

RoC & PoS Study

Bobby Freer – P141566

Lectures – Ashely Pullen / Peter Charlton

Submission: 26/02/2017

Abstract

This paper investigates the Non-Destructive Testing (NDT) statistical methods, probability of sizing (PoS), and receiver operator characteristic curves (ROC) used throughout NDT industry. Three methods are analysed using the probability of sizing exercise, one using magnetic flux leakage and the other two being ultrasonic methods. Three unknown methods of NDT were also analysed using receiver operator characteristic curves. From the recorded data and statistical analysis, it was found that in the PoS under calibrated and over calibrated recorded data points from different systems could be used to create a set of data points which follow the true size line. From RoC analysis it was found that method three had the highest chance of detecting a flaw whereas method one had the highest chance of finding a false alarm.

Contents

Contents.....	2
1. Introduction	3
2. Literature Review	4
2.1. PoS	4
2.2. ROC.....	6
3. Probability of Sizing.....	7
3.1. Methodology.....	7
3.2. Results & Analysis	7
3.2.1. Line & Correlation Comparisons	10
3.2.2. Solutions / Compromises	11
4. Receiver operator Characteristics.....	13
4.1 Methodology.....	13
4.2 Results & Analysis	14
5. Discussion.....	16
6. Conclusion.....	17
Appendix	18

Figures & Tables

Figure 1 PoS Graph.....	4
Figure 2 Confidence limits.....	5
Figure 3 ROC Curve	6
Figure 4 MFL PoS.....	7
Figure 5 UT-A Graph.....	8
Figure 6 UT-B PoS.....	9
Figure 7 R-Squared Comparison	10
Figure 8 MFL &UT-A Average.....	11
Figure 9 MFL Error Average	12
Figure 10 RoC Graph	14
Table 1 RoC	13
Table 2 RoC Values.....	15

1. Introduction

Non-destructive testing processes and procedures can be ineffective unless the correct method is applied to the task in question. The optimum process and procedure can be defined for a given task by using statistical analysis on experimental trials. Statistical analysis will tell of how accuracy from readings could be improved and help the operators chances of finding a flaw.

As NDT engineer's methods of statistics can be used to analyse data such as the mean, standard deviation and, confidence limits, these analytical methods will indicate how consistent tests may be. A method of analysing data using these tools is PoS (probability of sizing), PoS will tell how accurate a systems ability for flaw detection is and analyse operator capabilities of defining the size of a known defect using Probability of Sizing (PoS).

Another statistical method that will be used in the report will be ROC curves (receiver operator characteristic). ROC curves can be used to determine how likely a system is to detect a flaw and how likely a system is to call a false alarm, be able to analyse RoC data will allow the engineer to make a judgement on the recordings taken.

The ability to check how accurately an operator can define a defects dimensions, will allow designers to take into account the size of defects which could be missed in an inspection, therefore when the engineers are calculating fracture mechanics it will allow for them a safety factor to be determined for the component.

In this report two different sets of data will be analysed with PoS and ROC. The aim was to be able to use the statistical methods to determine what operator has sized the defects with the highest accuracy and what is the likelihood of recording a false alarm (no defect present but defect is recorded).

2. Literature Review

2.1. PoS

Probability of Sizing assesses an operator's ability to accurately size defects. Uncertainties exist when the operator must give flaw dimensions, different operators readings will vary plus/minus the actual size of the flaw, as is typical with all operator measurements. For example, if a group of 20 people were to use measure a block with a vernier calliper for *Length, Height, Width* results would differ due to how much pressure is applied to the calliper jaws, where the measurement is taken, experience, mechanical or digital vernier gauge possible rounding errors (www.nde-ed.org, 2015).

PoS is important as sizing inconsistency will relate back to fracture mechanics already calculated in design. When an NDT operator is asked to make a judgment about the size of the flaw the result is then compared to an allowable dimension chart, operator under-sizing could ignore a potential failure and oversizing could cause concern to a within tolerance component. The graph on the next page has an Ideal Sizing line, this would have been calculated through either destructive testing or a known defect size, the recordings from the operator can be compared to what is an allowable deviation and the true size of the PoS graph.

