2, for constructing a CI for the change in the *PPV* of the EHH. Table 7-3 shows the effect of varying Th_{SP}. '*True'* is the number of records from the GSSLR in which the version of the health history linked the WO to the same outage as that in the GSSLR and '*False'* is the number in which it linked the WO to a different outage. TUB uses these results to estimate the uncertainty of the difference that it makes.

		Th_{SP} = 1.7	
		True	False
Th_{SP} = 0	True	14	0
	False	2	13

Table 7-3, Effect on the Quality of Record Linkage of Varying the Parts Training Data Score Threshold (Th_{SP})

Equation 7.5 defines the Increase in the *PPV* of the EHH that would by yielded by a Th_{SP} of 1.7 over that yielded by a Th_{SP} of zero (${}^{PPV}_{Th_{SP}} I^{0}_{1.7}$). It uses the Probability distribution of estimates of the *PPV* of the EHH that would be yielded by a Th_{SP} of 1.7 ($P(PPV \mid Th_{SP} = 1.7)$) and the same distribution that would be yielded by a Th_{SP} of zero ($P(PPV \mid Th_{SP} = 0)$)

$$ {}^{PPV}_{Th_{SP}} I^{1.7}_{0} = \frac{P(PPV \mid Th_{SP} = 1.7)}{P(PPV \mid Th_{SP} = 0)} - 1 \tag{7.5} $$

Equation 7.6 evaluates ${}^{\widehat{PPV}}_{Th_{SP}} I^{1.7}_{0}$, a point estimate of ${}^{PPV}_{Th_{SP}} I^{1.7}_{0}$, using the data from Table 7-3.

$$ {}^{\widehat{PPV}}_{Th_{SP}} I^{1.7}_{0} = \frac{14 + 2}{14 + 0} - 1 = 14\% \tag{7.6} $$

This research used TUB to estimate the uncertainty of ${}^{PPV}_{Th_{SP}} I^{1.7}_{0}$, which yielded a 95% CI of (0, 45%). Again, this research predicts that this uncertainty will be drastically reduced by an upcoming innovation in maintenance record keeping that will be discussed in chapter 9.

To better understand this uncertainty, this research investigated the distribution of TUBs list of estimates of ${}^{PPV}_{Th_{SP}} I^{1.7}_{0}$ using techniques presented in section 6.3. The results are shown in Figure 7-9(a). The most important features of this distribution are the proportion of negative estimates and the proportion of zero estimates because these indicate respectively that the quality of record linkage has reduced and that it has not increased, as detailed in Figure 7-9(b). The probability that a Th_{SP} of 1.7 would yield a lower *PPV* of the EHH than a Th_{SP} zero is estimated by the proportion of bootstrap samples for which ${}^{PPV}_{Th_{SP}} I^{1.7}_{0} \leq 0$, which this research found to be 13%. This indicates that there is a significant risk that optimising Th_{SP} using a GSSLR of the size available does not improve the quality of record linkage.

Figure 7.9 (b) also shows that the proportion of bootstrap samples for which ${}^{PPV}_{Th_{SP}} I^{1.7}_{0} < 0$ is 0. As explained in section 6.3, bootstrapping is conditional on the original sample, but this indicates that there is little risk that optimising *TTOA* using a GSSLR of the size available reduces the quality of record linkage.

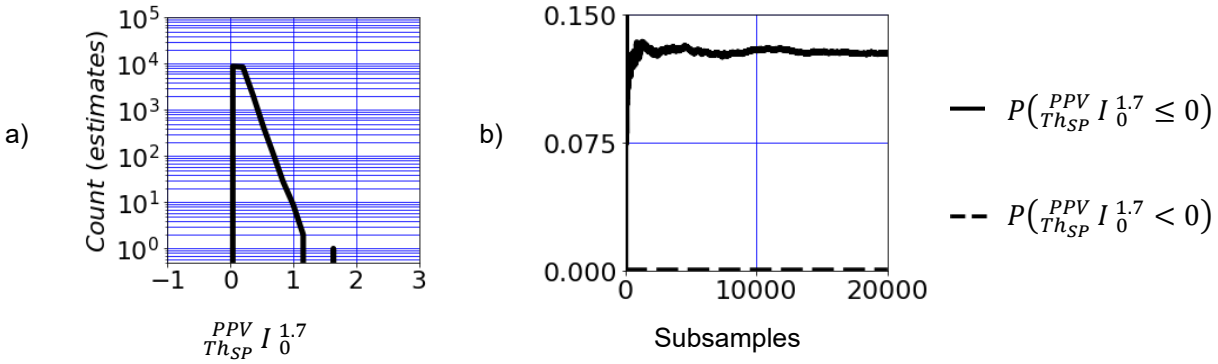


Figure 7-9, (a) Frequency of ${}^{PPV}_{Th_{SP}} I^{1.7}_{0}$ for 20,000 Subsamples, (b) Proportion of Bootstrap Samples for which ${}^{PPV}_{Th_{SP}} I^{1.7}_{0} \leq 0$ and for which ${}^{PPV}_{Th_{SP}} I^{1.7}_{0} < 0$ against Number of Subsamples

The conclusion of this section is to recommend keeping Th_{SP} at 1.7.

7.5 Optimisation of the Parts Score Threshold

This section will present the effect of varying the Parts score Threshold (Th_{Pa}) presented in section 4.4.4.3.3. It was described in that section that the PEOHH records agreement when *PS* > Th_{Pa} and that Th_{Pa} was set to a nominal value of 0.

Section 7.6.3 will present results that indicate that the effect of this feature and of all of the set of features that use the failure mode as part of an ensemble of features for record linkage comparison are too small to measure with a GSSLR of the size available. For that reason, this section will consider this feature used alone.

This research investigated the effect of varying Th_{Pa} to maximise the PPV. The results are shown in Figure 7-10 which shows *PPV* against Th_{Pa}. Other weights and thresholds are kept constant. The Agreement Weight for the Parts feature using Alarms and Outages ($AW_{Pa_{Ou}}$, $AW_{Pa_{Al}}$) are set to 1 and their Disagreement Weights ($DW_{Pa_{Ou}}$, $DW_{Pa_{Al}}$) are set to -1 while all other agreement and disagreement weights are set to 0. This has the effect of looking at the parts feature alone. It shows that a range of values of Th_{Pa} (1 to 3 and also -3), all yield the same value of *PPV* of 0.517. This thesis recommends that Th_{Pa} be set to 2. This section will consider the uncertainty of this recommendation.

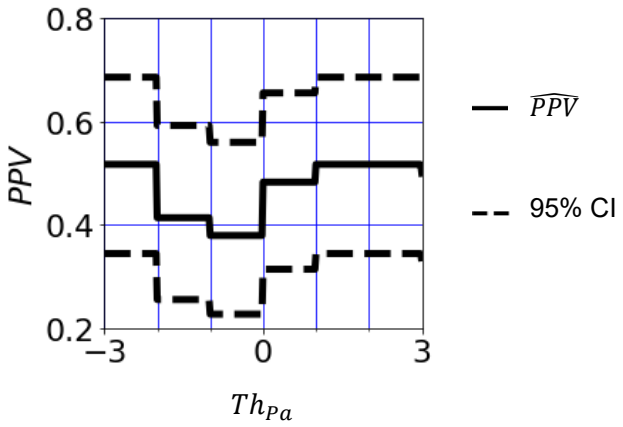


Figure 7-10, Positive Predictive Value (PPV) and 95% CI against Parts Score Threshold (Th_{Pa})

Figure 7-10 shows the 95% CI for the *PPV* of the EHH identified using the Wilson interval which was discussed in section 6.1.3. The small size of the GSSLR, only 29 POLRs, means that, for a *PPV* of the GSSLR of 0.517, the 95% CI of the *PPV* of the EHH is (0.344, 0.686).

Again, this section will use TUB, presented in section 6.2, for constructing a CI for the change in the *PPV* of the EHH. Table 7-4 shows the effect of varying Th_{Pa}. '*True'* is the number of records from the GSSLR in which the version of the health history linked the WO to the same outage as that in the GSSLR and '*False'* is the number in which it linked the WO to a different outage. TUB uses these results to estimate the uncertainty of the difference that it makes.

		Th_{Pa} = -0.5	
		True	False
Th_{Pa} = 2	True	11	0
	False	4	14

Table 7-4, Effect on the Quality of Record Linkage of Varying the Parts Score Threshold (Th_{Pa})

Equation 7.7 defines the Increase in the *PPV* of the EHH that would by yielded by a Th_{Pa} of 2 over that yielded by a Th_{Pa} of -0.5 ($^{PPV}_{Th_{Pa}} I^{-0.5}_{2}$). It uses the Probability distribution of estimates of the *PPV* of the EHH that would be yielded by a Th_{Pa} of 2 ($P(PPV \mid Th_{Pa} = 2)$) and the same distribution that would be yielded by a Th_{Pa} of -0.5 ($P(PPV \mid Th_{Pa} = -0.5)$)

$$^{PPV}_{Th_{Pa}} I^{2}_{-0.5} = \frac{P(PPV \mid Th_{Pa} = 2)}{P(PPV \mid Th_{Pa} = -0.5)} - 1 \tag{7.7}$$

Equation 7.8 evaluates $\widehat{^{PPV}_{Th_{Pa}} I^{2}_{-0.5}}$, a point estimate of $^{PPV}_{Th_{Pa}} I^{2}_{-0.5}$, using the data from Table 7-4.

$$\widehat{^{PPV}_{Th_{Pa}} I^{2}_{-0.5}} = \frac{11 + 4}{11 + 0} - 1 = 36\% \tag{7.8}$$

This research used TUB to estimate the uncertainty of $^{PPV}_{Th_{Pa}} I^{2}_{-0.5}$, which yielded a 95% CI of (7% , 100%). Again, this research predicts that this uncertainty will be drastically reduced by an upcoming innovation in maintenance record keeping that will be discussed in chapter 9.

To better understand this uncertainty, this research investigated the distribution of TUBs list of estimates of $^{PPV}_{Th_{Pa}} I^{2}_{-0.5}$ using techniques presented in section 6.3. The results are shown in Figure 7-11 (a). The most important features of this distribution are the proportion of negative estimates and the proportion of zero estimates because these indicate respectively that the quality of record linkage has reduced and that it has not increased, as detailed in figure 7.11 (b). The probability that a Th_{Pa} of

2 would yield a lower *PPV* of the EHH than a Th_{Pa} of -0.5 is estimated by the proportion of bootstrap samples for which ${}^{PPV}_{Th_{Pa}} I^{\,2}_{-0.5} \leq 0$, which this research found to be 1.5%. This indicates that there is little risk that optimising Th_{Pa} using a GSSLR of the size available does not improve the quality of record linkage.

Figure 7-11(b) also shows that the proportion of bootstrap samples for which ${}^{PPV}_{Th_{Pa}} I^{\,2}_{-0.5} < 0$ is 0. As explained in section 6.3, bootstrapping is conditional on the original sample, but this indicates that there is little risk that optimising Th_{Pa} using a GSSLR of the size available reduces the quality of record linkage.

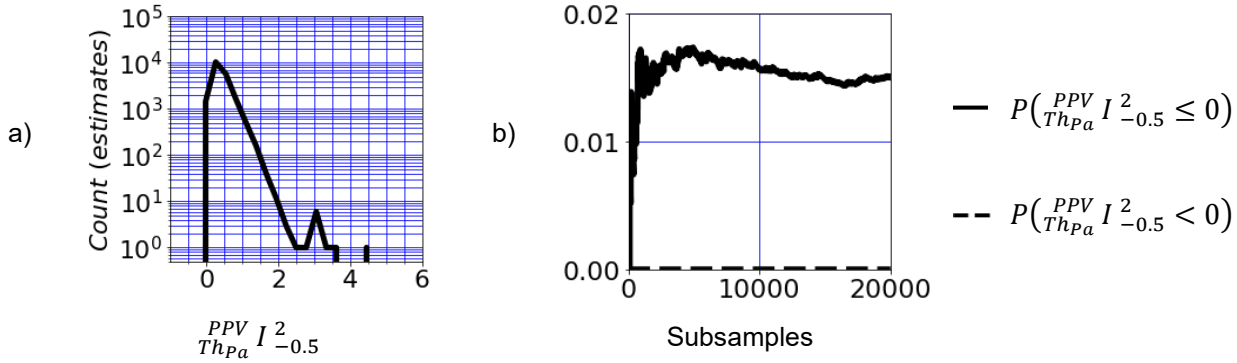


Figure 7-11, (a) Frequency of ${}^{PPV}_{Th_{Pa}} I^{\,2}_{-0.5}$ for Subsamples = 20,000, (b) Proportion of Bootstrap Samples for which ${}^{PPV}_{Th_{Pa}} I^{\,2}_{-0.5} \leq 0$ and for which ${}^{PPV}_{Th_{Pa}} I^{\,2}_{-0.5} < 0$ against Number of Subsamples

The conclusion of this section is to recommend setting Th_{Pa} at 2.

7.6 Optimisation of the Weights and of the Remaining Thresholds

This section presents the optimisation of the 28 remaining weights and thresholds used in the PEOHH that have not already been addressed. These are:

- The Agreement and Disagreement Weight for each of the 12 Features (AW_{Fe}, DW_{Fe}) presented in section 4.1.
- The Threshold for each of four time Features (Th_{Fe}) presented in section 4.4.1.

This optimisation will search for the maximum value of *PPV*. If the optimised AW_{Fe} or DW_{Fe} for a feature is not zero then this research will have found that the hypothesis that the feature can be used for record linkage comparison is true. It will use Differential Evolution (DE), which was reviewed in section 2.4.1.2, for this purpose.

This section will again use techniques from chapter 6 to quantify the uncertainty of this optimisation and will go on to consider how this uncertainty limits what conclusions can be drawn.

