8th International Symposium on NDT in Aerospace, November 3-5, 2016

Practical Experiences in POD Determination for Airframe ET Inspection

Virkkunen, I.[1] and Ylitalo, M.[2]

[1] Trueflaw Ltd., e-mail: iikka@trueflaw.com
[2]Patria Aviation Oy

Abstract
Evaluation of NDT reliability has received increasing emphasis in recent times. In particular, quantifying the probability of detection (POD) attained in routine inspections have become more widespread. Although there are good guidelines and standards for POD determination, the process is still far from trivial. Various choices made during the experimental set-up may have significant effect on the results. Also, the cracked samples used are often limited necessitating various compromises in the analysis.

Patria performed a set of POD studies for eddy-current inspections performed on various parts of typical metal airframe. The project included manufacturing of cracked samples, organizing the inspection of these samples and final analysis of the results. Several inspectors from different organizations took part in the exercise. The project was done in collaboration with Finnish and international partners.

The data showed various unlikely events (small hits, big misses and poor separation), which necessitated adjustment for the standard methodology. Contrary to expectation, the false call rate did not show significant correlation with the inspection performance. When the â vs. a and hit/miss analyses could be directly compared, they showed surprisingly poor correlation and caution is advised in using â vs. a analysis for manual inspections such as the ones shown here.

Keywords: Probability of detection (POD), Eddy current inspection (ET)

1. Introduction

The best practices of estimating probability of detection (POD) in non-destructive evaluation (NDE) are now well established. The venerable MIL-HDBK-1823A (most recent release from 2009) [1] is used extensively in the aerospace industry and is now finding increasing use also in other areas, like the rail industry and even nuclear industry. The methods have recently been standardized by ASTM (ASTM-E2862 [2] and ASTM-E3023 [3]) and these standards are congruent with the current MIL-HDBK methodology.

Despite the now standardized methodology and significant tradition in POD determination, the process still offers some practical challenges. The requirements for cracked test pieces are sometimes difficult or costly to fulfill, the statistical analysis may prove demanding and, perhaps most importantly, justifying that the various assumptions behind the methodology are fulfilled to sufficient extend may prove challenging.

The standard practice offers two variant of POD curve estimation, the â vs. a approach and the hit/miss approach. The â vs. a approach models, in simple terms, the NDE reliability as kind of measurement system problem, where the quantity to be measured (crack size a) give rise to measured signal (â) proportional to the measured quantity and the task is to determine the possible existence of the signal with decreasing a (and thus decreasing â). The system has noise both related to the signal and independent of the signal. That is, â varies due to factors other than a, like crack orientation and tortuosity, which results in noisy â vs. a relation. In addition, there's noise, that is independent of a, e.g. electric noise on the signal path. Thus, the task is to find a decision threshold (â value), that minimizes false calls from the noise and, in parallel, maximizes the number of cracks found (i.e. cracks with â above the threshold), given the

variation in the â-vs-a relation. This is done by fitting a linear function through the â-vs-a data, computing prediction intervals to take the noise and statistical uncertainty into account. The resulting best-fit and confidence limit lines are then compared to the set detection threshold and the corresponding POD curves computed. For input, the â vs. a analysis requires a set of representative flaws (at least 40) and measurements of signal strength â and corresponding crack size a. In addition, noise independent of crack size needs to be evaluated either with additional measurements of crack-free samples or in connection with the same sample set measurement.

The hit/miss approach, in contrast, does not deal with signal values, but estimates the POD curve based on binary results, that is hits (correctly found cracks) and misses (cracks not found in the inspection). Because the data contains less information (regarding the correlation between crack size and signal strength or "ease of detection") more samples are needed for reliable POD determination. The POD curve is solved using generalized linear model and a chosen link function (typically logit), that gives the shape of the POD curve using maximum likelihood fit to the data. The corresponding confidence limits are then obtained by the likelihood-ratio method, where a likelihood surface near the maximum likelihood value is interrogated, POD curves with likelihoods corresponding to the chosen confidence interval computed and the lower (and upper) limit curves resolved. For input, the hit/miss analysis requires a set of representative flaws (at least 60) and hit/miss results for each crack. In addition, the hit/miss results should exhibit a range with "unlikely to find" cracks, a range with "likely to find" cracks and transition in between. Otherwise, the model does not describe the data and, while a fit may in some cases be obtained, it does not describe the underlying probability of detection.

