Virkkunen, Iikka; Koskinen, Tuomas; Papula, Suvi; Sarikka, Teemu; Hänninen, Hannu

Comparison of â Versus a and Hit/Miss POD-Estimation Methods : A European Viewpoint

Published in:
Journal of Nondestructive Evaluation

DOI:
10.1007/s10921-019-0628-z

Published: 01/12/2019

Document Version
Publisher's PDF, also known as Version of record

Journal of Nondestructive Evaluation (2019) 38:89
https://doi.org/10.1007/s10921-019-0628-z

Comparison of â Versus a and Hit/Miss POD-Estimation Methods: A European Viewpoint

Iikka Virkkunen[1] · Tuomas Koskinen[2] · Suvi Papula[1] · Teemu Sarikka[1] · Hannu Hänninen[1]

Received: 23 March 2019 / Accepted: 23 August 2019

Abstract
For estimating the probability of detection (POD) in non-destructive evaluation (NDE), there are two standard methods, the so-called â versus a approach and the hit/miss approach. The two approaches have different requirements for the quality and quantity of input data as well as for the underlying NDE method. There is considerable overlap between the methods, and they have different limitations, so it is of interest to study the differences arising from using each methodology. In particular, if the dataset is not ideal, the methodologies may exhibit different problems dealing with various limitations in the data. In this paper, a comparison between â versus a and hit/miss analysis was completed for two different data sets, a manual aerospace eddy-current inspection and a nuclear industry phased array ultrasonic weld inspection using a simplified online tool. It was found that the two standard methods (â vs. a and hit/miss) may give significantly different results, if the true hit/miss decision is based on inspector judgement and not automated signal threshold. The true inspector hit/miss performance shows significant variance that is not attributable to signal amplitude. Model-assisted POD was not able to model the inspector performance due to lack of representative amplitude threshold and difficulties in capturing true signal variance. The paper presents experience from practical cases and may be considered a European viewpoint.

Keywords Non-destructive testing · NDT · NDE · Probability of detection · POD · Reliability

1 Introduction

The best practices of estimating probability of detection (POD) in non-destructive evaluation (NDE) are now well established. The longstanding MIL-HDBK-1823A [1] is used extensively in the aerospace industry [2–4] and is now finding increasing use also in other areas, like the rail industry [5, 6] and Nuclear industry [7]. The methods have recently been standardized by ASTM [8, 9] and these standards are congruent with the current MIL-HDBK methodology.

The standard practice offers two variant of POD curve estimation, the so-called â versus a approach and the hit/miss approach. The â versus 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 on the signal (â varies due to factors other than a), which results in noisy â versus a relation. In addition, there's noise, that is independent of a. 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 versus a relation. The ASTM-E3023 (and MIL-HDBK) solve this by fitting a linear function through the â versus a data, computing prediction intervals to take the notice 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. Improvements to the classical Berens [10] model have been proposed to behave better with very limited data sets, e.g. by Syed Akbar Ali et al. [11], Syed Akbar Ali and Rajagopal [12] and Le Gratiet et al. [13].

For input, the â versus a analysis requires a set of representative flaws (at least 30) 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

✉ Iikka Virkkunen
iikka.virkkunen@aalto.fi

[1] Aalto University, PL 14200, 00076 AALTO Espoo, Finland

[2] VTT Technical Research Centre of Finland, PL 1000, 02044 VTT Espoo, Finland

Published online: 12 September 2019

measurements of crack-free samples or in connection with the same sample set measurement.

Recently, the cost and lack of representative test pieces has been alleviated by using simulated inspection results in lieu of actual physical test samples and measurements. This is called model-assisted POD (MA-POD), and is widely applied in different contexts. Typically, a simulation is used to provide â versus a data for the inspection case of interest. Formulating the computation involves several simplifications to the physical reality to make the modelling effort feasible. These include simplified physical models to describe the inspection signal (e.g. wave propagation laws), simplified material data (e.g. homogeneous and isotropic material instead of the actual inhomogeneous material) and simplified flaw description (e.g. a simplified notch-like reflector instead of tortuous and branching crack). Due to these simplifications, the simulated data is normally free from noise and exhibits no variation for given flaw size and configuration. The variance in â versus a is introduced by varying flaw configuration parameters, e.g. flaw tilt or skew angle and location in relation to geometric features. The variation in flaw configuration then produces variation in â versus a dependence, which is then used to compute â versus a POD curves using standard methodology. Simulations enable computation of large number of cases and thus the statistical sampling error in the results can be decreased to arbitrary low values.

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 (at least 60 [14]). Some statisticians have recommended sample sized over 300 for hit/miss [15], especially if the 95% confidence limits on POD curves are calculated according to MIL-HDBK-1823. There are also improved statistical methods proposed in the literature, allowing POD curves to be reliably determined from data sets of even as few as 50 hit/miss observations [16]. The POD curve is solved using generalized linear model and a chosen link function (typically logistic, but sometimes probit), that gives the shape of the POD curve using maximum likelihood fit to the data. The corresponding confidence limits are the 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 solved. The number of samples and the flaw size distribution in relation to the actual POD curve also affect the width of the confidence bounds [14].