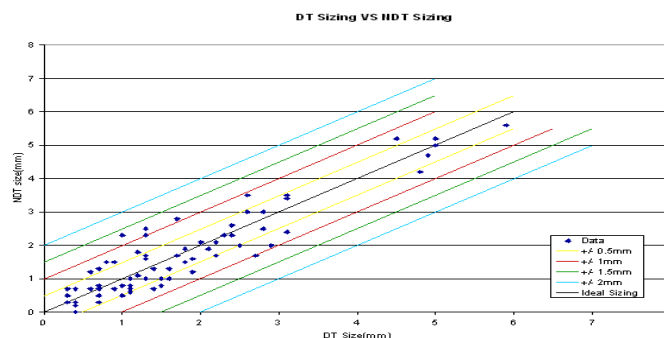


Figure 1 PoS Graph

(www.ndt.net-POD, 2006)

PoS graphs are created by comparing the true size against the operators estimated readings. From these values the operators error can be calculated using the following;

$$\text{Operator Error} = \text{Operator Measurement} - \text{True Size}$$

This value is used to calculate the standard deviation of the operator's errors allowing for the 95% confidence limits to be determined. Confidence Limits are constructed at set parameters by the user such as 95% (www.rfcafe.com, 2017). This means that if a random sample of data is taken from the population on numerous occasions all samples should sit within the 95% confidence limits for 95 % of the cases. The confidence interval will form a bracket above and below the point of interest (i.e. the mean), known as the upper and lower limits as the figure below displays;

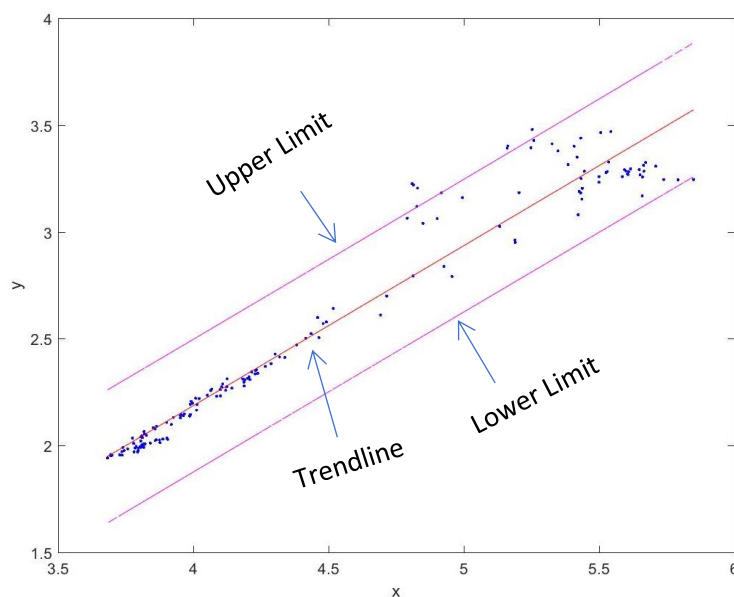


Figure 2 Confidence limits

Confidence limits can be calculated using values from the equation of a straight line (which could be found the trendline in this case) and the standard deviation.

$$\pm 95\% \text{ Limit} = mx + (C \pm (1.96 * \sigma))$$

(1.96 comes from fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean (itl.nist.gov, 2013).

Confidence will vary based on the data which it is created from, a wider spread of data results will create a wider 95% confidence band, similarly a tight set of data will produce a narrow confidence band.

2.2. ROC

Receiver operator characteristic (ROC) were first used in WW2 by the British army. ROC curves were developed to differentiate in the radar system between an enemy aircraft and signal noise, i.e. flocks of birds (statisticshowto.com, 2011). The curve will give indication on how likely is that that the reading is a false alarm (False Positive) or in the case of NDT a flaw (true positive).

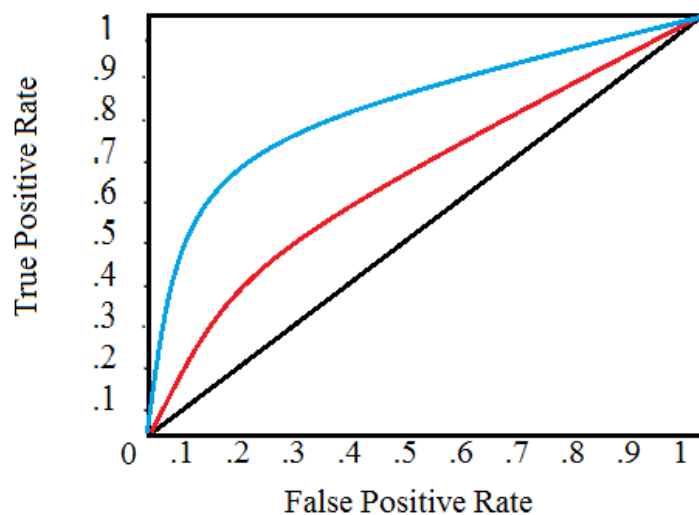


Figure 3 ROC Curve

(statisticshowto.com, 2011)

Graph above shows curves representing the likelihood between a false positive and true positive, if we were to take the red and blue curve as NDT methods, it could be assumed that the blue curved line represents a more accurate system as reaches higher values on the true positive axis.