The optimisation process described in this section constrained all AW_{Fe} between zero and one and all DW_{Fe} between zero and negative one. These weights are relative to each other, so this is equivalent to constraining all the agreement weights between zero and + infinity and the disagreement weights between zero and - infinity. This means that any positive value of relative agreement weight and any negative value of relative disagreement weight is within the bounds of the search.

Section 4.1 presented the PEOHH:

> The PEOHH calculates a score for each POLR (S_{POLR}), the sum across all the features of the Weight for that Feature and for that POLR ($W_{Fe_{POLR}}$) using equation 4.2.

$$S_{POLR} = \sum_{Fe} W_{Fe_{POLR}} \tag{4.2}$$

7.6.1 Initial Optimisation

This section will use Differential Evolution (DE)[70] to optimise a set of weights and thresholds in the PEOHH. It will define a function (*fun*) that, for each POLR, uses the weights and thresholds to calculate S_{POLR} and, for each WO, finds the highest scoring POLR. For any combination of weights and thresholds, *fun* identifies an EHH that might be the same or that might be different to the EHH

[70] Section 2.4.1.2 reviewed DE.

identified using another combination of weights and thresholds. *Fun* uses the GSSLR to yield a value of the *PPV* for the given combination of weights and thresholds.

This section will present results obtained by DE for the optimisation of the PEOHH. It will present the effect on *PPV* of disregarding individual features and sets of features. The results from this section will inform the conclusions that this chapter will draw about which of the new and existing methods for record linkage presented in this chapter are beneficial for the application of wind turbine health history enrichment.

The maximum number of iterations was not reached in this research as instead the convergence criteria were reached.

For each weight and each threshold to be optimised together, the optimisation process defined a range of values. The optimisation process ran the Python tool 'differential_evolution' from the library 'scipy.optimize' on the function *fun*. This process yielded an optimised combination of weights and thresholds.

This research investigated the effect of adjusting the settings of the DE optimiser. DE with lower *CR* and *Mu* and higher *Di* and *PSi* has the effect of widening the search radius; increasing the chance of finding the optimum; but it also has the effect of slowing convergence (Piotrowski, 2017). This research ran DE with such control parameter settings *(CR = 0.1, Mu = 0.5, Di = 1, PSi = 100)* but stopped the solving after it had run without convergence for 5 days. That run found the optimum value of *PPV* of 0.828. This research will use standard parameter values for DE, as defined by Table 7-5.

Parameter	Abbreviation	Value
Mutation	Mu	0.75
Dithering	Di	0.5
Population Size	PSi	15
Crossover Rate	CR	0.7

Table 7-5, Control Parameter Values

10 runs optimising the agreement and disagreement weights for the PEOHH all yielded the same value of *PPV* of 0.828. These results repeat the result from using wide search radius control parameters, indicating that this value is genuinely the global maximum, as opposed to the solver being stuck at a local maximum, such as an elevated crater in 28-dimensional space. Figure 7-12 shows that the optimised results represent a wide range of agreement and disagreement weight values. This wide range indicates uncertainty, caused again by the small size of the GSSLR, about what the optimum values are. This research predicts that this uncertainty will be drastically reduced by an innovation in

maintenance record keeping that will be discussed in chapter 9. The following sections will recommend values for these weights to be used by the PEOHH.

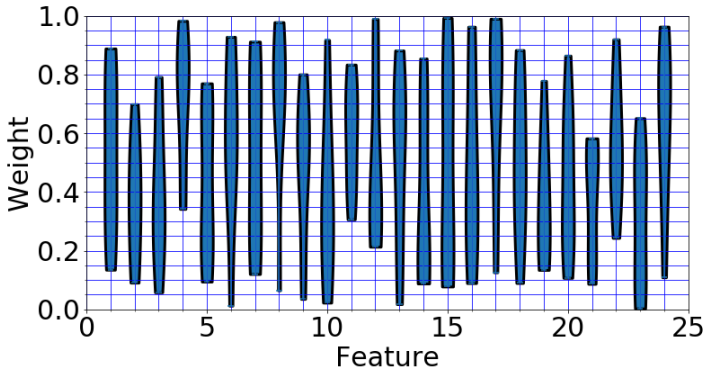


Figure 7-12, Vectors of Values of Agreement Weight and of the Negative of Disagreement Weight that all Yield Optimum Results

7.6.2 Positive Predictive Value

Section 7.6.1 optimised the weights and thresholds to the extent achievable with a GSSLR of the size available. It yielded the maximum value of *PPV* of 0.828. This value of *PPV* was calculated using a small sample of the health history and will therefore be over fitted to the data. To measure a more accurate value of *PPV*, this research split the GSSLR into training and testing data sets.

Overfitting is an error that occurs when a theoretical model is too closely fitted to a small set of data points. Rather than optimising the weights and thresholds by training on the whole GSSLR, training on a sample of the already small GSSLR increases the amount of overfitting to the data. This research will then then calculate the *PPV* of the remainder of the GSSLR that was not used for training. Overfitting to the training data means that this technique will under-estimate *PPV* against the test data.

This research investigated the effect of varying the Size of the Test data sample (*TeS*) between 0.1 and 0.5 of the GSSLR. It used for training all of the data in the GSSLR that will not be used for testing, so the Size of the Training data sample (*TrS*) = 1 – *TeS*, that is that this research varied *TrS* between 0.9 and 0.5. This research repeated the optimisation of *PPV* 5 times for each value of *TeS* using DE. This technique is an example of bootstrapping, which refers to a test that uses random sampling with replacement.

Figure 7-13 presents *PPV* against *TeS* for (a) training and (b) testing as a violin plot which shows the distribution and median value for *PPV*. Figure 7-13 (b) shows that the median *PPV* for testing does not vary with *TeS*. It might be expected that overfitting would cause *PPV* to tend to increase for training and reduce for testing with a smaller *TrS*, that is for larger *TeS*, but that did not happen, showing that the amount of overfitting does not change significantly within this range of *TeS*.

The median value of *PPV* for training is 0.889, compared to 0.571 for testing. The higher value of *PPV* for training is due to overfitting. It follows that the values of *PPV* for testing will be under fitted to the data. Each value of *PPV* calculated here is a point estimate of the *PPV of the EHH*. 0.571 is therefore a low point estimate of it. The next chapter will look at whether the health history has been enriched and will include discussion of the *PPV*.

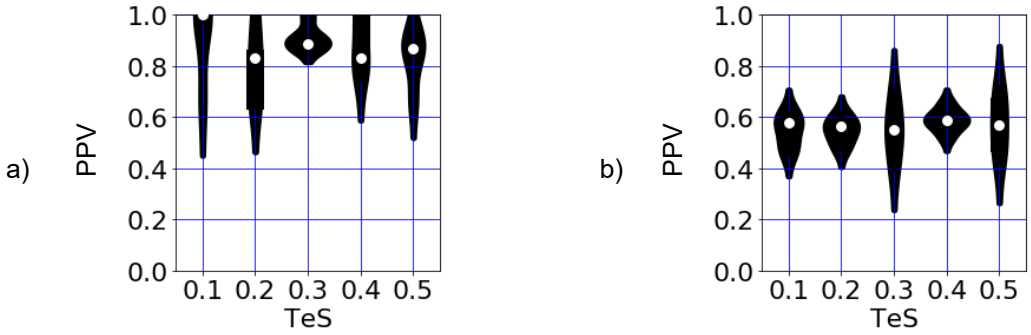


Figure 7-13, Positive Predictive Value (PPV) against Test Size (TeS) for (a) Training and (b) Testing

This research will use the 95th percentile of the distribution of the *PPV* of the GSSLR to estimate the 95% CI of the *PPV* of the EHH. It is (0.667, 1) for training and (0.416, 0.706) for testing and so this research estimates the 95% CI of the *PPV* of the EHH as (0.416, 0.706). This estimation did not require the techniques for interval estimation of a binomial proportion described in section 6.1 because it has instead used the distribution of results from repeated trials.

7.6.3 Negational Positive Predictive Value

This section will now investigate the relative effectiveness of each of the techniques for the enrichment of the health history that were presented in section 4.3. It will use all of the GSSLR to optimise the weights and thresholds, investigating the effect of setting some of the agreement and disagreement weights to zero. If, when the agreement and disagreement weights for a feature are set to zero, the maximum value of $\hat{p}$ yielded is not lower than with them varied, then that feature is not contributing anything and therefore health history enrichment can be done without it without any drop in quality. The small size of the GSSLR means that the following results are uncertain but they are advice based on the best evidence available. Further work could increase this work with a larger GSSLR to find whether the techniques that seem from these results to not add quality do actually add quality to the PEOHH.

Section 3.2.3 presented the Negational Positive Predictive Value (*NPPV*):

> Where the weight assigned to a dimension of the comparison vector is set to zero, the corresponding feature is disregarded for record linkage comparison. This thesis will refer to setting the weight to zero as '*disregard'.* Consider disregarding one of the features: the effect of not using a feature is a useful measure of the effectiveness of that feature. This thesis presents the Negational Positive Predictive Value (*NPPV*); the Positive Predictive Value (*PPV*) calculated using all the features except one feature or set of features that is disregarded.

Table 7-6 presents optimised values of *NPPV.* When the PEOHH used all the features, that is that none of them were disregarded, DE yielded the maximum value of 0.828. DE also yielded this maximum value when the start time feature, defined in section 4.4.1, the type feature, defined in section 4.4.2, the visits feature, defined in section 4.4.3 or all but one of the failure mode features defined in section 4.4.4 was disregarded. When the PEOHH did not use each of the other features, DE yielded lower values of *NPPV*. This result indicates that those features where DE yielded a lower value of $\widehat{p_N}$ when they were disregarded, that is the finish, notification and part posting time features defined in section 4.4.1, may tend to be the most important features for wind turbine maintenance record linkage.

Set of Features		Feature	*NPPV*
All features present			0.828
Time		w/o Start	0.828
		w/o Finish	0.793
		w/o Notification	0.724
		w/o Part Posting	0.759
w/o Type			0.828
w/o Visits			0.828
Failure Mode	Outage	w/o Description	0.828
		w/o Alarm Code	0.828
		w/o Parts	0.793
	Alarms	w/o Description	0.828
		w/o Alarm Code	0.828
		w/o Parts	0.828

Table 7-6, Optimised Values of NPPV, Calculated Disregarding Selected Features

The next step, after disregarding individual features, is to disregard selected sets of features. It would have been of interest to find the optimum *PPV* for each permutation of features. For 12 features, this would be 2^{12} = 4096 permutations which would unfortunately have taken too long.

Table 7.7 presents optimised values of *NPPV* calculated disregarding selected sets of features. When the PEOHH used all the features, that is that none of them was disregarded, DE yielded the maximum value of $\widehat{p_N}$ of 0.828. DE also yielded this maximum value when the type feature defined in section 4.4.2 and the visits feature defined in section 4.4.3 were disregarded, when all the failure mode

features defined in section 4.4.4 were disregarded or when all the failure mode features and the start time feature defined in section 4.4.1 were disregarded. When the PEOHH disregarded each of the other sets of features listed, DE yielded lower values of *NPPV*.

The failure mode features are relatively computationally expensive and so it might be inferred that, if there is no evidence of them improving the quality of record linkage, unnecessary computation should be avoided by not using them. The next section will consider the limits of that conclusion.

This result indicates that wind turbine maintenance record linkage should use an ensemble of features made up of the finish, notification and part posting time features, defined in section 4.3.1, the type feature, defined in section 4.4.2 and the visits feature, defined in section 4.4.3, and that it should avoid unnecessary computation by not using the failure mode features, defined in section 4.4.4 and the start time feature, defined in section 4.4.1.

Sets of Features	*NPPV*
All features present	0.828
w/o Type and Visits	0.828
w/o Failure Mode	0.828
w/o Failure Mode and Start	0.828
w/o Failure Mode and Finish	0.793
w/o Failure Mode and Type	0.793
w/o Failure Mode, Type and Visit	0.793
w/o Failure Mode and Part Posting	0.759
w/o Failure Mode and Visits	0.759
w/o Failure Mode, Finish and Part Posting	0.759
w/o Failure Mode, Type and Finish	0.759
w/o Failure Mode and Notification	0.724
w/o Failure Mode, Type and Part Posting	0.724
w/o Time	0.517

Table 7-7, Optimised Values of NPPV, Calculated Disregarding Selected Sets of Features

There are two sources of uncertainty in this result.

Firstly, the *PPV* of the GSSLR is a point estimate of the *PPV* of the EHH. A confidence interval (CI) is a region with a specified, nominal probability that it contains a feature of interest. The small size of the GSSLR, only 29 POLRs, means that, for a *PPV* of the GSSLR of 0.828, the 95% CI of the *PPV* of the EHH, identified using the Wilson interval which was discussed in chapter 6, is (0.655, 0.924). The previous section recognised that this is an over estimate of PPV caused by over fitting and that a better estimate of the *PPV* of the EHH is 0.571 with a 95% CI of (0.416, 0.706). The significance of the CI of the over fitted *PPV* is that it quantifies the uncertainty of the optimisation presented in this section.

Secondly, DE does not always find the optimum vector. As the number of dimensions increases, the optimisation problem becomes more difficult as the space required to compute solutions increases exponentially with the number of dimensions. This is referred to in the literature as the 'curse of dimensionality' (Bellman, 1956, Rust 1997). The more features the PEOHH disregards, the easier the optimisation problem becomes. The following section will consider the uncertainty of the optimisation presented here and will conclude from this and from further investigation that all of the features should be used for record linkage.

Table 7.8 presents the optimised weights and thresholds. Table 7.8 does not record optimised values for the time Threshold for the start time Feature (Th_{St}), the description similarity threshold (Th_{De}) or the parts score threshold (Th_{Pa}) because, with the agreement and disagreement weights for these features set to zero, these thresholds are meaningless.

Where the PEOHH uses the weights and thresholds set out in Table 7.8 the sums of the columns equal the limits of possible values of Score for each POLR (S_{POLR}). The sum of the column of agreement weights is 3.127. The sum of the column of disagreement weights is -2.643.