In both cases, the basic assumptions underlying both POD models should be fulfilled: the POD should be an increasing function of the crack size and should reach 100% with sufficient crack size. If the data contains signs of violation of these assumptions (e.g. a miss with big crack length indicating that the POD does not reach 100% even with large crack size), the standard models are not applicable and alternate model must be sought.

Patria performed a set of POD studies for eddy-current inspections performed on various parts of typical metal airframe. The project included manufacturing of cracked samples, organizing the inspection of these samples and final analysis of the results. Some of the cracked samples were provided by collaborating organizations. Several inspectors took part in the exercise from different organizations. The project was done in collaboration with Finnish and international partners. This study provided an opportunity to study various practical aspects of the POD determination process and to compare POD results obtained in different settings.

2. Materials and methods

The study was divided in three cases, as described in Table 1. Each case had different set of cracked samples and was completed in one "go". The analyses were completed with the openly available mh1823 software package [4].

Table 1. Summary of studied inspection cases.

Case	Description	Cracks	Inspectors
A	Typical fillet	49	7
B	Rivet hole	58	11
C	Rivet hole	68	7

3. Results and discussion

For each case, the practical difficulties obtained were somewhat different and are analyzed on case-by-case basis below.

3.1 Case A

For case A a hit/miss analysis was completed. The sample set contained somewhat smaller number of cracks (49) than required by the MIL-HDBK [1]. However, the crack sizes were well distributed in terms of hits and misses and showed no adverse behavior in the statistical analysis. Thus, hit/miss analysis was deemed appropriate for the data. The lack of sufficient cracked samples increases the uncertainty of the analysis and thus the reported $a_{90/95}$ values are expected to be greater than what would be obtained with additional number of samples. The cracks used were produced using mechanical fatigue. Produced cracks were inspected using automated ET and selected cracks were destructively examined.

A typical obtained POD curve is shown in Figure 1. The data shows good separation between crack sizes likely to be missed, a transition zone and crack sizes likely to be found. On several cases, the inspectors also found some very small cracks. This unlikely hit significantly changed the confidence bounds and, paradoxically, increased the size of the computed $a_{90/95}$. Such curves were re-analyzed with the small hit changed to miss (thus "worsening" the inspection behavior) and smaller $a_{90/95}$ values were obtained.

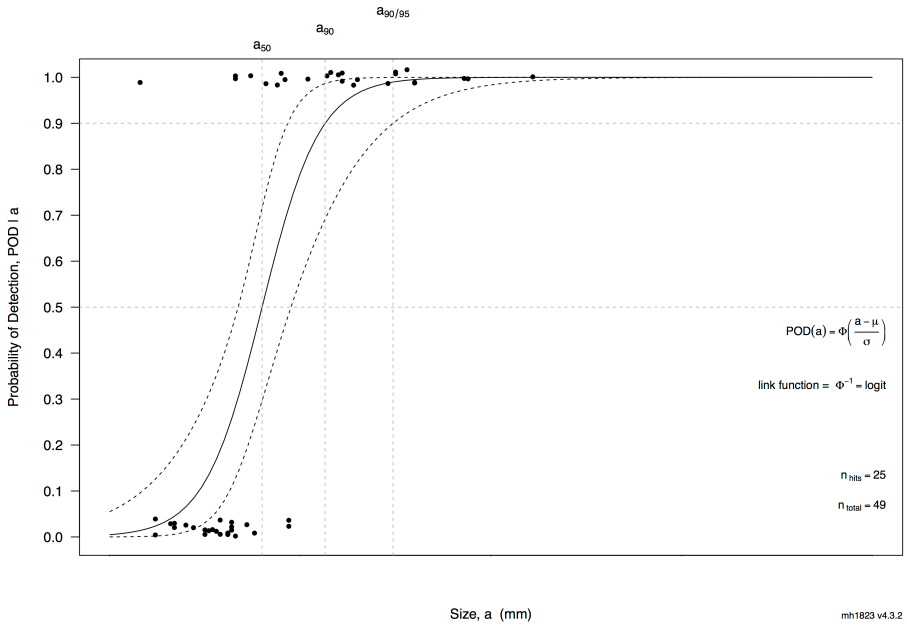

Figure 1. Typical POD curve for case A. The curve includes an unlikely small hit, which widens the confidence bounds and paradoxically decreases the measured performance. With this hit changed to miss, the $a_{90/95}$ value decreased by 14%.