3. Probability of Sizing

An excel sheet has been provided of percentage wall loss in oil storage tank plates where corrosion has caused this percentage wall loss. The data recorded was taken with three different methods, the first being magnetic flux leakage, and the other two being described as UT (ultrasonic test) operator A and B. Once the plate had been retrieved after testing all defects with the stated NDT methods the plate was now mechanically pit gauge tested which then gave a *true size* to use for calculations and comparison with the readings.

3.1. Methodology

The excel data was organised (see appendix 1) where an error value was calculated, plus and minus 95% values and some basic statistics on the error limits such as the mean and standard deviation.

From these statistical methods the graphs in the results were plotted.

3.2. Results & Analysis

The graphs were now graphically plotted from the excel data. Graphs were kept with equal x and y axis to give accurate visual representation of the data. Values of R^2 and $y = mx + c$ are also displayed on the graph to give more data.

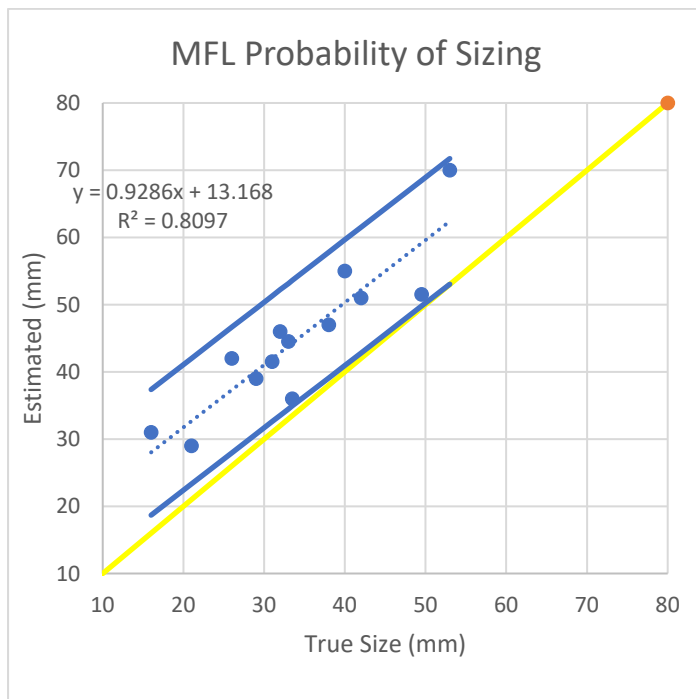


Figure 4 MFL PoS

MFL results are displayed above. Here the measures have been consistently over the true line (displayed in yellow) which suggests an issue with the machine being over calibrated. The reason the assumption of calibration being an issue and not operator skill level has been made is that on the whole the data points recorded are fairly consistent, whereas if a lack of operator skill level was being displayed there would more than likely be less consistency and wide confidence bands.