The PEOHH yields its best performance with the weights and thresholds set to the values in Table 7-8, however, the following section will present some alternative weights and thresholds at which, to the measurement accuracy yielded with the size of the GSSLR that is available, the PEOHH yields an equally good performance but that this thesis advises should be used instead.

Set of Features		Feature	Agree (*AW*)	Disagree (*DW*)	Threshold (days)
Time		Start	0	0	
		Finish	0.995	-0.602	2.534
		Notification	0.588	-0.088	1.312
		Part Posting	0.668	-0.846	2.260
Type			0.095	-0.280	
Visit			0.781	-0.827	
Failure Mode	Outage	Description	0	0	
		Alarm Code	0	0	
		Parts	0	0	
	Alarms	Description	0	0	
		Alarm Code	0	0	
		Parts	0	0	
Sum			3.127	-2.643	

Table 7-8, Optimised Weights and Thresholds

7.6.4 Using all the Features

The previous section tested techniques using an ensemble of features and found that if any of the failure mode based techniques has a useful effect then it is too subtle to be measured with a GSSLR of the size available. This section will use a different validation approach to investigate whether it might be advantageous to use all of the features, rather than disregarding the computationally expensive failure mode based features.

This section investigates the effect of using all the features, dividing them into two sets; the minor features; those that are indicative of the failure mode and the WO start time feature, and the major features; the remainder of the features. It gives the major features their optimum values from the

previous section. It gives all the Minor Features the agreement Weight We_{MF} and a disagreement weight of negative We_{MF}.

Using all of the features is more computationally expensive than using only the major features, however, because section 1.2 showed that an EHH is valuable to a wind farm developer, if it does yield an even slightly better quality of record linkage then the more computationally expensive approach might be justified.

This research investigated the effect of varying We_{MF} between zero and 1. A We_{MF} of zero yields the agreement and disagreement weights detailed in Figure 7-14 in which a selection of features have their weights set to zero and so are effectively not used. Where We_{MF} is not zero, all the features are used by the PEOHH with the weights that in that figure are set to zero replaced by We_{MF}. The results are shown in Figure 7-14.

The maximum *PPV* in figure Figure 7-14 (0.828) occurs throughout the range We_{MF} = zero to 0.33. With a GSSLR of the size available, the process for the validation of the EHH cannot distinguish between these values of We_{MF}. That means that all the features can be used within this range with no measurable drop in record linkage quality. This section will consider the uncertainty of this recommendation and will report the results of a small additional test of what happens to the EHH when the PEOHH uses a We_{MF} of 0.1. It will find that using a We_{MF} of 0.1 would most likely improve the quality of record linkage, which is consistent with the figure.

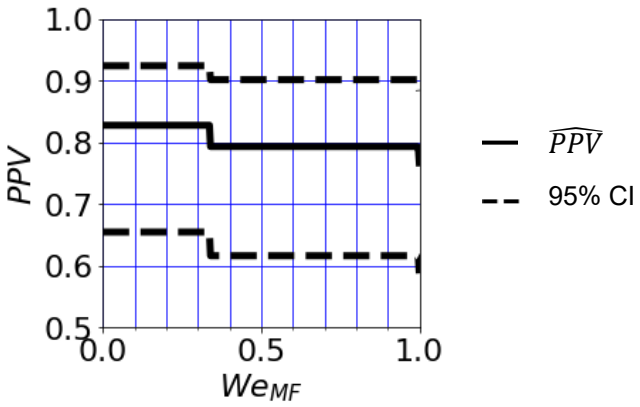


Figure 7-14, Positive Predictive Value (PPV) and 95% CI against Minor Feature Weight (We_{MF})

Figure 7-14 shows the 95% CI for the *PPV* of the EHH identified using the Wilson interval which was discussed in section 6.1.3. The small size of the GSSLR, only 29 POLRs, means that, for a *PPV* of the GSSLR of 0.828, the 95% CI of the *PPV* of the EHH is (0.655, 0.924).

Again, this section will use TUB, presented in section 6.2, for constructing a CI for the change in the *PPV* of the EHH. Table 7.9 shows the effect of varying We_{MF}. '*True'* is the number of records from the GSSLR in which the version of the health history linked the WO to the same outage as that in the GSSLR and '*False'* is the number in which it linked the WO to a different outage. TUB uses these results to estimate the uncertainty of the difference that it makes.

		We_{MF} = 0	
		True	False
We_{MF} = 1	True	23	0
	False	1	5

Table 7-9, Effect on the Quality of Record Linkage of Varying the Minor Feature Weight (We_{MF})

Equation 7.9 defines the Increase in the *PPV* of the EHH that would by yielded by a We_{MF} of zero over that yielded by a We_{MF} of 1 (${}^{PPV}_{We_{MF}}I^{0}_{1}$). It uses the Probability distribution of estimates of the *PPV* of the EHH that would be yielded by a We_{MF} of zero $(P(PPV \mid We_{MF} = 0))$ and the same distribution that would be yielded by a We_{MF} of 1 $(P(PPV \mid We_{MF} = 1))$.

$$ {}^{PPV}_{We_{MF}}I^{0}_{1} = \frac{P(PPV \mid We_{MF} = 0)}{P(PPV \mid We_{MF} = 1)} - 1 \tag{7.9} $$

Equation 7.10 evaluates ${}^{\widehat{PPV}}_{We_{MF}}I^{0}_{1}$, a point estimate of ${}^{PPV}_{We_{MF}}I^{0}_{1}$, using the data from Table 7-9.

$$ {}^{\widehat{PPV}}_{We_{MF}}I^{0}_{1} = \frac{23+1}{23+0} - 1 = 4\% \tag{7.10} $$

This research used TUB to estimate the uncertainty of ${}^{PPV}_{We_{MF}}I^{0}_{1}$, which yielded a 95% CI of (0, 16%). Again, this research predicts that this uncertainty will be drastically reduced by an upcoming innovation in maintenance record keeping that will be discussed in chapter 9.

To better understand this uncertainty, this research investigated the distribution of TUBs list of estimates of ${}_{We_{MF}}^{PPV} I_{1}^{0}$ using techniques presented in section 6.3. The results are shown in Figure 7-15(a). The most important features of this distribution are the proportion of negative estimates and the proportion of zero estimates because these indicate respectively that the quality of record linkage has reduced and that it has not increased, as detailed in Figure 7-15(b). The probability that a We_{MF} of zero would yield a lower *PPV* of the EHH than a We_{MF} of 1 is estimated by the proportion of bootstrap samples for which ${}_{We_{MF}}^{PPV} I_{1}^{0} \leq 0$, which this research found to be 36%. This indicates that there is a significant likelihood that reducing We_{MF}, from 1 to zero would not increase the quality of record linkage. There is a 63% chance that such a reduction would increase the quality of record linkage.

Figure 7-15(b) also shows that the proportion of bootstrap samples for which ${}_{We_{MF}}^{PPV} I_{1}^{0} < 0$ is 0. As explained in section 6.3, bootstrapping is conditional on the original sample, but this indicates that there is little likelihood that reducing We_{MF} from 1 to zero would increase the quality of record linkage.

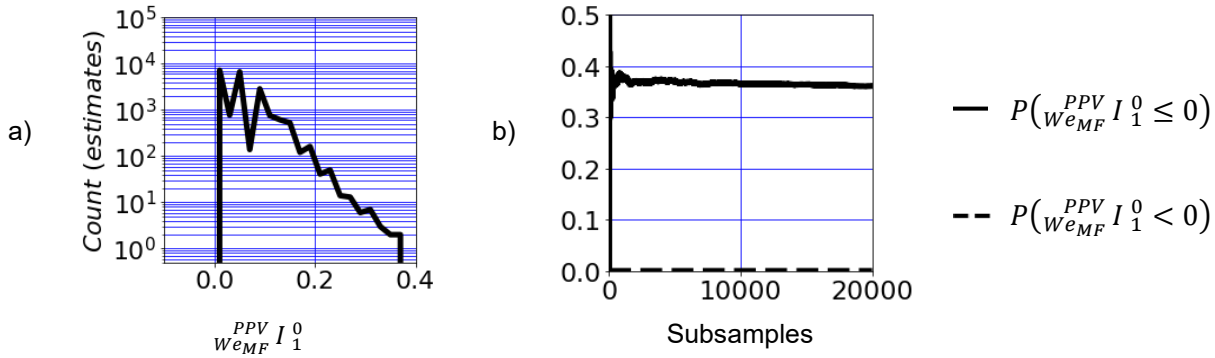


Figure 7-15, (a) Frequency of ${}_{We_{MF}}^{PPV} I_{1}^{0}$ for 20,000 Subsamples, (b) Proportion of Bootstrap Samples for which ${}_{We_{MF}}^{PPV} I_{1}^{0} \leq 0$ and for which ${}_{We_{MF}}^{PPV} I_{1}^{0} < 0$ against Number of Subsamples

This section presented the effect of varying We_{MF} and identified that this effect is too subtle to measure accurately with a GSSLR of the size available.

This section will test what happens to the EHH when the PEOHH uses a We_{MF} of 0.1. It will compare that version of the EHH (EHH 1) to the version of the EHH that is yielded when the baseline We_{MF} value from the previous section of zero is used (EHH 0). When compared to EHH 0, out of 9820 WOs, EHH 1 has 74 different POLRs. This section will refer to this set of 74 WOs that link to Different outages in EHH 1 to those that they link to in EHH zero as DiWO. Using the whole EHH rather than

just the GSSLR gives much more detail about such relatively subtle differences in results because it is the whole population of the health history data rather than being a sample from it.

If there is any benefit in using all the features then it is most likely to derive from those POLR that yielded the highest quality of record linkage, where *TP*s are more common. This section therefore selects for investigation the best WO from DiWO. It selects as a sample the WO from DiWO that yields the highest S_{POLR} when We_{MF} = *0.1*. This sampling method is systematic, not random, and of course the sample is not representative of the population. In EHH 1 the example WO is linked to an outage with the alarm code description of 'Yaw hydraulic oil level low' while in EHH 2 it is 'Manual idle stop - yawing'. The WO description is identical to the outage alarm code description in EHH 1; 'Yaw hydraulic oil level low'. This authors experience of the data and of participation in the validation meetings is that, in this case, EHH 1 has obviously yielded a higher quality record linkage result than EHH 0.

In this one systematically selected case, EHH 1 has outperformed EHH 0, that is that We_{MF} = *0.1* has outperformed We_{MF} = *0*. This thesis therefore recommends that the PEOHH most likely yields its best performance with the weights and thresholds set to the values in Table 7.10.

Set of Features		Feature	Agree (*AW*)	Disagree (*DW*)	Threshold (days)
Time		Start	0.100	0.100	2.000
		Finish	0.995	-0.602	2.534
		Notification	0.588	-0.088	1.312
		Part Posting	0.668	-0.846	2.260
Type			0.095	-0.280	
Visit			0.781	-0.827	
Failure Mode	Outage	Description	0.100	-0.100	
		Alarm Code	0.100	-0.100	
		Parts	0.100	-0.100	
	Alarms	Description	0.100	-0.100	
		Alarm Code	0.100	-0.100	
		Parts	0.100	-0.100	

Table 7-10, Recommended Weights and Thresholds

7.6.5 Simplifying the Parts Frequency Techniques

The parts frequency techniques presented in this thesis are computationally expensive. For a single farm, it took 1 day to calculate the Parts Score for Outages (PS_O) for each POLR and 2 weeks to calculate the Parts Score for Alarms (PS_A) for each POLR. PS_A takes longer to calculate than PS_O because there can be multiple alarms linked to an outage.

This technique is computationally expensive because it uses as training data all the records except for the one being tested; each time a record is tested, the techniques generate a new set of training data from all the other records and this requires additional computation.

A less computationally expensive alternative would be to split the data in two; set A and set B. Set A could be tested on a model trained using set B and set B could be tested on a model trained using set A. That simpler and quicker technique was developed as part of this research. The results are not presented because the technique used by the PEOHH is more comprehensive and, despite it being trained on more data, its effect could not be measured with a GSSLR of the size available.

7.7 Conclusion to the Optimisation of the Weights and Thresholds

This chapter has presented the results from the validation of the PEOHH. It has presented the effect on the quality of record linkage of varying the weights and thresholds that the PEOHH uses. It has presented the optimisation of the agreement and disagreement weights (defined in section 4.1) for each feature (defined in section 4.4).

Chapter 4 presented the PEOHH. It presented 12 novel record linkage techniques that can be used as part of the PEOHH. This chapter presented the result that an ensemble of five of these techniques yielded an optimal Enriched Health History (EHH). The identification of which outage to match to each WO was achieved to the same quality whether using the full ensemble of twelve techniques or using all of them except for the six techniques indicative of the failure mode. The six techniques using the failure mode could be disregarded with the PEOHH still yielding an optimum quality of record linkage. This result indicates that if any of the failure mode-based features has a useful effect then it is too subtle to be measured with a GSSLR of the size available. This result is counter-intuitive because the manual methods that are frequently and consistently used by practitioners for linking WOs to outages do make use of the WO *description* field, one of the features indicative of the failure mode. Practitioners do not tend to use the WO *finish date*, the *notification date* or the *part posting date* for manual record linkage and this result indicates that these features can substitute for the features indicative of the failure mode.

The size of the GSSLR, only 29 POLRs, was constrained by the amount of expert time that was available. This chapter quantified the consequent uncertainty of its estimate of the *PPV* of the EHH. It reported considerable uncertainty as to the optimum values of the weights and thresholds. This research predicts that this uncertainty will be drastically reduced by an innovation in maintenance record keeping that will be discussed in chapter 9; the automatic linking of new WOs to outages will result in a larger GSSLR and consequently more certainty in this estimate. Given that uncertainty, this research advises using all the features rather than disregarding those features that use the failure mode.

Operators should initially use the weights and thresholds for the EHH detailed in Table 7-10 but should re-assess this when a larger GSSLR becomes available.

Chapter 8 will assess the EHH's usefulness.