The inspectors also showed strong variation in false call rates. Interestingly, the false call rate was not correlated with inspection performance (as measured by the $a_{90/95}$). Figure 2. shows the obtained $a_{90/95}$ values in comparison to the false call rate of the inspectors.

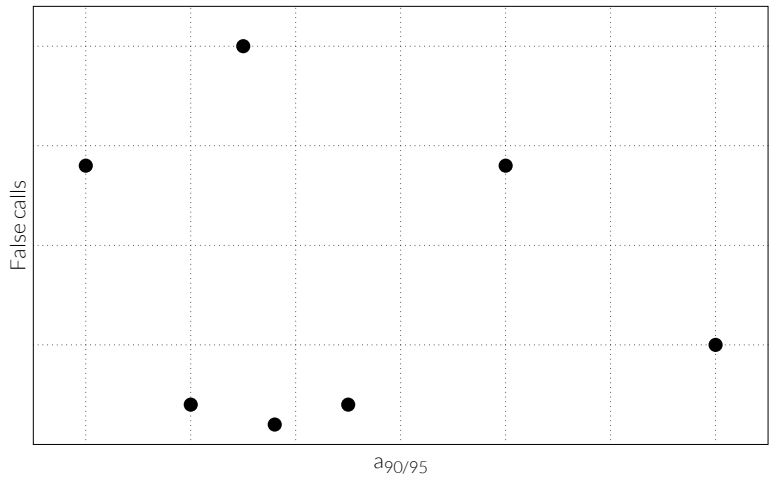

Figure 2. False call rate as a function of obtained $a_{90/95}$ *values. Negative correlation would be expected, but there's no clear correlation observed.*

The overall performance was not quite as good as was hoped. This was attributed partly to significant time pressure during the inspection.

3.2 Case B

For case B a hit/miss analysis was completed. The sample set contained somewhat smaller number of cracks (58) than required by the MIL-HDBK [1]. However, the crack sizes were well distributed in terms of hits and misses and showed no adverse behavior in the statistical analysis. Thus, hit/miss analysis was deemed appropriate for the data. A typical obtained POD curve is shown in Figure 3. As in the case A, the data shows good separation between crack sizes likely to be missed, a transition zone and crack sizes likely to be found. The overall results were significantly better than for Case A.

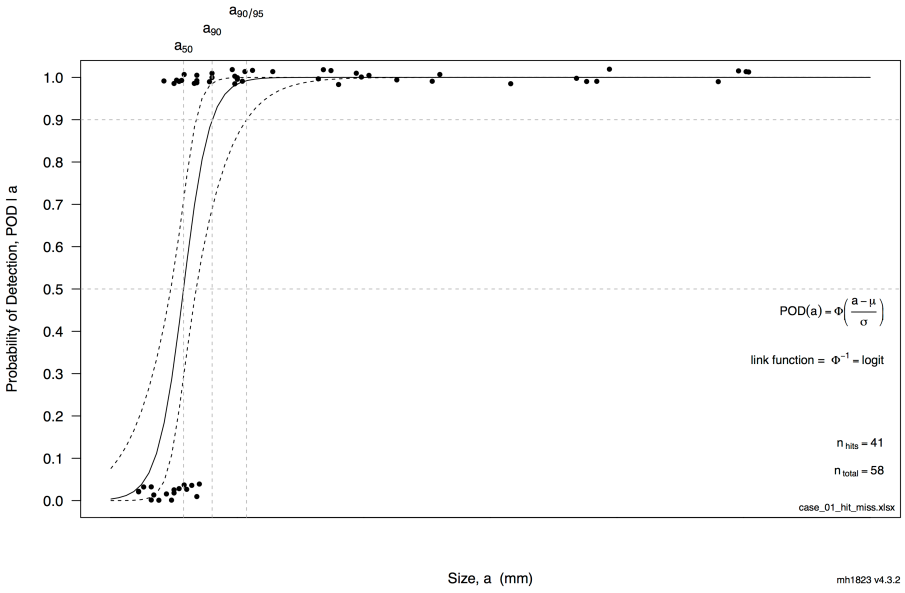

Figure 3. Typical POD curve for case B.

In case of one inspector, a single large (or larger than other) crack was missed. This unlikely event significantly changed the maximum likelihood curve and widened the confidence bounds. The resulting $a_{90/95}$ value was quite large for this data set, but with the single big miss changed

to hit, decreased by 52%. Furthermore, with the one outlier, the confidence bounds do not seem to sufficiently cover the variability: the biggest miss is still significantly over the computed $a_{90/95}$ value and would be highly unlikely, according to the computed POD curve. Thus, the single big miss effectively calls to question the applicability of the POD model used. A comparison is shown in Figure 4.