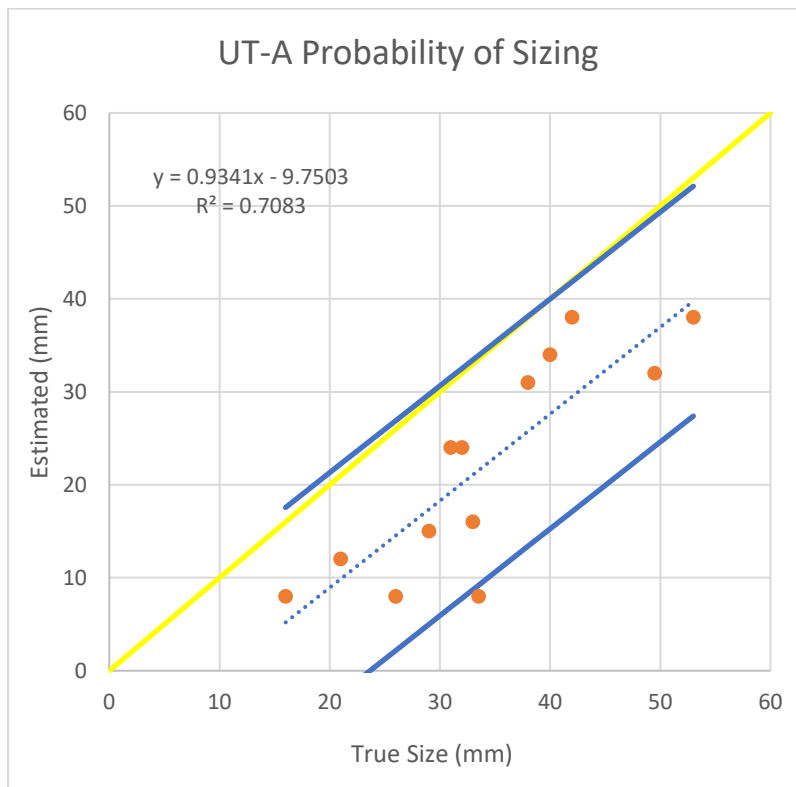


Figure 5 UT-A Graph

Graph above of ultrasonic UT-A shows that results have been consistently undersized as all points lay below the true line. There is also a wider confidence band in comparison to the MFL results which displays a high standard deviation which indicates that there has been inconsistency in the readings. Again, here it could be considered a case more of machine calibration rather than operator experience as the results are all consistently below the true size rather than some above and some below, but the wide confidence bands indicate a degree of operator inconsistency.

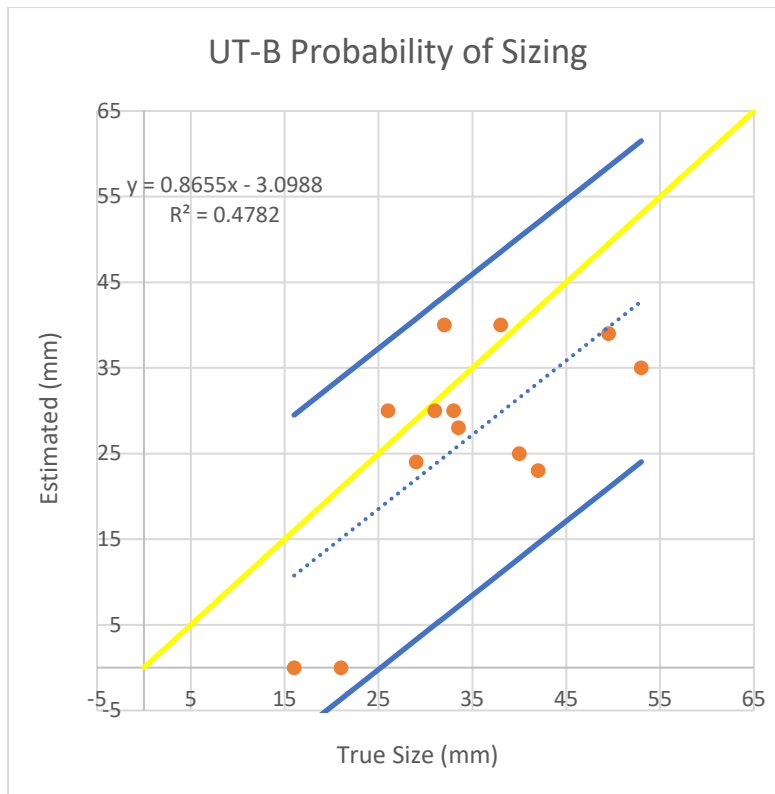


Figure 6 UT-B PoS

Ultrasonic operator-B displays points recorded both above and below the true line, at what appears to be random locations. The wide confidence bands and scattered data points can be put down as operator inexperience and inconsistency. The plot is showing the majority of the points below the true line, so an argument could be made for under calibration, but with the inconsistency of recorded data points the overall problem with the recorded data points could be considered to be operator inexperience/ inconsistency.

3.2.1. Line & Correlation Comparisons

It can be proven which system has recorded the most accurate data not just by visual analysis of the graphs but by comparing values of R^2 and $y = mx + C$. The R^2 values from each graph tell of

R^2 values demonstrate the predicative capabilities of a linear graph (people.duke.edu, 2001). The R^2 value is a statistical measure of how close the data points are fitted to the trend line, it is defined as the percentage of the response variable variation in a linear model. 0% would indicate that the model explains none of the variation, 100% would indicate that the model explains all variability of the data around it's mean. Therefore, it is the higher the R^2 value the better the model fits the data (minitab.com, 2013).