8 Results: Has the Health History been Enriched?

To investigate whether the Process for the Enrichment of wind turbine Health History (PEOHH) developed in this research does what it is for, that is to enrich health history, this chapter will address RQ3:

RQ3 How can the richness of historical data on wind turbine health be measured?

This chapter will present four measures of the richness of historical data on wind turbine health. They will be defined later in this section but will be referred to as the number of WO records, the number of MLI records, the number of MLIs per alarm code and the prevalence of POLRs in the EHH where the health history would recommend all the required parts.

Section 8.1.2 will define those WOs already labelled with an alarm code as the unenriched health history. This chapter will measure the richness of both the unenriched and the enriched health history and compare them to quantify how much enrichment has been achieved. It will then measure the effect of enrichment in a more practical way by assessing the EHHs usefulness for the application of troubleshooting, described in section 1.2.4. If the EHH has become more useful for troubleshooting then it can be claimed that the health history has been enriched. This research selected troubleshooting as the application to consider because health history enrichment will have a very direct impact on advising which parts might be required to repair each fault. This research expects that a health history that has, by enrichment, become more useful for the application of troubleshooting will also be more useful for the applications of maintenance scheduling, described in section 1.2.2, Condition Based Maintenance (CBM), described in section 1.2.3 and for the measurement of maintenance effectiveness, described in section 1.2.5. This expectation is based on extensive discussions with wind turbine experts conducted as part of this research but could be tested by further work, applying the EHH in practice.

8.1 What is Enrichment?

By enrichment, this thesis means enhancing the utility of the EHH for maintenance, to maximise productivity. The following sections will consider how operators might quantify the impact of the application of the EHH by defining some richness metrics; increases in these metrics imply enrichment.

Operators try to quantify the impact of any innovations to their maintenance practice and they use such measures to assess the value added as a return on their investment in innovation. The measurement of productivity, defined in section 1.2.1, before and after the implementation of an innovation is not a useful measure because there are too many stochastic variables to account for, such as wear out failures, human error and changing weather. The same would be true of measuring the cost of energy, capacity factor, *time-based availability* or *OPEX*, which are all also defined in section 1.2.1, or any other overall metric. All the applications of the EHH that this research has identified are designed to help maximise productivity, but its impact on each application would be measured by a different technique.

The following section will propose richness metrics that would be available after the implementation of the PEOHH. Section 8.1.2 will review the data that are available to this thesis for the identification of richness metrics.

8.1.1 Further Work on Richness Measurement

Section 1.2 identified 4 potential applications for the EHH and these could be used to identify richness metrics:

- Section 1.2.2 reviewed maintenance scheduling. Operators could assess the value added by implementing the application of the EHH to maintenance scheduling. They could record their inputs to their decision support tools prior to and after enrichment and estimate the consequent change in productivity.
- Section 1.2.3 reviewed CBM. Operators could assess the value added by implementing the application of the EHH to using the notifications that prognostic models generate for each farm on which they are implemented. These notifications are used to initiate Work Orders (WO) and it is this remedial work that creates value by avoiding faults that can cause power outages. The value of a notification is the product of the number of notifications that it will create and the average productivity of those notifications. After a model has been implemented the estimated productivity is compared to the productivity that it did create by looking at how many of the notifications were used to create WOs.
- Section 1.2.4 reviewed troubleshooting. Operators could assess the value added by implementing the application of the EHH to troubleshooting by identifying those maintenance

operations where the required part was not brought. They could count each time that the troubleshooting guide consulted did not list the correct part. It would be possible to use information on lost production, linked to WOs by the PEOHH, to calculate the value added by this work.

- Section 1.2.5 reviewed the measurement of maintenance effectiveness. Operators could assess the value added by implementing the application of the EHH to the measurement of maintenance effectiveness. An EHH of the machinery under study could be used by engineers and data scientists to develop more accurate measures of maintenance effectiveness and a better understanding of the confidence of these measures. They could then estimate the value added by these innovations.

The applications of the EHH described above have not been implemented yet but this research advises that such implementation is the most immediate further work that it has enabled. This chapter will assess the richness of the health history before and after enrichment.

8.1.2 Richness Data Review

This section will review the data that are available to this thesis for the identification of richness metrics.

Section 1.2.4 reviewed troubleshooting on offshore wind turbines:

> "Data scientists analyse failure histories and derive troubleshooting guides from them. These guides are key to the technicians' diagnosis of faults, alongside their expert knowledge. Technicians refer to a trouble shooting guide for advice on how to diagnose the failure mode and the repair activity that is most likely to be effective… Each troubleshooting guide applies to a single alarm code, rating and manufacturer."

Section 2.1.3.6 presented the alarm code data in the database WOs:

> "Some WOs are labelled with an alarm code. This label is used to identify the failure mode."

Section 2.1.2 presented Ørsted's database of outages:

> "Ørsted label each outage with an alarm code indicative of the failure mode."

When Ørsted produce their troubleshooting guides they use those WOs that are already labelled with an alarm code to identify which parts might be required to repair each failure mode. This chapter will use those WOs already labelled with an alarm code as the unenriched health history and will use them to measure the richness of the unenriched health history. By joining each WO to an outage, the

PEOHH links each WO with an alarm code. This chapter will use all the WOs in the farm's history, linked to outages by the PEOHH, to measure the richness of the EHH. This will inform a comparison of the enriched health history with the unenriched health history.

Section 2.2 presented existing record linkage techniques:

> "Probabilistic record linkage techniques join two databases together to create a new database in which each row represents one Pair Of Linked Records (POLR)."

Section 4.1 presented the PEOHH:

> "The PEOHH calculates a score for each POLR (S_{POLR}), the sum across all the features of the Weight for that Feature and for that POLR ($W_{Fe_{POLR}}$) using equation 4.2."

$$S_{POLR} = \sum_{Fe} W_{Fe_{POLR}} \tag{4.2}$$

Section 4.3 presented the method for the validation of the PEOHH:

> "To identify the GSSLR, this research randomly selected a set of corrective WOs from the database of WOs. Random selection was used with the intention of getting a sample representative of the population."

This research filtered the EHH and the GSSLR to only include POLRs with an S_{POLR} above a Score Threshold (Th_S) or that are already labelled with an alarm code. A Th_S greater than the maximum possible value of S_{POLR} excludes all the linked records from the filtered EHH, leaving only those WOs already labelled with an alarm code, and yields the unenriched health history. A Th_S less than the minimum possible value of S_{POLR} excludes none of the linked records from the filtered EHH and yields the full EHH. Varying Th_S will enable this chapter to compare the unenriched health history with the enriched health history and with the partially enriched health history.

Table 7-8 presented the optimised weights and thresholds that will be used in this chapter. It showed that the sum of the column of agreement weights was 3.127 and that the sum of the column of disagreement weights was -2.643. These values are the limits of possible values of Score for each POLR (S_{POLR}). Figure 8-1 presents the distribution of S_{POLR} in the EHH and in the GSSLR. It shows that the values of S_{POLR} are within this range. A Th_S of 4 yields the unenriched health history and a Th_S of -4 yields the EHH.

The PEOHH calculates these values of S_{POLR} using values of the weights and thresholds optimised using the GSSLR. Please recall that overfitting is an error that occurs when a theoretical model is too

closely fitted to a small set of data points. Overfitting would be expected to cause higher values of S_{POLR} in the GSSLR than in the EHH. Figure 8-1 shows that the distributions are broadly similar, each having a mode at 0.5, but that the GSSLR is more weighted towards higher values indicating that there is some overfitting. Further work with a larger GSSLR would reduce the amount of over fitting but, as the distributions are broadly similar, valid conclusions can be based on the data used in this thesis.

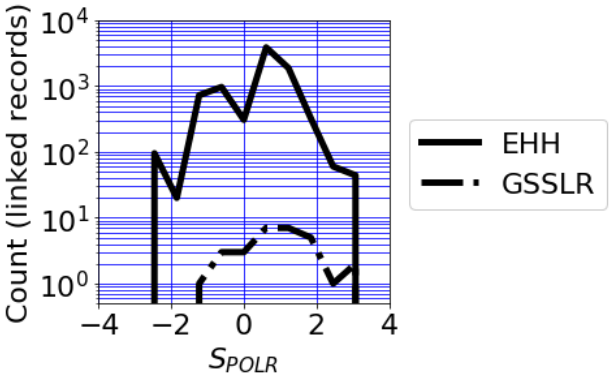


Figure 8-1, Frequency of the Score for each POLR (S_{POLR}) in the EHH and in the GSSLR

A higher Th_S means that the criterion for recommending parts is more stringent. A lower Th_S means that the criterion for recommending parts is less stringent. It would recommend parts that might be needed but this would be less likely. A Th_S between these extremes yields a partially enriched health history. This technique models the effect of filtering the EHH to only include higher scoring POLRs to increase the quality of the EHH. If the number of POLRs in the EHH is its *quantity* and the minimum S_{POLR} in the EHH is its *quality* then, with a high Th_S, the *quality* of the EHH is higher but its *quantity* is lower. This section will present the results of filtering the EHH to only include higher scoring POLRs to increase the quality of the EHH.

The optimum value of Th_S depends on the application of the EHH, where some applications benefit from a higher quality EHH and other applications benefit from a larger quantity of EHH data. For example, it is important to operators that maintenance technicians have confidence in the analytic tools that are made available to them. An analytic tool that recommended that technicians take parts that they would be unlikely to need might not win the confidence of technicians. In that case, a higher value of Th_S would help to win the confidence of the technicians. On the other hand, if the tool were well understood by the technicians then they might choose to set a lower value of Th_S so as to bring parts that, while unlikely, might turn out to be required for repairing their OWTs. Section 1.2.1 defines

the productivity of OWT maintenance. Variations in technicians' confidence in the troubleshooting guides would have the consequence that the value of Th_S that can be expected to yield the most productive troubleshooting varies from site to site.

This section has defined enrichment as enhancing the utility of the EHH for maintenance, so as to maximise productivity. It has introduced a trade-off between *quality* and *quantity* any it has explained how this trade-off has different effects on productivity at different sites. Section 8.2 will measure the quality of record linkage and sections 8.3 to 8.5 will present four measures of the richness of the EHH. These measures will illustrate a trade-off between *quality* and *quantity*. This chapter will demonstrate that the wind turbine health history has been substantially enriched.

8.2 Quality of Record Linkage

The quality of record linkage is the likelihood that each WO is linked to its corresponding outage. Section 2.2 presented a record linkage quality metric:

> Positive Predictive Value (*PPV*), defined by equation 2.2, measures the proportion of classified matches that are correctly identified as such.

$$PPV = \frac{TP}{TP + FP} \tag{2.2}$$

This research uses the Process for the Validation of the EHH (PVEHH)[71] for assessing the quality of record linkage. The PVEHH identifies the *PPV* of the EHH and of the unenriched health history by the comparison of these linked records with the GSSLR. Any errors in the GSSLR would therefore lead to errors in the estimation of the *PPV* of the linked records. This research considers the risk of systematic errors to be insignificant because the GSSLR was identified by one wind turbine expert, was checked by another wind turbine expert and was then re-checked as part of this research.

This research investigated the effect on record linkage quality of varying the Score Threshold (Th_S). The results are shown in Figure 8-2(a) which shows the *PPV* of the GSSLR ($\widehat{PPV}$) and the 95% confidence interval (CI) of the *PPV* of the EHH against Score Threshold (Th_S). A higher value of Th_S yields a higher value of $\widehat{PPV}$; that is a higher quality of health history. Increasing Th_S increases $\widehat{PPV}$ from 0.828 to 1. A the $\widehat{PPV}$ of 1 means that all the POLRs in the GSSLR are true matches, indicating a high prevalence of true matches in the EHH, that is a high quality EHH.

Figure 8-2 shows the 95% confidence interval (CI) of the *PPV* of the EHH, calculated using the Wilson interval which was presented in section 6.1. A higher value of Th_S yields a greater CI, showing reduced confidence caused by the smaller number of samples remaining in the filtered GSSLR.

In Figure 8-2, values of Th_S above 3 yield no results as the number of samples remaining in the filtered GSSLR is reduced to zero. Figure 8-2(a) presents *PPV* calculated using the same data, the GSSLR, for training and for testing. These results are over fitted to the data, yielding an over-estimate of the *PPV* of the EHH. Section 7.6.2 presented the result, based on training and testing on different samples from the GSSLR, that 0.571 is a more realistic estimate of *PPV*. Figure 8-2(a) uses the optimised values of the PEOHH's weights and thresholds which yield the over fitted estimate of *PPV*. Figure 8-2 (b) calculates values of *PPV* using the number of true matches adjusted by an OverFitting

[71] Section 3.1 presented the PVEHH.

Factor (*OFF*); *OFF* = 0.571 / 0.828 = 0.690. Figure 8-2(b) presents a more representative estimate of *PPV*.

Figure 8-2 shows the effect of varying Th_S on the quality of the EHH. A higher value of Th_S yields a higher point estimate of *PPV* but a lower estimate of the lower bound of the 95% CI of *PPV*; that is a higher quality EHH but a lower confidence in that quality.