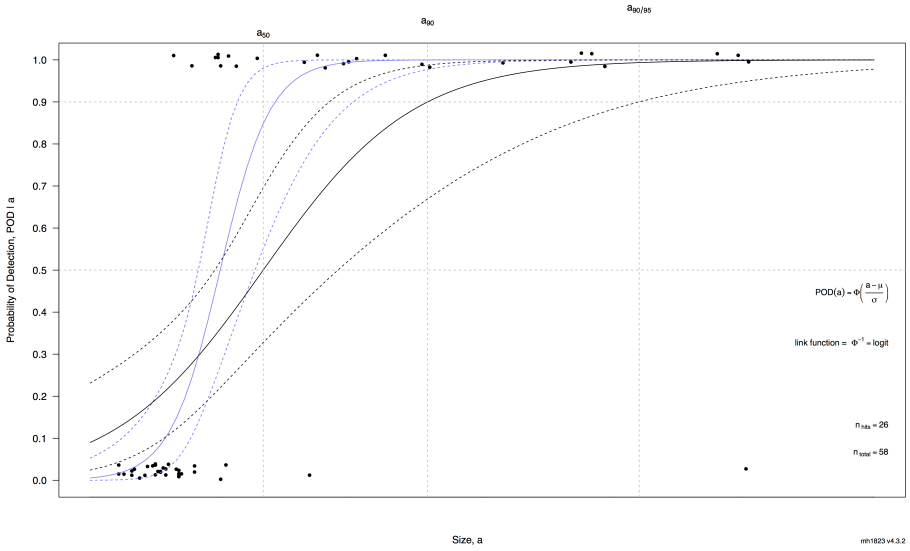

Figure 4. Atypical POD curve, where a single big miss has significantly altered the obtained POD curve. For comparison, curves computed with the same data except the biggest miss changed to hit are shown in light-blue. With the modified data, the computed $a_{90/95}$ decreased by 52%.

Again, the inspectors also showed strong variation in false call rates and the false call rate was not correlated with inspection performance (as measured by the $a_{90/95}$). Figure 5. shows the obtained $a_{90/95}$ values in comparison to the false call rate of the inspectors.

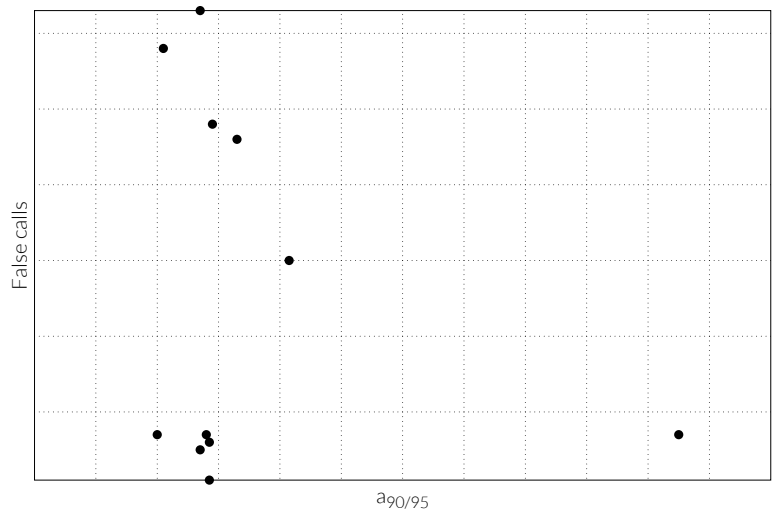

Figure 5. False call rate as a function of obtained $a_{90/95}$ values. Negative correlation would be expected, but there's no clear correlation observed.

3.3 Case C

For case C both â vs a and a hit/miss analysis were completed. The sample set contained 68 cracks. However, the inspection performance in this case was better than expected and most inpsectors found most of the cracks and some inspectors found all of the cracks. Paradoxically, this good performance caused numerical difficulties with the POD curve determination (for the hit/miss analysis) since now the crack sizes were not well distributed in terms of hits and misses, despite large population of small crack sizes. To obtain maximum likelihood estimates for the hit/miss analysis, a single small miss was added to the data. (Had there been such a small crack in the samples, it would have likely been missed by both the inspection under investigation and the previous after-manufacturing inspection. Thus assuming such a miss is not unrealistic.) This single miss allowed the maximum likelihood estimate to converge and provided $a_{90/95}$ results. In practice, the induced single miss forced the POD curve to steep curve between the artificial miss and the smallest found crack and the lower-limit estimate near the second-smallest crack missed. Thus the shape of the POD curve does not carry much information, but the obtained $a_{90/95}$ results are justifiable. A typical POD curve is shown in Figure 6.