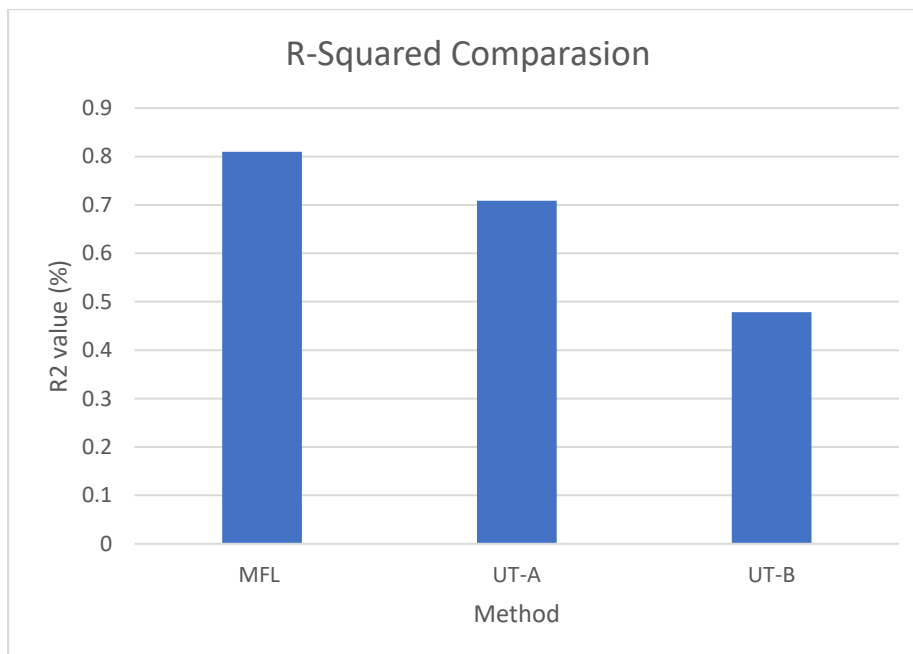


Figure 7 R-Squared Comparison

The graph above shows that the MFL recordings have the highest R^2 value therefore indicating the closest correlation of data between the thee methods. The data from UT-B recordings has the lowest R^2 result, this would be expected as there was little correlation between the data recorded.

3.2.2. Solutions / Compromises

Taking into account that there was an issue with the MFL system being over calibrated and UT-A under calibration a solution could be made from the test by drawing an average from the two tests. The average error was also taken away from the over calibrated MFL recordings also to assess how close a figure this would give to the true line. The graph below displays this average of the two sets of data.