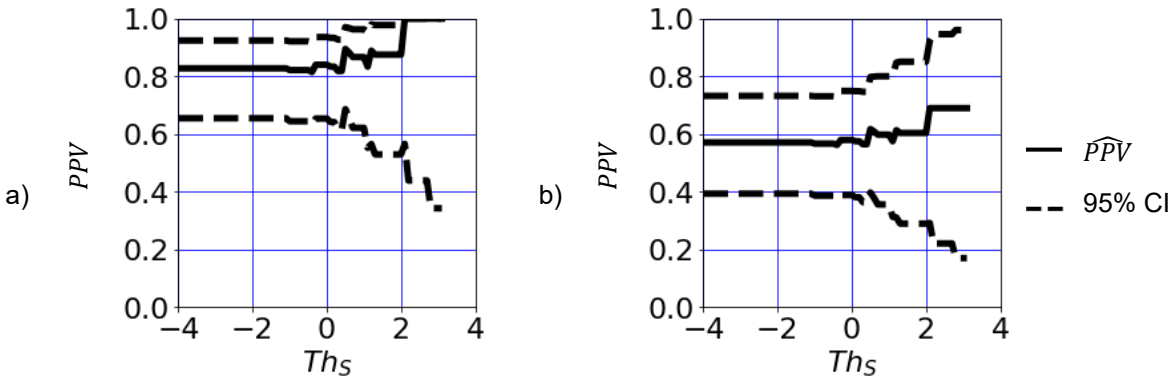


Figure 8-2, Positive Predictive Value (PPV) and 95% CI against Score Threshold (Th_S)
(a) Raw and (b) Adjusted by the OverFitting Factor (OFF)

8.3 Number of WO and of Material Line Item Records

This section will use the database of material consumption, described in section 2.1.4. It was described in that section that:

> "The material consumption database lists what parts were used in the maintenance of the OWTs. Each Material consumption Line Item (MLI) refers to a single part number and is assigned to an order number. Some WOs have no material consumption line items assigned to them while others have many. Materials include replacement parts as well as consumables such as oil, grease or paint."

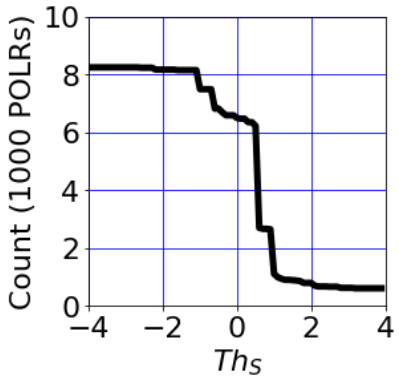
This research investigated the effect of varying Th_S through a full range of values from -4 (enriched) to 4 (unenriched). The results are shown in Figure 8-3.

Figure 8.3 (a) shows the count of POLRs in the EHH against Score Threshold (Th_S). It shows that there are 606 POLRs in the unenriched health history compared to 8246 in the EHH. Each POLR counted here represents one WO joined to an outage, that means, for the application for example of troubleshooting, that any material consumption associated with the WO is made available for the improvement of troubleshooting guides.

Figure 8-3 (b) shows the count of MLIs in the EHH against Th_S. It shows that there are 2642 MLIs in the unenriched health history compared to 13377 in the EHH. Each MLI counted here means, for the application for example of troubleshooting, that there is one more instance of a part number being associated with a specific failure mode. Figure 8.3 indicates that the health history has been enriched.

Figure 8-3(a) and (b) appear similar to each other because, as more POLRs are included in the filtered EHH, these are associated with more MLIs.

a)



b)

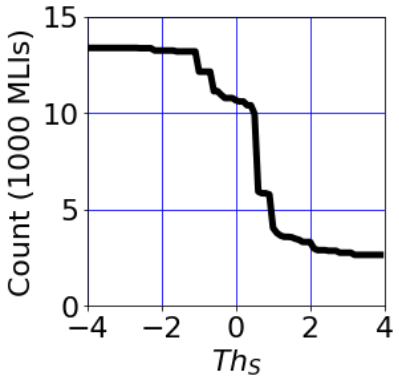


Figure 8-3, (a) Count of POLRs, (b) Count of Material Line Items (MLIs), in the EHH against Score Threshold (Th_S)

8.4 Number of Material Line Items per Alarm Code

Many maintenance activities require a specific set of replacement parts. Offshore wind farm operators decide which parts to take to each job, for example how to stock a Crew Transfer Vessel (CTV) for a day's maintenance work. Section 1.2 identified the typical capacity of a CTV as 2 to 3 t and this places an upper limit on the mass of spare parts that can be selected. They weigh such considerations against the risk of an extended interval of downtime caused by a required part not being available.

This section presents, as a measure of richness, the use of the number of spare parts associated with each alarm code. If more spare parts are associated with an alarm code then, when that alarm code is associated with a fault that requires a visit to a wind turbine to repair it, the repair technicians can bring that larger set of spare parts to the wind turbine to affect a repair. If that larger list of spare parts includes some parts that could need replacing to repair the fault then the likelihood that the technicians have brought the parts necessary to repair the wind turbine is improved. Health history enrichment can help to avoid costly power outages by better informing the decision of what to load.

This section uses the median number of parts per alarm code as a measure of the richness of the EHH. It uses the median rather than the mean so as to avoid placing undue weight on outliers. It also identifies the effect of health history enrichment on outlying alarm codes using percentiles of the number of parts per alarm code.

This research investigated the effect of varying Th_S through a full range of values from -4 (enriched) to 4 (unenriched). The results are shown in Figure 8-4 and Figure 8-5. Figure 8-4 shows percentiles of the number of MLIs in each Alarm Code in the EHH against Th_S while Figure 8-5 shows the distribution of the number of MLIs for each alarm code in the EHH by Th_S. This is another way of looking at the results presented in Figure 8-4 and the findings from the two figures are consistent.

Both figures show that there are more MLIs at lower values of Th_S. Each MLI counted here means, for the application for example of troubleshooting, that there is one more instance of a part number being associated with a specific failure mode and therefore that the health history has been enriched for that failure mode. Figure 8-5shows that the distribution of the number of MLIs for each alarm code is similar at different values of Th_S which shows that the PEOHH enriches both those failure modes that require a wide range of parts to repair them and those that require fewer.

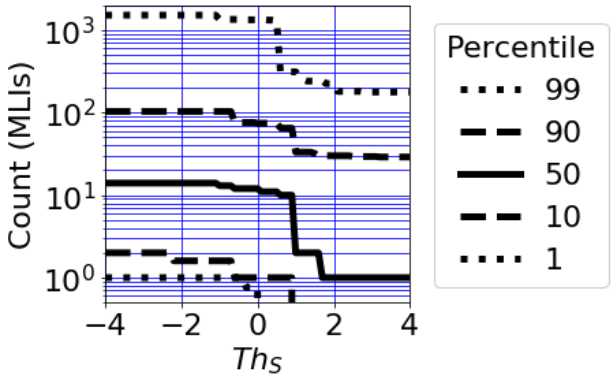


Figure 8-4, Percentiles of the number of Material Line Items (MLI) in each Alarm Code in the EHH against Score Threshold (Th_S)

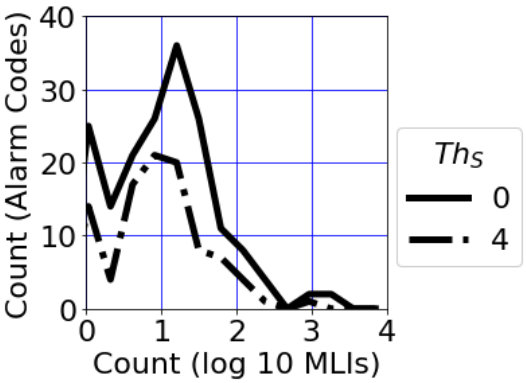


Figure 8-5, Frequency of the number of Material Line Items (MLI) for each Alarm Code in the EHH by Score Threshold (Th_S)

8.5 Application to Troubleshooting

This chapter has explained that, to effect a repair, the correct parts are required. This section will use the application of health history to OWT troubleshooting to assess whether the health history has been enriched.

Consider the practice of troubleshooting prior to the identification of the EHH. When an OWT fails, its maintenance team use the alarm log to make a preliminary diagnosis of the failure mode. At this point, the alarm log may indicate more than one possible failure mode. The maintenance team consult their troubleshooting guide for each of these failure modes. Each troubleshooting guide includes a list of parts that it recommends be brought to the OWT in case one of them might be required to effect the repair. If the maintenance team visit an OWT but have not brought a part required to make the repair then they need to come back with the outstanding part. The OWT may be out of availability until the required part is in place.

Now, consider the practice of troubleshooting after the identification of the EHH. Troubleshooting guides will be used in the same way but, alongside other potential improvements not considered in this section, they may now contain a more appropriate list of parts. This would reduce the risk that a required part would not be available at the OWT and would consequently reduce the risk that the outage would be extended unnecessarily.

This section will make the simplifying assumption that only one troubleshooting guide would be consulted for each fault. This assumption is not realistic, but it does simplify the analysis. Applying it consistently to both the unenriched health history and to the enriched health history is a valid comparison and thus a valid approach to the assessment of whether the health history has been enriched.

This section will analyse each WO in the GSSLR to identify whether all the parts that were used would have been brought according to a simulated troubleshooting guide based on the health history. It will use the GSSLR which should be representative of the corrective WOs from the selected wind farm since it was randomly selected from them. The degree to which the GSSLR is representative is subject to its constrained size which was discussed in section 4.3.2.

Consider a POLR in which the simulated troubleshooting guide for the outage alarm code lists all the parts associated with the WO. In this case, this research assumes that the health history would recommend all the required materials. Now consider a POLR in which the simulated troubleshooting guide for the outage alarm code is missing one or more of the parts associated with the WO. In this case, this research assumes that the health history would not recommend all the required materials.

8.5.1 Example of the Application to Troubleshooting

Table 8-1 presents the MLIs used in WO 80109138, in which a processor (the *M-System*) and a signal input module for a vibration sensor were replaced. The process based on the unenriched health history does not predict that any parts would be required. The process based on the EHH, on the other hand, predicts all but one of them. In this example, based on the assumptions stated above, neither process would identify all the parts that it is assumed would be required to complete the work and consequently it is assumed that the health history has not been usefully enriched in this case.

In practice, the part that was not predicted using the EHH is one of the three cables. The decision of which parts to bring is not really based purely on this process; planners pack an ensemble of spare parts based mostly on their experience. It is quite possible that the repair might have been achievable without this spare part.

This repair required three visits to the wind turbine, in which the processor was reset before being replaced. A repair strategy that utilises the EHH would inform the planner before the first visit that these parts might be required, and this would mean that this type of repair could be carried out in one visit.

Material	Material Description	Reserved	Unit
A9B00030597	Sensor cable M12 8POL 15M	1	PC
A9B00030733	Sensor cable M12 3POL 10M	1	PC
A9B00030734	Sensor cable M12 3POL 15M	1	PC
A9B00552369	ADAPTER F VIBRATION SENSOR	1	PC
A9B10001484	VIBRATION SENSOR ICP-MODULE WR12000M8	1	PC
A9B10043236	M-System for ******** (8 channel)	1	PC
A9B10144676	Cables ******** upgrade kit 8 ch	1	PC

Table 8-1, MLIs used in WO 80109138

8.5.2 Results

This section presents the concept of the estimated Proportion of POLRs in a set where the health history would recommend all the required Materials (*PM*), a measure of the quality of simulated troubleshooting guides. *PM* is defined by equation 6.17 using the number of POLRs in the EHH where the Health History Would Recommend all the required parts (WR_{EHH}) and the number of POLRs in the EHH where the Health History would Not Recommend all the required parts (NR_{EHH}).

$$PM = \frac{WR_{EHH}}{WR_{EHH} + NR_{EHH}} \tag{6.15}$$

Chapter 6 discussed the Bernoulli distribution. *PM* is a Bernoulli distribution; the *PM* of the GSSLR is analogous to the *PPV* of the GSSLR; and the *PM* of the EHH is analogous to the *PPV* of the EHH. The *PM* of the GSSLR is a point estimate of the *PM* of the EHH. This thesis will refer to the *PM* of the EHH as *PM* and to the *PM* of the GSSLR as $\widehat{PM}$.

This research investigated the effect of varying Th_S through a full range of values from -4 (enriched) to 4 (unenriched). The results are shown in Figure 8-6.

There are 29 POLRs in the GSSLR and 28 of them have parts associated with them. 28 is therefore the maximum number of POLRs in the GSSLR for which the Health History could recommend all the required parts, which would indicate a prefect set of troubleshooting guides.

The maximum *PM* in Figure 8-6 (0.643) occurs at a range of values of Th_S between -4 and -0.5. This means that, with a GSSLR of the size available, changing Th_S between these values does not change *PM*.

Figure 8-6 shows that *PM* is higher at lower values of Th_S, which indicates that health history enrichment increases *PM*, improving the proportion of repairs made potentially successful by the recommendation of the correct spare parts. This indicates that the PEOHH does enrich the health history in a way that is potentially useful for troubleshooting. This section will consider the uncertainty of this conclusion.

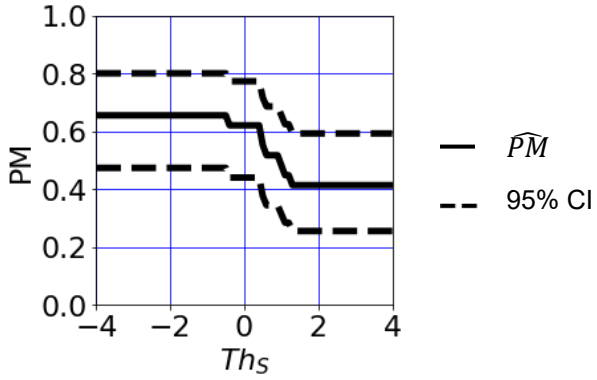


Figure 8-6, Estimated Proportion of POLRs where the Health History Would Recommend all the Required Materials (PM) and 95% CI against Score Threshold (Th_S)

Section 4.3.2 presented the method for the validation of the techniques for health history enrichment used in this thesis. Please recall from the chapter that the size of the sample “Gold Standard” Set of Linked Records (GSSLR) was constrained by the amount of expert time that was available. A Confidence Interval (CI) is a region with a specified, nominal probability that it contains a feature of interest. Figure 8-6 shows the 95% CI for the *PM* of the EHH identified using the Wilson interval which was discussed in section 6.1.3. The small size of the GSSLR, only 28 POLRs with materials assigned

to them, means that, for a *PM* of the GSSLR of 0.643, the 95% CI of the *PM* of the EHH is (0.458, 0.793).