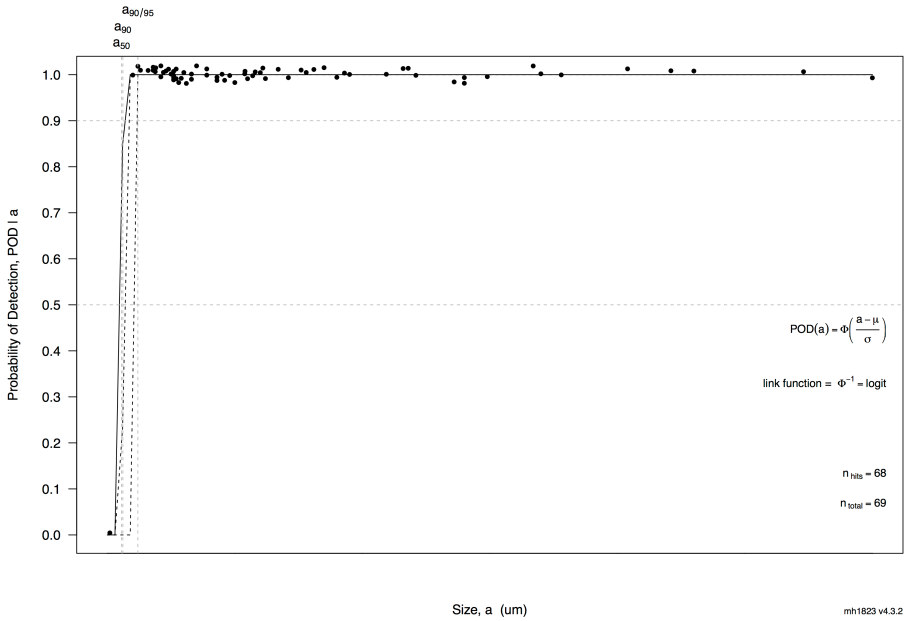

Figure 6. Typical hit/miss POD curve for case C.

For case C, results enabled both â vs. a and hit/miss analyses to be completed and allowed direct comparison between the two methodologies. Figure 7. shows typical POD curve obtained from â vs. a analysis.

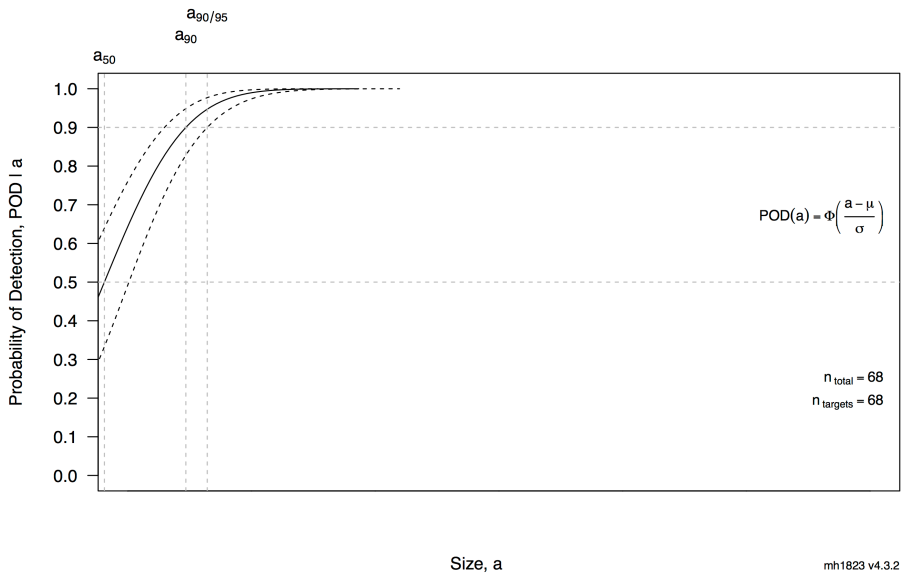

Figure 7. Typical â vs. a POD curve for case C.(Same inspector as for Figure 6 for direct comparability.)