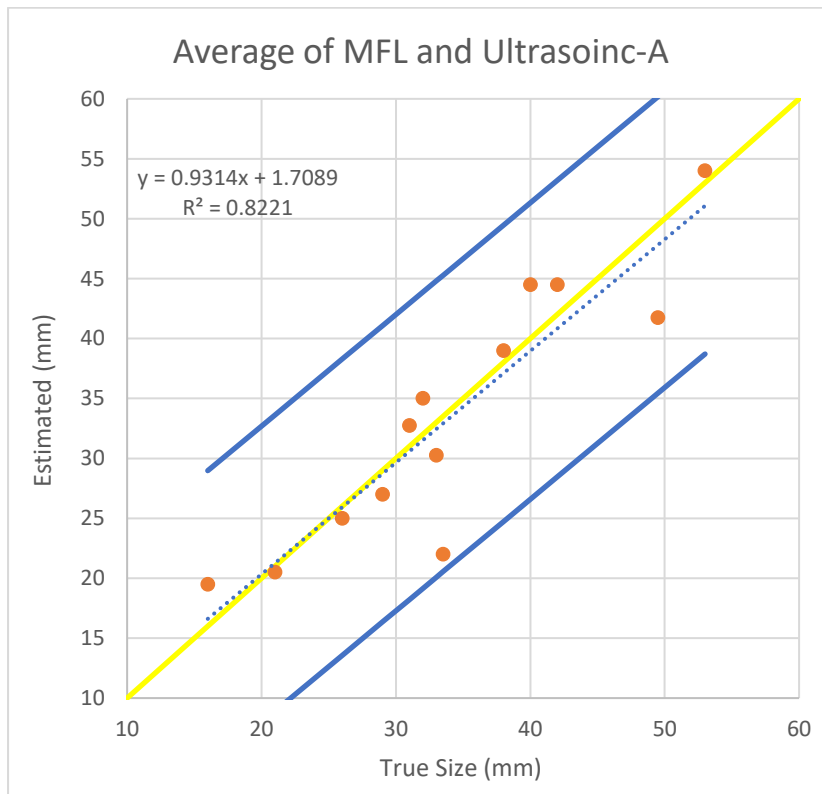


Figure 8 MFL & UT-A Average

From analysing the graphs of MFL and UT-A it could be seen that the MFL had oversized all measurements and the UT-A had undersized all measurements. With this in mind, an average of the undersized and oversized measurements could be taken which would then give figures closer to the true size line. Here a higher degree of accuracy has been achieved, but wide confidence bands still appear.

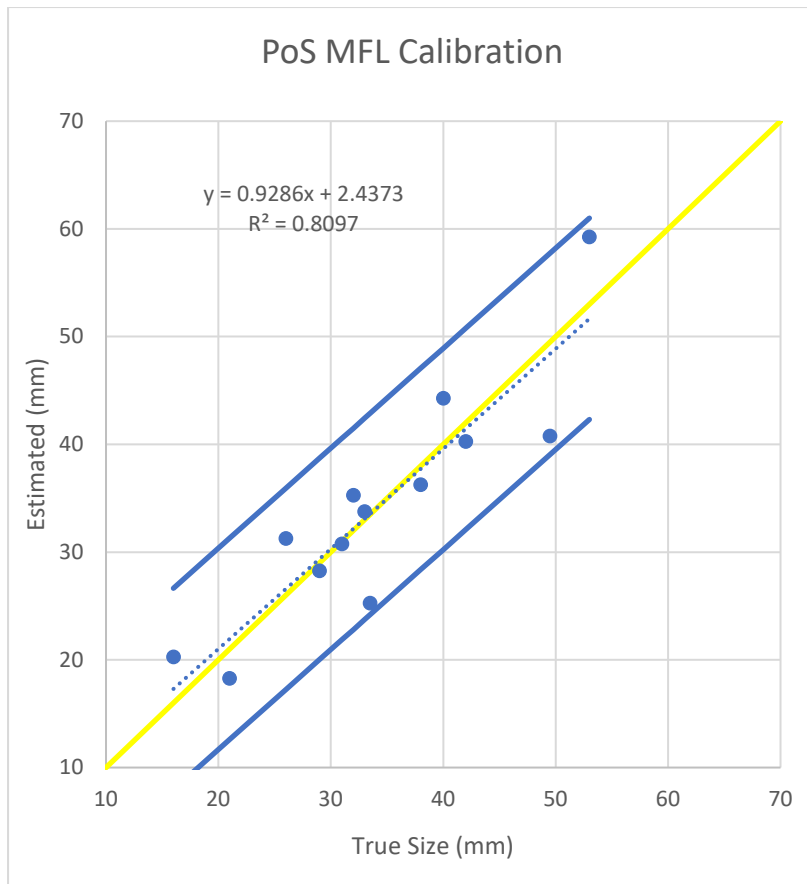


Figure 9 MFL Error Average

The graph above shows how the MFL PoS could be calibrated. Here the average error of the MFL readings were taken away from the actual recorded MFL readings as the equation below displays.

$$\text{MFL Calibrated PoS} = \text{MFL Recorded Values} - \text{Average of MFL Error}$$

The reason the average error was taken away from the readings is because in the first instance all recordings were oversized. The graph above displays how using information on the error involved in the MFL recordings could give a high level of accuracy with data points close to the true line.

4. Receiver operator Characteristics

Receiver operator characteristic (ROC) data has been given in excel (see appendix 2) from an aircraft wing inspection, it describes whether a defect was found or not found from 19 locations tested. By categorising the hits and misses of the defects found in the test the information could then be graphically plotted to define ROC characteristics.

4.1 Methodology

When inspecting a given sample with a particular NDT method it would appear there will either be readings of positive (defect found) or negative (no defect). But in reality, there can be four possibilities;

Table 1 RoC

True Positive (TP)	There is a defect present and the system has found it
False Positive (FP)	There is no defect present, but the system has recorded a defect
True Negative (TN)	There is a defect present and the system has recorded it
False Negative (FN)	There is a defect present, but the system has not recorded it

PoD (probability of detection) and PoFA (probability of false alarm) were used to analyse a set of data on an aircraft wing for three different NDT methods. PoD indicates how likely the system is to detect a flaw. PoFA identifies the likelihood of the system reading a false alarm. The two equations below were used to get the data plots for graphical analysis;

$$PoD = \frac{\text{Total number of positive calls}}{\text{Total opportunities for rejection}}$$

Where, total number of positive calls is the number of opportunities where defects are found, the total opportunities for rejection is the total number of flaws within the specimen.

$$PoFA = \frac{\text{Total number of false positive calls}}{\text{Total opportunities for false acceptance}}$$

Where, total number of false positive calls is the total number of sections where there is no defect present and a defect is found, Total opportunities for false acceptance is the total number of section where there are no defects present.