Section 6.2 considered techniques for interval estimation for a change in a proportion. It showed that the Technique Assuming Independence (TAI) yielded very poor coverage and that the Technique Using Bootstrapping (TUB) yielded acceptable coverage. In Figure 8-6, the width of the CI is greater than the difference between the maximum and minimum values of the *PPV* of the GSSLR. Intuitively, one might infer from this that there is a significant risk that the optimum value has not been identified. Such an inference would however be based on the invalid assumption that the measurements of the *PPV* of the GSSLR are independent of each other. This section will instead use TUB for constructing a CI for the change in the *PPV* of the EHH.

Table 8-2 shows the effect of varying Th_S. Please recall that *WR* is the number of POLRs in the GSSLR where the simulated troubleshooting guide Would Recommend all the required parts and that *NR* is the number of POLRs in the GSSLR where that same guide would Not Recommend all the required parts, given the assumptions that have been described earlier in this section. TUB uses these results to estimate the uncertainty of the difference that it makes.

		Th_S = -4	
		WR	*NR*
Th_S = 4	*WR*	11	0
	NR	7	10

Table 8-2, Effect on the Richness of the Health History of Varying the Score Threshold (Th_S)

Equation 8.1 defines the Increase in the *PM* of the EHH that would by yielded by a Th_S of -4 over that yielded by a Th_S of 4 (${}^{PM}_{Th_S}I^{-4}_{4}$). It uses the Probability distribution of estimates of the *PM* of the EHH that would be yielded by a Th_S of 4 $(P(PM \mid Th_S = -4))$ and the same distribution that would be yielded by a Th_S of -4 $(P(PM \mid Th_S = 4))$

$$ {}^{PM}_{Th_S}I^{-4}_{4} = \frac{P(PM \mid Th_S = -4)}{P(PM \mid Th_S = 4)} - 1 \tag{8.1} $$

Equation 8.2 evaluates ${}^{PM}_{Th_S}\widehat{I^{-4}_{4}}$, a point estimate of ${}^{PM}_{Th_S}I^{-4}_{4}$, using the data from Table 8-2.

$$ {}^{PM}_{Th_S}\widehat{I^{-4}_{4}} = \frac{11+7}{11+0} - 1 = 64\% \tag{8.2} $$

This research used TUB to estimate the uncertainty of ${}^{PM}_{Th_S} I^{-4}_{4}$, which yielded a 95% CI of (20%, 167%). Again, this research predicts that this uncertainty will be drastically reduced by an upcoming innovation in maintenance record keeping that will be discussed in chapter 9; the automatic linking of new WOs to outages will result in a larger GSSLR and consequently in more certainty for this estimate.

To better understand this uncertainty, this research investigated the distribution of TUBs list of estimates of ${}^{PM}_{Th_S} I^{-4}_{4}$ using techniques presented in section 6.3. The results are shown in Figure 8-7 (a). The most important features of this distribution are the proportion of negative estimates and the proportion of zero estimates because these indicate respectively that the quality of record linkage has reduced and that it has not increased, as detailed in Figure 8-7 (b). The probability that a Th_S of -4 would yield a lower PM of the EHH than a Th_S of 4 is estimated by the proportion of bootstrap samples for which ${}^{PM}_{Th_S} I^{-4}_{4} \leq 0$, which this research found to be 0.03%. This indicates that there is little risk that the PEOHH does not enrich the health history.

Figure 8-7 (b) also shows that the proportion of bootstrap samples for which ${}^{PM}_{Th_S} I^{4}_{-4} < 0$ is 0. As explained in section 6.3, bootstrapping is conditional on the original sample, but this indicates that there is little risk that the PEOHH reduces the richness the health history.

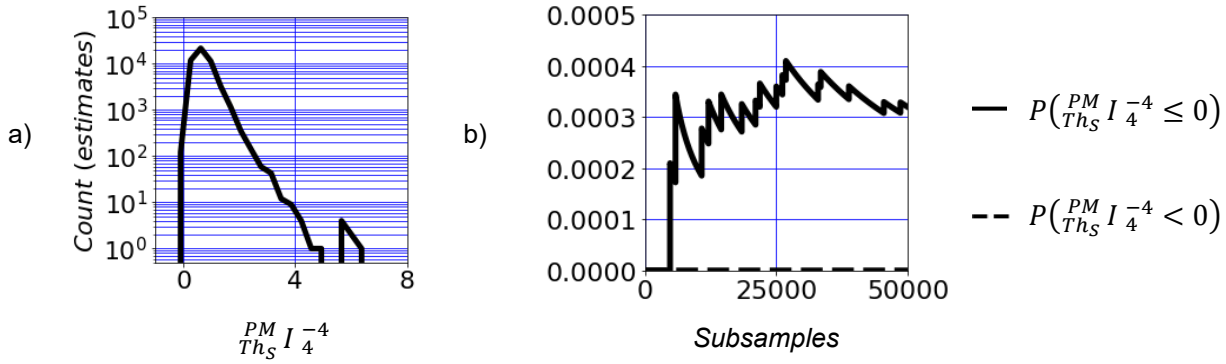


Figure 8-7, (a) Frequency of ${}^{PM}_{Th_S} I^{4}_{-4}$ for 50,000 Subsamples, (b) Proportion of Bootstrap Samples for which ${}^{PM}_{Th_S} I^{-4}_{4} \leq 0$ and for which ${}^{PM}_{Th_S} I^{-4}_{4} < 0$ against Number of Subsamples

This section showed that the PEOHH does enrich the health history. It estimated that the health history has become more useful for troubleshooting by 64% with a 95% CI of (20%, 167%). This result, based on the GSSLR, can be considered in combination with the results from sections 8.3 and 8.4 which are based on the entire health history of the farm. Those previous results do not suffer from the same uncertainty caused by a small sample size and show without uncertainty that the health history has been enriched.

8.6 Conclusion to ‘Has the Health History been Enriched?’

To investigate whether the PEOHH does what it is for, that is to enrich health history, this chapter has addressed RQ3:

RQ3 How can the richness of historical data on wind turbine health be measured?

This chapter has presented four measures of the richness of historical data on wind turbine health. It has measured the richness of both the unenriched and the enriched health history. Each of the four measures showed that enrichment has been achieved. For each measure, this chapter has quantified how much enrichment has been achieved. Section 8.5 predicted the effect of enrichment by assessing the EHHs usefulness for the application of troubleshooting. It estimated that the health history has become more useful for troubleshooting by 64% with a 95% CI of (20%, 167%).

9 Critical Review of this Research

This short chapter will review and evaluate this research. Salient points are the chronological order of work in this research and the size of the GSSLR. It will recommend a change of maintenance record keeping procedure.

This research benefitted from privileged access to Ørsted's records. It has not had direct access to other operators records which, like Ørsted's records, are also confidential. To apply the Process for the Enrichment of OWT Health History (PEOHH) developed in this research, other wind turbine operators, as well as operators in other sectors, would need to adapt it to their record keeping system.

This research has shown the benefit of linking records in the wind turbine maintenance sector. Dunn, 1946, observed that probabilistic record linkage techniques require a sample "Gold Standard" Set of Linked Records (GSSLR) to determine the optimum weighting for each feature. The identification of the GSSLR for this research was carried out in a series of meetings around 3 years after the start of the project. It was done so late with the intention that the validation of the PEOHH should be done after the development of the PEOHH. Such a sequence would ensure that the process of validation was independent of the process of development. In retrospect, it would have been advisable to identify the GSSLR as early as possible. The techniques used in this research that split the GSSLR into separate sets for testing and for training would have been sufficient to make testing independent of training without a requirement to develop the PEOHH before identifying the GSSLR.

The GSSLR represents expert experience and this thesis assumed that it is a set of true matches, although it could of course contain errors. This would mean that true and false linkages were incorrectly identified as such which in turn would cause errors in the optimisation of the PEOHH and in the measurement of its qualities. Such errors will be rectified by the following innovation in maintenance record keeping.

The size of the GSSLR, only 29 POLRs, was constrained by the amount of expert time that was available. The small size of the GSSLR meant that the results of the optimisation of the weights and thresholds used in the PEOHH is so uncertain that it is not clear whether all of the record linkage techniques presented in this thesis should be used or whether any of them can be disregarded without risking a drop in the quality of record linkage. It also means that the optimised weights and thresholds used by the PEOHH are approximate.

The uncertain work of linking WOs to outages could be made unnecessary for records generated in the future by a change of maintenance record keeping procedure. A record linking these two databases should be generated by the maintenance teams themselves. One way to do this would be

to give each outage a unique identifier; the outage number. Then, when maintenance technicians were back on the boat after completing a WO, they could note the outage number on the WO record.

Such labelling would negate the need for probabilistic methods to link future WOs to outages but it would not negate the need for probabilistic methods to link the existing records of outages to WOs. Older records are important to operators because improvements in Condition-Based Maintenance (CBM) can reduce the failure rate, more so for those failure modes that have effective CBM models. The study of how to more accurately link unlinked, historical maintenance records will therefore be of continuing importance.

Ready-linked maintenance records may in the future be used to train and to test techniques for linking unlinked, historical maintenance records. If and when new, very large GSSLRs become available, the uncertainty in the estimation of the *PPV* of the EHH that this research worked with will vastly reduce, improving confidence in the results. The record linkage techniques presented in this thesis should then be re-assessed to identify what combination of techniques yields the highest quality of record linkage and consequently the most useful EHH.

When searching for the global maximum of an irregular, multi-dimensional surface, it is problematic to determine whether the maximum identified is a global maximum or a local maximum. A more powerful computer can assess more combinations of features in the same time and so has a higher probability of identifying the true global maximum. To comply with Ørsted's confidentiality requirements, this research had to run the optimisation of weights and thresholds[72] on a standard laptop computer[73]. For future studies, the uncertainty of finding the global maximum would be reduced if it were reproduced using a more powerful processor.

[72] Chapter 7 presented the optimisation of weights and thresholds.

[73] Section 4.5 detailed the computer used for this research.

10 Conclusions to the Thesis

This research has developed new techniques for linking existing records of offshore wind turbine health history together. These techniques identify an Enriched Health History (EHH) with the aim of enabling improvements in the maintenance of offshore wind turbines. The productivity of a wind energy project depends on the price of electricity and on the suitability of the weather, both beyond the control of a maintenance team, but also on how much of its potential production of electricity is lost to outages and on the costs of maintenance and of operation. The EHH will enable improvements in maintenance scheduling, CBM and troubleshooting, and in the measurement of maintenance effectiveness. The wind farm maintenance sector can use the intelligence embodied in an EHH to increase productivity. Maintenance record linkage will also be of interest for the maintenance of equipment in other sectors.

This research developed new techniques for linking existing offshore wind turbine health history records together by joining WOs to outages. Previous authors (Leahy et al., 2017) have studied how to use wind turbine alarm logs to identify outages and generate a database of outages labelled with a failure mode while Papatzimos et al., 2017 use offshore wind turbine Work Orders (WO) to build a health history database. Multi-feature record linkage techniques are an established technology when applied to linking medical records (Sayers et al., 2015, Nasseh and Stausberg, 2016, Oliveira et al., 2016), address data (Churches et al., 2002, Comber et al., 2019, Lin et al., 2019), census data (Jaro, 1989, Smith et al., 2016) or genealogical records (Wilson, 2011) and have been used to detect duplicate internet search results (Hajishirzi et al., 2010). However, this thesis is the first to link records of maintenance data using multi-feature techniques.

In the only publication identified in the literature survey on the linkage of maintenance records, Papatzimos et al., 2017, link offshore wind turbine WOs to records of control system alarms but using a single feature (the timestamp). This thesis has shown that multi-feature record linkage techniques outperform single-feature record linkage techniques and has taken a significant step forward from Papatzimos by measuring the quality of the record linkage. The literature on record linkage does recognise that a small gold standard set of linked records can only be used to yield an uncertain estimate of the quality of record linkage but it does not quantify this uncertainty. This thesis has shown how uncertainty can be quantified.

The quality of record linkage was measured using a well-established method (Dunn, 1946), which compares the generated set of linked records to a gold standard set of linked records identified by human expertise. The Process for the Enrichment of OWT Health History (PEOHH) developed in this research requires a vector of weights and thresholds and the agreement and disagreement weights for each feature indicate the importance of the feature to the quality of record linkage. If the PEOHH can achieve the same quality of record linkage with the weights for a set of features set to zero then it can

disregard those features, avoiding the need for their computation. This research used differential evolution (Storn and Price, 1997) to globally optimise this vector of weights and thresholds.

To achieve a better quality of record linkage, this research developed and tested new record linkage innovations specific to the application of linking WOs to outages. There is inevitably some uncertainty associated with the measurement of the quality of record linkage, and consequently with the optimum values for the weights and thresholds; this research has not only measured the quality of record linkage but also estimated the uncertainty associated with that quality.

This research has identified an innovation in maintenance record keeping that would drastically reduce this uncertainty. Whilst a consequence of this uncertainty is that this thesis cannot offer conclusive advice as to which features to include in an ensemble of features for record linkage comparison, the thesis can recommend the use of all of the features presented until new maintenance record linkage practices can remedy the uncertainty.

This research has defined enrichment and has quantified the extent to which the process for the enrichment of the health history actually enriches the health history. It estimated that the PEOHH will improve offshore wind turbine fault troubleshooting by 64% with a 95% CI of (20%, 167%) and anticipates similar improvements in maintenance scheduling, in condition-based maintenance, and in the measurement of maintenance effectiveness.

This research used a database of material consumption in which each material line item is labelled with an order number. It developed two techniques that use these data as part of an ensemble of features for record linkage comparison. The first technique used a Bernoulli Naïve Bayes (BNB) classifier to predict the probability for each pair of Linked Records (POLR) that the parts used in the WO correspond to the alarm code of the outage. Historical data on machinery failures tends to be unbalanced; some failure modes feature more than others. As a result of this unbalance, this thesis has shown that BNB classification is not an appropriate method for this application. The second technique is simpler, checking whether the records for the outage failure mode contain each part assigned to the WO. It does not suffer from the problem with unbalanced data because it only looks at the relevant event code. This thesis has shown that the second technique is an appropriate method for this application.