To compare the POD values obtained from â vs. a and hit/miss analysis, the values were compared inspector-by-inspector. The results are shown in Figure 8. Although both results were obtained with standard methodology, they present significant variation and the overall correlation is not very good. For the very small $a_{90/95}$ sizes (where inspectors found all or almost all the cracks), the hit/miss analysis shows smaller (better) $a_{90/95}$ values indicating, that the inspectors included factors other than the signal strength â for their judgement (e.g. signal stability in repeated measurements were sited). Conversely, for the larger $a_{90/95}$ values, the â vs. a results show smaller $a_{90/95}$ values. This can be attributed to the â vs. a methodology failing to account sufficiently to the larger missed cracks in the data. Furthermore, the â vs. a is sensitive to variation in the â vs. a relation, which in this case was also affected by inspector reporting practices. Values of â were read from the equipment screen and there may have been differences of accuracy between inspectors in this respect. This accuracy did not affect the inspector performance (as shown in hit/miss), but it did affect the confidence bounds obtained from â vs. a and thus measured performance. On one occasion, the reported â vs. a relation showed significant non-linearity and the reliability of the â vs. a was questionable (despite this non-linearity having no effect on actual inspector performance). In conclusion, the hit/miss method seems to better describe the present manual inspection case and caution is advised if using â vs. a for such cases.

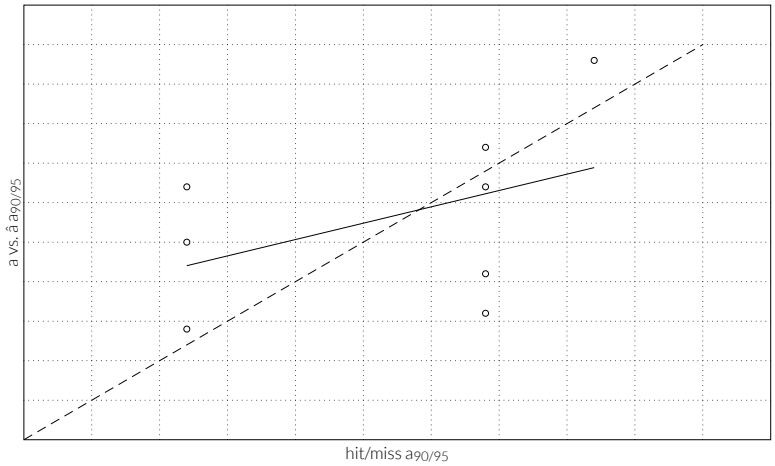

Figure 8. Comparison of â vs. a and hit/miss $a_{90/95}$ results.

Finally, case C was, on the surface, very similar in inspection arrangement with case A. However, inspectors showed significantly better performance in Case C. This can be attributed to smaller time-pressure during case C, possible learning from earlier cases and differences in samples. This indicates, that rather small changes in the set-up or inspector feedback can have significant impact to the obtained POD performance.

4. Conclusions

Three separate POD exercises were completed, each showing separate experimental challenges and solutions. The following conclusions may be drawn from this study:

- Contrary to expectation, the false call rate did not show significant correlation with the inspection performance.
- The data showed various unlikely events (small hits, big misses and poor separation), which necessitated adjustment for the standard methodology. This shows, that the standard POD methodology can not be used as a "black box", and must be accompanied by careful analysis of the underlying technical and statistical factors leading to the obtained POD values.
- â vs. a and hit/miss analyses showed surprisingly poor correlation and caution is adviced in using â vs. a analysis for manual inspections such as the ones shown here.
- Small changes in the inspection set-up (e.g. time pressure) can have significant impact to the obtained POD performance.

Acknowledgements

This work was supported by several partners who provided test pieces and inspection results for the study. Their support is gratefully acknowledged. In particular, the authors wish to thank RUAG aviation, Switzerland for providing some of the test samples for this study.

References

1. Anon. 2009. Nondestructive Evaluation System Reliability Assessment. Department of Defense Handbook. MIL-HDBK-1823A. 171 p.
2. Anon. 2012. Standard Practice for Probability of Detection analysis for Hit/Miss Data. American Society for Testing and Materials, ASTM E2862-1
3. Anon. 2015. Standard Practice for Probability of Detection Analysis for â Versus a Data. American Society for Testing and Materials, ASTM-E3023
4. Annis, C., 2015. mh1823 R software package, version 4.3.2. Available online: http://statisticalengineering.com/mh1823/mh1823-algorithms.html