4.2 Results & Analysis

The results of the three NDT methods are shown in the form of an ROC plot in the graph below. The graph displayed does not show the *ROC Curve* spoke of in the literature review but displays single points plotted for different systems worked of the same principle.

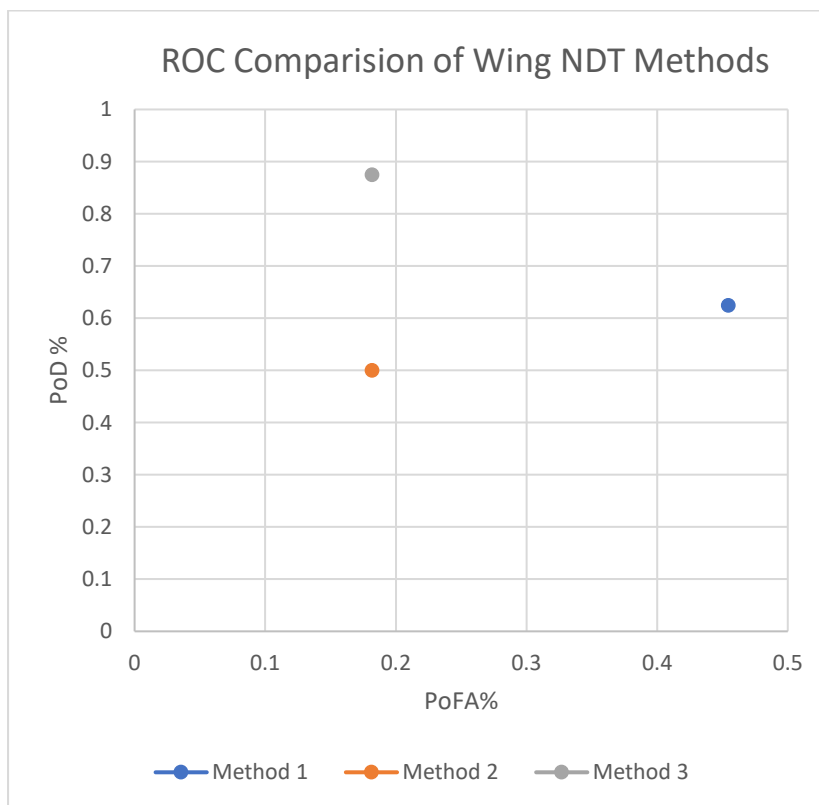


Figure 10 RoC Graph

From the graph above method one has the highest chance of a probability of false alarm, and method 3 has the highest chance of probability of detecting a flaw. Method two is displaying a lower probability of detection than method one but is less likely to give false alarm, therefore would be a better system to use for tasks even though it only has a 50% chance of finding a flaw.

Table 2 RoC Values

Method	PoFA Percentage Values	PoD Percentage Values
Method 1	0.45	0.63
Method 2	0.18	0.5
Method 3	0.18	0.88

The table above shows the values of PoD and PoFA percentage values, the numbers display how method three has the highest chance of probability of detection, here method one is showing high values of both PoD and PoFA meaning that there will be inconsistent results. Method two shows the same value of PoFA as method three but has a lower value, therefore with the trade off between

5. Discussion

By analysing operator recordings with statistical methods, it was proved that it could be defined which operator had recorded the most accurate data when these statistical methods were graphically displayed.

In the PoS section it was displayed how when the operators were recording consistent data narrow confidence bands would be displayed and a high value of R^2 . The R^2 gave reference to how well the data points were correlated to one another and justified what could be seen in the graphical display that the MFL system has recorded the most accurate data but oversized slightly, UT-A recorded slightly undersized data. Once this had been determined that an average was made of the two recordings, a graph was then also made by taking away the average error of the MFL recordings, as the MFL was deemed to be already the most accurate set of data taking away the average error gave the closest set of data points to what the true line was displaying.