This research has shown the potential that linking offshore wind turbine maintenance records together has in enabling improvements in maintenance practice. Further work linking together maintenance records from other sectors will be informed by the record linkage techniques presented in this thesis.

Each wind energy operator holds their own set of health history data. This research will enable them to use their data to identify an EHH and this will enable further innovations in maintenance scheduling, CBM, troubleshooting and the measurement of maintenance effectiveness.

The EHH will provide valuable insights into historical costs of maintenance and of lost production. These insights could be used to further optimise maintenance scheduling from a logistical perspective by providing more robust information to the models.

The EHH will provide more detailed labels than currently available for the training and for the testing of CBM models. This will enable developers to develop new models and to improve existing ones.

The EHH will extend the troubleshooting guide parts list and this can avoid the lost production caused by the right part not being available. Bringing the correct parts to the OWT will be of increasing importance as the distance to shore increases and it could help to avoid the cost of an offshore spare parts store.

The EHH integrates records of what work has been done on each turbine with failure mode specific information on the failure rate. If a repair was successful, then the time to failure for the failure modes effected by the repair would tend to increase. The EHH makes it possible to use this integrated data for the first time to measure maintenance effectiveness.

To summarise the contributions made by this research:

- Applied multi-feature record linkage techniques to maintenance data for the first time.
- Applied statistical techniques for the interval estimation of a binomial proportion to record linkage techniques for the first time.
- Estimated the distribution of the coverage error of statistical techniques for the interval estimation of a binomial proportion for the first time.

The resulting main contribution of this research is a process for the enrichment of offshore wind turbine health history.

References

Aalen, O.O., 1989. A linear regression model for the analysis of life times. Statistics in medicine, 8(8), pp.907-925.

Abbas, M., Memon, K.A., Jamali, A.A., Memon, S. and Ahmed, A., 2019. Multinomial Naive Bayes classification model for sentiment analysis. International Journal of Computer Science and Network Security, 19(3), p.62.

Abbasa, Q., Ahmadb, J. and Jabeenc, H., 2017. The analysis, identification and measures to remove inconsistencies from differential evolution mutation variants. SCIENCEASIA, 43, pp.52-68.

Al-Aidaroos, K.M., Bakar, A.A., Othman, Z., 2012. Medical Data Classification with Naive Bayes Approach. Information Technology Journal, 11: pp.1166-1174.

Almeida, T.A., Almeida, J. and Yamakami, A., 2011. Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of Internet Services and Applications, 1(3), pp.183-200.

Altheneyan, A.S. and Menai, M.E.B., 2014. Naïve Bayes classifiers for authorship attribution of Arabic texts. Journal of King Saud University-Computer and Information Sciences, 26(4), pp.473-484.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbour nonparametric regression. *The American Statistician*, *46*(3), pp.175-185

anky, 2019, Repeat an Experiment and Store the Results in a Dataframe - Stack Overflow, https://stackoverflow.com/questions/55848879/repeat-an-experiment-and-store-the-results-in-a-dataframe

Arar, Ö.F. and Ayan, K., 2017. A feature dependent Naive Bayes approach and its application to the software defect prediction problem. Applied Soft Computing, 59, pp.197-209.

Artigao, E., Koukoura, S., Honrubia-Escribano, A., Carroll, J., McDonald, A., Gómez-Lázaro, E., 2018. Current Signature and Vibration Analyses to Diagnose an In-Service Wind Turbine Drive Train. Energies 11, no. 4: 960.

Arts, J., Basten, R. and van Houtum, G.J., 2019. Maintenance service logistics. In Operations, Logistics and Supply Chain Management (pp. 493-517). Springer, Cham.

Bach-Andersen, M., Winther, O. and Rømer-Odgaard, B., 2015. Scalable systems for early fault detection in wind turbines: a data driven approach. In Annual Conference of the european Wind Energy Association.

Bach-Andersen, M., Winther, O. and Rømer-Odgaard, B., 2015. Scalable systems for early fault detection in wind turbines: a data driven approach. In Proceedings of the annual conference of the European wind energy association.

Bach-Andersen, M., Rømer-Odgaard, B. and Winther, O., 2017. Flexible non-linear predictive models for large-scale wind turbine diagnostics. Wind Energy, 20(5), pp.753-764.

Bach-Andersen, M., Rømer-Odgaard, B. and Winther, O., 2018. Deep learning for automated drivetrain fault detection. Wind Energy, 21(1), pp.29-41.

Balci, O., 1994. Validation, verification, and testing techniques throughout the life cycle of a simulation study. Annals of operations research, 53(1), pp.121-173.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances

Bellman, R., 1956. Dynamic programming (Vol. 295). RAND CORP SANTA MONICA CA.

BENY, 2019, Column contains column 4 - Stack Overflow, https://stackoverflow.com/questions/56276574/column-contains-column-4

Boote, D., Galleggioni, F., Colaianni, T., McCartan, S., Thompson, T., Iliopulos, F., McFarlane, I., Rose, D., Verheijden, B., Anderberg, C. and Phalm, H., 2015. DESIGN-DRIVEN INNOVATION: NEXT GENERATION WIND FARM MOTHERSHIP FOR THE NORTH SEA. In Marine Design 2015.

BP Statistical Review of World Energy, 2021.

Brown, L.D., Cai, T.T. and DasGupta, A., 2001. Interval estimation for a binomial proportion. Statistical science, pp.101-117.

Cameron, E., 2011. On the estimation of confidence intervals for binomial population proportions in astronomy: the simplicity and superiority of the Bayesian approach. Publications of the Astronomical Society of Australia, 28(2), pp.128-139.

Carroll, J., McDonald, A. and McMillan, D., 2015. Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. Wind Energy.

Carroll, J., Koukoura, S., McDonald, A., Charalambous, A., Weiss, S. and McArthur, S., 2019. Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques. Wind Energy, 22(3), pp.360-375.

Christen, P., 2012. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media.

Churches, T., Christen, P., Lim, K. and Zhu, J.X., 2002. Preparation of name and address data for record linkage using hidden Markov models. BMC Medical Informatics and Decision Making, 2(1), p.9.

Cipollini, F., Oneto, L., Coraddu, A., Murphy, A.J. and Anguita, D., 2018. Condition-Based Maintenance of Naval Propulsion Systems: Data Analysis with Minimal Feedback. Reliability Engineering & System Safety.

Clopper, C.J. and Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26(4), pp.404-413.

Cock, P., 2005. RPy: A Simple and Efficient Access to R from Python. URL http://rpy. sourceforge. net.

Comber, S. and Arribas-Bel, D., 2019. Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. Transactions in GIS, 23(2), pp.334-348.

Cortes, C. and Vapnik, V., 1995. Support vector machine. Machine learning, 20(3), pp.273-297.

Cox, D.R., 1958. Two further applications of a model for binary regression. Biometrika, 45(3/4), pp.562-565.

Crabtree, C.J., Feng, Y. and Tavner, P.J., 2010, April. Detecting incipient wind turbine gearbox failure: a signal analysis method for on-line condition monitoring. In Proceedings of European Wind Energy Conference (EWEC 2010), Warsaw, Poland (pp. 20-23).

Dagnely, P., Tsiporkova, E., Tourwé, T., Ruette, T., De Brabandere, K. and Assiandi, F., 2015, July. A semantic model of events for integrating photovoltaic monitoring data. In Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on (pp. 24-30). IEEE.

Dao, C., Basten, R. and Hartmann, A., 2018. Maintenance scheduling for railway tracks under limited possession time. Journal of Transportation Engineering, Part A: Systems, 144(8), p.04018039.

Davis, Neil; Badger, Jake; Hahmann, Andrea N.; Hansen, Brian Ohrbeck; Olsen, Bjarke Tobias; Mortensen, Niels Gylling; et al. (2019): Global Wind Atlas v3. figshare. Dataset.

Dawar, D. and Ludwig, S.A., 2014, December. Differential evolution with dither and annealed scale factor. In 2014 IEEE Symposium on Differential Evolution (SDE) (pp. 1-8). IEEE.

Dean, N. and Pagano, M., 2015. Evaluating confidence interval methods for binomial proportions in clustered surveys. Journal of Survey Statistics and Methodology, 3(4), pp.484-503.

Dharmadhikari, S.C., Ingle, M. and Kulkarni, P., 2011. Empirical studies on machine learning based text classification algorithms. Advanced Computing, 2(6), p.161.

DNV GL, 2017, Definitions of Availability Terms for the Wind Industry

Dong, A., Zhao, Y., Liu, X., Shang, L., Liu, Q. and Kang, D., 2017, August. Fault Diagnosis and Classification in Photovoltaic Systems Using SCADA Data. In Sensing, Diagnostics, Prognostics, and Control (SDPC), 2017 International Conference on (pp. 117-122). IEEE.

Doostparast, M. and Doostparast, M., 2018. Prediction of corrosions in Gas and Oil pipelines based on the theory of records. arXiv preprint arXiv:1801.00959.

Dorai-Raj, S., 2014, Binomial Confidence Intervals For Several Parameterizations

Doyle, P., 1973. The use of automatic interaction detector and similar search procedures. Journal of the Operational Research Society, 24(3), pp.465-467.

Dunn, H.L., 1946. Record linkage. American Journal of Public Health and the Nations Health, 36(12), pp.1412-1416.

Efron, B., 1979. Computers and the theory of statistics: thinking the unthinkable. SIAM review, 21(4), pp.460-480.

Espinoza, D. and Morris, J.W., 2013. Decoupled NPV: a simple, improved method to value infrastructure investments. Construction management and economics, 31(5), pp.471-496.

European Patent Application EP 2 339 174 A1, 2011

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

Fellegi, I.P. and Sunter, A.B., 1969. A theory for record linkage. Journal of the American Statistical Association, 64(328), pp.1183-1210.

Feng, Y., Tavner, P.J. and Long, H., 2010. Early experiences with UK Round 1 offshore wind farms. Proceedings of the Institution of Civil Engineers: energy., 163(4), pp.167-181.

Feng, Y., Qiu, Y., Crabtree, C.J., Long, H. and Tavner, P.J., 2013. Monitoring wind turbine gearboxes. Wind Energy, 16(5), pp.728-740.

Festa, A., Williams, K., Hanley, A.J., Otvos, J.D., Goff, D.C., Wagenknecht, L.E. and Haffner, S.M., 2005. Nuclear magnetic resonance lipoprotein abnormalities in prediabetic subjects in the Insulin Resistance Atherosclerosis Study. Circulation, 111(25), pp.3465-3472.

Fix, E. and Hodges, J.L., 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), pp.238-247.

Friedlander, 2019, Time difference between Time Period and Instant - Stack Overflow, https://stackoverflow.com/questions/55006135/time-difference-between-time-period-and-instant-1

Garcia, M.C., Sanz-Bobi, M.A. and del Pico, J., 2006. SIMAP: Intelligent System for Predictive Maintenance: Application to the health condition monitoring of a windturbine gearbox. Computers in Industry, 57(6), pp.552-568.

Godwin, J.L. and 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.

Gonzalez, E., Reder, M. and Melero, J.J., 2016, September. SCADA alarms processing for wind turbine component failure detection. In Journal of Physics: Conference Series (Vol. 753, No. 7, p. 072019). IOP Publishing.

Gonzalez, E., Nanos, E.M., Seyr, H., Valldecabres, L., Yürüşen, N.Y., Smolka, U., Muskulus, M. and Melero, J.J., 2017. Key performance indicators for wind farm operation and maintenance. Energy Procedia, 137, pp.559-570.

Gray, C.S. and Watson, S.J., 2010. Physics of failure approach to wind turbine condition based maintenance. Wind Energy, 13(5), pp.395-405.

Gray, C.S. and Watson, S.J., 2010. Physics of failure approach to wind turbine condition based maintenance. Wind Energy, 13(5), pp.395-405.

Gross, J. and Groß, J., 2003. Linear regression (Vol. 175). Springer Science & Business Media.

Hahn, B., Durstewitz, M., Rohrig, K.: ‘Reliability of wind turbines – experience of 15 years with 1500 WTs’, in ‘Wind energy’ (Springer, 2007), pp. 329–332

Hajishirzi, H., Yih, W.T. and Kolcz, A., 2010, July. Adaptive near-duplicate detection via similarity learning. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 419-426).

Hicks, J.L., Uchida, T.K., Seth, A., Rajagopal, A. and Delp, S.L., 2015. Is my model good enough? Best practices for verification and validation of musculoskeletal models and simulations of movement. Journal of biomechanical engineering, 137(2), p.020905.

horseoftheyear, 2020, Coverage of the 99% Confidence Interval of a Binomial Proportion - Stack Overflow, https://stackoverflow.com/questions/62230610/coverage-of-the-99-confidence-interval-of-a-binomial-proportion

Hu, R.L., Leahy, K., Konstantakopoulos, I.C., Auslander, D.M., Spanos, C.J. and Agogino, A.M., 2016. Using Domain Knowledge Features for Wind Turbine Diagnostics. In 15th IEEE International Conference on Machine Learning and Applications (ICMLA)

Imani, M., 2021. Integration of the k-nearest neighbours and patch-based features for PolSAR image classification by using a two-branch residual network. Remote Sensing Letters, 12(11), pp.1112-1122.

Jiang, L., Wang, D. and Cai, Z., 2012. Discriminatively weighted naive Bayes and its application in text classification. International Journal on Artificial Intelligence Tools, 21(01), p.1250007.

Isobe, T., Feigelson, E.D., Akritas, M.G. and Babu, G.J., 1990. Linear regression in astronomy. The astrophysical journal, 364, pp.104-113.

Jaro, M.A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84(406), pp.414-420.

Jeffreys, H., 1973. Scientific inference. Cambridge University Press.

Joachims, T., 1999, June. Transductive inference for text classification using support vector machines. International Conference on Machine Learning, (Vol. 99, pp. 200-209).

Kim, S.B., Han, K.S., Rim, H.C. and Myaeng, S.H., 2006. Some effective techniques for naive bayes text classification. IEEE transactions on knowledge and data engineering, 18(11), pp.1457-1466.

Koster, J. and McElreath, R., 2017. Multinomial analysis of behavior: statistical methods. Behavioral Ecology and Sociobiology, 71(9), pp.1-14.

Koukoura, S., Carroll, J., Weiss, S. and McDonald, A., 2017, August. Wind turbine gearbox vibration signal signature and fault development through time. In Signal Processing Conference (EUSIPCO), 2017 25th European (pp. 1380-1384). IEEE.

Kuncheva, L.I., 2006. On the optimality of Naïve Bayes with dependent binary features. Pattern Recognition Letters, 27(7), pp.830-837.

Kusiak, A. and Li, W., 2011. The prediction and diagnosis of wind turbine faults. Renewable energy, 36(1), pp.16-23.

Kusiak, A. and Verma, A., 2012. Analyzing bearing faults in wind turbines: A data-mining approach. Renewable Energy, 48, pp.110-116.

Labroquere, J., Héritier, A., Riccardi, A. and Izzo, D., 2014, September. Evolutionary Constrained Optimization for a Jupiter Capture. In International Conference on Parallel Problem Solving from Nature (pp. 262-271). Springer, Cham.

Leahy, K., Gallagher, C., Bruton, K., O'Donovan, P. and O'Sullivan, D.T., 2017, November. Automatically Identifying and Predicting Unplanned Wind Turbine Stoppages Using SCADA and Alarms System Data: Case Study and Results. In Journal of Physics: Conference Series (Vol. 926, No. 1, p. 012011). IOP Publishing.

Leahy, K., Gallagher, C., O'Donovan, P., Bruton, K. and O'Sullivan, D.T., 2018. A robust prescriptive framework and performance metric for diagnosing and predicting wind turbine faults based on SCADA and alarms data with case study. Energies, 11(7), p.1738.

Leahy, K., Gallagher, C., O'Donovan, P. and O'Sullivan, D.T., 2019. Issues with data quality for wind turbine condition monitoring and reliability analyses. *Energies*, *12*(2), p.201.

Lee, Y., 1991. Handwritten digit recognition using k nearest-neighbor, radial-basis function, and backpropagation neural networks. Neural computation, 3(3), pp.440-449.

Levenshtein, V.I., 1966, February. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).

Lin, P., Yuan, X.X. and Tovilla, E., 2019. Integrative modeling of performance deterioration and maintenance effectiveness for infrastructure assets with missing condition data. Computer-Aided Civil and Infrastructure Engineering, 34(8), pp.677-695.

Lin, Y., Kang, M. and He, B., 2019. Spatial pattern analysis of address quality: A study on the impact of rapid urban expansion in China. Environment and Planning B: Urban Analytics and City Science, p.2399808319895272.

Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L. and Huang, T., 2011, June. Large-scale image classification: fast feature extraction and svm training. Conference on Computer Vision and Pattern Recognition, (pp. 1689-1696). IEEE.

Liu, Z., Meyendorf, N. and Mrad, N., 2018, April. The role of data fusion in predictive maintenance using digital twin. In AIP Conference Proceedings (Vol. 1949, No. 1, p. 020023). AIP Publishing.

Luo, B. and Lin, L., 2018, April. Multi-objective decision-making model based on CBM for an aircraft fleet. In AIP Conference Proceedings (Vol. 1955, No. 1, p. 040106). AIP Publishing.

Lydia, M., Selvakumar, A.I., Kumar, S.S. and Kumar, G.E.P., 2013. Advanced algorithms for wind turbine power curve modeling. IEEE Transactions on sustainable energy, 4(3), pp.827-835.

Mantalos, P. and Zografos, K., 2008. Interval estimation for a binomial proportion: a bootstrap approach. Journal of Statistical Computation and Simulation, 78(12), pp.1251-1265.

Marine Traffic Website, accessed 2020, https://www.marinetraffic.com/

McCallum, A. and Nigam, K., 1998, July. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

Mellit, A. and Kalogirou, S.A., 2011. ANFIS-based modelling for photovoltaic power supply system: A case study. Renewable energy, 36(1), pp.250-258.

Metsis, V., Androutsopoulos, I. and Paliouras, G., 2006, July. Spam filtering with naive bayes-which naive bayes?. In CEAS (Vol. 17, pp. 28-69).

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR

Naseem, I., Togneri, R. and Bennamoun, M., 2010. Linear regression for face recognition. IEEE transactions on pattern analysis and machine intelligence, 32(11), pp.2106-2112.

Nasseh, D. and Stausberg, J., 2016. Evaluation of a Binary Semi-supervised Classification Technique for Probabilistic Record Linkage. Methods of information in medicine, 55(02), pp.136-143.

Newcombe, R.G., 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in medicine, 17(8), pp.873-890.

Nielsen, J.J. and Sørensen, J.D., 2011. On risk-based operation and maintenance of offshore wind turbine components. Reliability engineering & system safety, 96(1), pp.218-229.

Oliveira, G.P.D., Bierrenbach, A.L.D.S., Camargo Júnior, K.R.D., Coeli, C.M. and Pinheiro, R.S., 2016. Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. Revista de saude publica, 50, p.49.

Olson, B., Hashmi, I., Molloy, K., and Shehu1, A., Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules, Advances in Artificial Intelligence, Volume 2012

Papatzimos, A.K., Dawood, T. and Thies, P.R., 2017, June. An integrated data management approach for offshore wind turbine failure root cause analysis. ASME.

Papatzimos, A.K., Thies, P.R. and Dawood, T., 2019. Offshore wind turbine fault alarm prediction. Wind Energy, 22(12), pp.1779-1788.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.

Pfaffel, S., Faulstich, S. and Sheng, S., 2019, October. Recommended key performance indicators for operational management of wind turbines. In Journal of Physics: Conference Series (Vol. 1356, No. 1, p. 012040). IOP Publishing.

piRSquared, 2018, Lookup with Missing Labels - Stack Overflow, https://stackoverflow.com/questions/53417435/lookup-with-missing-labels

Pinar Pérez, J.M., García Márquez, F.P., Tobias, A., et al.: 'Wind turbine reliability analysis', Renew. Sustain. Energy Rev., 2013, 23, pp. 463–472

Porter, M.F., 1980. An algorithm for suffix stripping. Program, 14(3), pp.130-137.

Prässler, T. and Schaechtele, J., 2012. Comparison of the financial attractiveness among prospective offshore wind parks in selected European countries. Energy Policy, 45, pp.86-101.

Price, K., Storn, R.M. and Lampinen, J.A., 2006. Differential evolution: a practical approach to global optimization. Springer Science & Business Media.

Qiu, Y., Feng, Y., Tavner, P., Richardson, P., Erdos, G. and Chen, B., 2012. Wind turbine SCADA alarm analysis for improving reliability. Wind Energy, 15(8), pp.951-966.

Qiu, Y., Feng, Y., Sun, J., Zhang, W. and Infield, D., 2016. Applying thermophysics for wind turbine drivetrain fault diagnosis using SCADA data. IET Renewable Power Generation, 10(5), pp.661-668.

R Core Team, 2021, R: A Language and Environment for Statistical Computing

Rezaeian, N. and Novikova, G., 2020. Persian text classification using naive bayes algorithms and support vector machine algorithm. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 8(1), pp.178-188.

Ribrant, J., Bertling, L.M.: 'Survey of failures in wind power systems with focus on Swedish wind power plants during 1997–2005', IEEE Trans. Energy Convers., 2007, 22, (1), pp. 167–173

Robinson, S., 1999. Simulation verification, validation and confidence: a tutorial. Transactions of the Society for Computer Simulation, 16(2), pp.63-69.

Rodriguez-Mier, P., Mucientes, M., Lama, M. and Couto, M.I., 2010. Composition of web services through genetic programming. Evolutionary Intelligence, 3(3-4), pp.171-186.

Rubenstein, H. and Goodenough, J.B., 1965. Contextual correlates of synonymy. Communications of the ACM, 8(10), pp.627-633.

Rust, J., 1997. Using randomization to break the curse of dimensionality. Econometrica: Journal of the Econometric Society, pp.487-516.

Sadinle, M. and Fienberg, S.E., 2013. A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. Journal of the American Statistical Association, 108(502), pp.385-397.

Sakakibara, Y., Misue, K. and Koshiba, T., 1993, March. Text classification and keyword extraction by learning decision trees. In Proceedings of 9th IEEE Conference on Artificial Intelligence for Applications (p. 466). IEEE.

Salo, E., McMillan, D. and Connor, R., 2019, September. Work orders-value from structureless text in the era of digitisation. In SPE Offshore Europe Conference and Exhibition. Society of Petroleum Engineers.

Sandell, 2020, What are 'population energies'? - Stack Overflow https://stackoverflow.com/questions/64321365/what-are-population-energies

Sargent, R.G., 2013. Verification and validation of simulation models. Journal of simulation, 7(1), pp.12-24.

Sayers, A., Ben-Shlomo, Y., Blom, A.W. and Steele, F., 2015. Probabilistic record linkage. International journal of epidemiology, 45(3), pp.954-964.

Schwefel, H.P. and Rudolph, G., 1995, June. Contemporary evolution strategies. In European conference on artificial life (pp. 891-907). Springer, Berlin, Heidelberg.

SergioR, 2020, Compare Methods in rpy2 - Stack Overflow, https://stackoverflow.com/questions/62149840/compare-methods-in-rpy2

Seyr, H. and Muskulus, M., 2019. Decision support models for operations and maintenance for offshore wind farms: A review. Applied Sciences, 9(2), p.278.

Sheng, G., Hou, H., Jiang, X. and Chen, Y., 2018. A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model. IEEE Transactions on Smart Grid, 9(2), pp.695-702.

Sheng, S.: ‘Report on wind turbine subsystem reliability – a survey of various databases’ (National Renew. Energy Lab, 2013)

Silenced Temporarily, 2019, Apply function to dataframe column of lists - Stack Overflow, https://stackoverflow.com/questions/55103428/apply-function-to-dataframe-column-of-lists

Smith, A, 2017, Python - Levenstein Distance Substring - Stack Overflow https://stackoverflow.com/questions/44398027/levenstein-distance-substring

Smith D, Shlomo N. Privacy preserving record linkage. University of Manchester, School of Social Sciences Working Paper, 2014.

Spinato, F., Tavner, P.J., Van Bussel, G.J.W. and Koutoulakos, E., 2009. Reliability of wind turbine subassemblies. IET Renewable Power Generation, 3(4), pp.387-401.

StackExchange, Stack Overflow, https://stackoverflow.com/, accessed 2021

StackExchange, Cross Validated, https://stats.stackexchange.com/, accessed 2021

Stock-Williams, C. and Swamy, S.K., 2019. Automated daily maintenance planning for offshore wind farms. Renewable Energy, 133, pp.1393-1403.

Storn, R. and Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4), pp.341-359.

Tang, M., Chen, W., Zhao, Q., Wu, H., Long, W., Huang, B., Liao, L. and Zhang, K., 2019. Development of an SVR Model for the Fault Diagnosis of Large-Scale Doubly-Fed Wind Turbines Using SCADA Data. Energies, 12(17), p.3396.

Takagi, N., 2006. An application of binary decision trees to pattern recognition. Journal of Advanced Computational Intelligence and Intelligent Informatics, 10(5), pp.682-687.

Tavner, P.J., Xiang, J., Spinato, F.: 'Reliability analysis for wind turbines', Wind Energy, 2007, 10, (1), pp. 1–18

Tierney, W.M., McDonald, C.J. and McCabe, G., 1985. Serum potassium testing in diuretic-treated outpatients: a multivariate approach. Medical Decision Making, 5(1), pp.89-104.

V.G.B. PowerTech, 2014. RDS-PP Application Guideline Part 32: Wind Power Plants. VGB Powertech eV VGB-S-823-32-2014-03-EN-DE.

Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical society, 54(3), pp.426-482.

Walford, C.A., 2006. Wind turbine reliability: understanding and minimizing wind turbine operation and maintenance costs (No. SAND2006-1100). Sandia National Laboratories.

Wang, J., Liu, X.Z. and Ni, Y.Q., 2018. A Bayesian Probabilistic Approach for Acoustic Emission-Based Rail Condition Assessment. Computer-Aided Civil and Infrastructure Engineering, 33(1), pp.21-34.

Wang, J., Liang, Y., Zheng, Y., Gao, R.X. and Zhang, F., 2020. An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. Renewable Energy, 145, pp.642-650.

Wickramasinghe, I. and Kalutarage, H., 2021. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), pp.2277-2293.

Wilkinson, M., Harman, K.: 'Measuring wind turbine reliability, results of the reliawind project'. EWEA Annual Conf. 2011, 2011

Wilson, D.R., 2011, July. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In The 2011 International Joint Conference on Neural Networks (pp. 9-14). IEEE.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158), pp.209-212.

WindEurope Annual Offshore Statistics 2018 https://windeurope.org/about-wind/statistics/offshore/european-offshore-wind-industry-key-trends-statistics-2018/

Xu, S., 2018. Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 44(1), pp.48-59.

Xu, S., Liu, S., Wang, H., Chen, W., Zhang, F. and Xiao, Z., 2021. A hyperspectral image classification approach based on feature fusion and multi-layered gradient boosting decision trees. Entropy, 23(1), p.20.

Yürüşen, N.Y., Rowley, P.N., Watson, S.J. and Melero, J., 2020. Automated wind turbine maintenance scheduling. Reliability Engineering & System Safety.

Zaher, A.S.A.E., McArthur, S.D.J., Infield, D.G. and Patel, Y., 2009. Online wind turbine fault detection through automated SCADA data analysis. Wind Energy, 12(6), pp.574-593.

Zhang, L., Jiang, L., Li, C. and Kong, G., 2016. Two feature weighting approaches for naive Bayes text classifiers. Knowledge-Based Systems, 100, pp.137-144.