The errors MFL and UT-A recordings were put down to either over calibration or under calibration issues, this could be stated as all recordings were either under or above the true size line. But UT-B recordings could not be attributed to calibration issues as recordings were above and below the true line, it was more likely a case of operator in experience.

In section four ROC was used to analyse the probability of sizing a defect for three operators, by creating a table of hits and misses for each operator coordinates could be plotted for percentages of probability of false alarms and probability of detection. This displayed that operator three had the best trade of between the chances of detecting a flaw and the chances of detecting a false alarm. By plotting the ROC points graphical representation gave reference to how much confidence could be placed in a system.

6. Conclusion

From the statistical analysis and research, it was proved that if PoS data is understood it will be possible to make informed decisions about the process i.e. calibration issues and operator competency. This information could be further investigated by comparing the operator experience levels to defect dimensions, then conclusions could be drawn upon such as how user friendly the equipment is and would it take an operator of high skill level to find a certain size defect, the statistical analysis helped tools such as confidence bands and R^2 values helped to highlight these.

Again, with the RoC there could be reference made to operator skill level and equipment used as to why certain operators performed better, by plotting this data it gave reference to how much confidence could be placed with an operator or system.

From the statistical methods it was able to make clear statements on which operators or systems recorded the most accurate sets of data during the report (MFL PoS, Operator three RoC) with the methods used. If more variables were to be added to the data sets such as operator skill level and operator experience we would further be able to analyse the usability of equipment and complexity of the task.

Bibliography

- itl.nist.gov. (2013, 10 30). *itl.nist.gov*. Retrieved from itl.nist.gov:
<http://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm>
- minitab.com. (2013, May 01). *minitab.com*. Retrieved from minitab.com:
<http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>
- people.duke.edu. (2001, 01 01). <https://people.duke.edu/~rnau/rsquared.htm>. Retrieved from <https://people.duke.edu/~rnau/rsquared.htm>:
<https://https://people.duke.edu/~rnau/rsquared.htm/~rnau/rsquared.htm>
- statisticshowto.com. (2011, January 01). *statisticshowto.com*. Retrieved from statisticshowto.com:
<http://www.statisticshowto.com/receiver-operating-characteristic-roc-curve/>.
- www.nde-ed.org. (2015, January 01). *www.nde-ed.org*. Retrieved from www.nde-ed.org:
<https://www.nde-ed.org/EducationResources/CommunityCollege/EddyCurrents/ProbesCoilDesign/ProbesModeOp.htm>

www.ndt.net-POD. (2006, May 01). *www.ndt.net-POD*. Retrieved from *www.ndt.net-POD*:
<http://www.ndt.net/article/v11n05/ginzel2/ginzel2.htm>

www.rfcafe.com. (2017, November 01). *www.rfcafe.com*. Retrieved from *www.rfcafe.com*:
<http://www.rfcafe.com/references/general/velocity-sound-media.htm>.

Appendix

1.

Defect	True Depth	UT - OpA	UT-A ERROR	Plus 95%	Minus 95%
1	16	8	-8	17.5591047	-7.1685047
2	21	12	-9	22.2296047	-2.4980047
3	26	8	-18	26.9001047	2.1724953
4	29	15	-14	29.7024047	4.9747953
5	31	24	-7	31.5706047	6.8429953
6	32	24	-8	32.5047047	7.7770953
7	33	16	-17	33.4388047	8.7111953
8	33.5	8	-25.5	33.9058547	9.1782453
9	38	31	-7	38.1093047	13.3816953
10	40	34	-6	39.9775047	15.2498953
11	42	38	-4	41.8457047	17.1180953
12	49.5	32	-17.5	48.8514547	24.1238453
14	53	38	-15	52.1208047	27.3931953
Mean		22.15385	-12		
STD DEV		11.61067	6.308063623		
±95%		22.75692	12.3638047		

2.

Aircraft Wing Inspection ROC Raw Data				
Location	Defect	Method 1	Method 2	Method 3
1	Y	1	0	1
2	Y	1	1	1
3	N	0	0	0
4	Y	1	1	1
5	N	1	0	0
6	Y	0	0	1
7	Y	0	0	1
8	N	0	0	0
9	N	1	0	0
10	N	0	0	0
11	N	1	1	1
12	N	1	1	1
13	Y	0	1	1
14	N	0	0	0
15	N	0	0	0
16	Y	1	1	0
17	N	0	0	0
18	N	1	0	0
19	Y	1	0	1