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 logistic (or probit) 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. Such alternate models exist, among others, for POD with limited maximum POD etc.

With the two models available, the user has a choice of method to make. In many cases, the choice is predetermined by the available data, i.e. signal values or hit/miss data. However, especially when designing a POD determination project, both methods may be available and they may give different results. The difference and, indeed, the validity of each method may be difficult to judge beforehand and if (as is often the case) only one is completed the possible difference remains unknown.

The two methods have different requirements for the underlying NDE method as well. The â versus an assumes, that the method can be modelled by a single detection threshold and â versus a correlation. In many cases, the inspectors use information other than the signal strength to judge crack existence, and in such cases the â versus a does not describe the true performance of the system. Even more disturbingly, the measurement of the signal may also be affected by the inspector. For example, in manual EC inspection, the inspector often may receive spurious signals from small aberrations on the surface etc., and will compensate by doing repeated measurements and reporting the "correct" signal. Thus, the true noise is not recorded and is already filtered in the inspector reporting. Similarly, when inspector judges a crack to be present, again repeated measurements are taken to find the "correct" signal strength. Again, the â versus a correlation is distorted by inspector judgement. Consequently, the â versus a methodology is mostly applicable for highly automated systems, where human intervention is insignificant and single detection threshold fully describes the crack detection process. However, in practice it may be difficult to assert the absence of human intervention. In addition, the â versus a methodology requires fewer samples than the hit/miss and thus there may be a preference for using it even when the inspection is not fully automated.

In contrast, the hit/miss analysis deals with the direct results of the inspection (i.e. hits and misses) and thus may incorporate information and variance of inspector judgement. Thus much less assumptions are made regarding the inspection method or hit/miss judgement and the method is more

robust in this respect. The inspection system can, in theory, be regarded as a "black box" and POD evaluation is done with the end results only. Thus, even if some of the aspects affecting inspector judgement are unknown, this does not jeopardize the validity of the analysis. At the same time, often the method is not a black box, and there may be significant information available, that describes the relevant "ease" of the detection (e.g. signal strength obtained), that is not incorporated into the analysis and thus, in effect wasted. Thus the analysis may seem wasteful.

Since there is considerable overlap between the two methods, and since they have different limitations, it is of interest to study the differences arising from using each methodology. In particular, if the dataset is not ideal, the methodologies may exhibit different problems dealing with limitations in the data. The source of variation in the POD estimation can be attributed to several distinct sources as follows:

- *Statistical sampling error:* the error caused by limited set of available data for the approximation of the POD curve.
- *Measurement variation:* for given crack and measurement set-up, there may be variation in the obtained â values due to operator variability, equipment calibration differences etc.
- *Configurational variation:* for given nominal crack size, there may be variation due variation in the crack orientation and location.
- *Variation in crack characteristics:* natural cracks may exhibit differing â for same nominal crack size a. Various factors besides the crack size may affect the obtained sample amplitude (e.g. crack path tortuosity, opening, surface roughness etc.). In some cases, this is modelled directly with so-called multivariate POD curves, where the POD curve is explicitly stated and modelled to be function of crack size and other chosen parameters. For standard analysis, these other features are represented as variation in the â versus a correlation and affect the confidence bounds calculated for the POD.
- *Inspector judgement variation:* for given â obtained from the inspection, in many cases there is an element of inspector judgement in translating the obtained signal strength to "cracks/no crack" -judgement. This may depend on the local data variance, noise surrounding the flaw, inspector variability etc. Even when there's an explicit detection threshold set, in manual inspection the inspector needs to separate the true indication from possible spurious signals. Thus the recorded signal may be affected by the inspector judgement and crack signals may be overlooked as artefacts.

Table 1 shows a comparison of â versus a and hit/miss in terms of how these sources of variation are handled. The statistical sampling error is explicitly handled with both â versus a methodology. The variation in crack characteristics and the measurement variation are directly measured in both methodologies and thus can be considered to be contained within the statistical scatter and confidence bounds, although there are differences, e.g. with the number of cracks used. The biggest difference is in the inspector judgement variation. In â versus a, this effect is assumed to be negligible, whereas for hit/miss the variation is included in the data. Thus, the main focus in selecting between â versus a and hit/miss is related to whether this effect can be assumed to be negligible.

Table 1 Comparison of â versus a and hit/miss in terms covered sources of variation

Source of variation	MA-POD	â versus a	Hit/miss
Statistical sampling error	N/A	YES	YES
Configurational variation	YES	YES	YES
Measurement variation	NO	YES	YES
Variation in crack characteristics	NO	YES	YES
Inspector judgement variation	NO	NO	YES

Finally, for the MA-POD case, the variation is assumed to come from configurational issues (flaw tilt, skew, etc.). The statistical error can be decreased by additional simulations, which is cheap in comparison to manufacturing physical test samples. On the other hand, variation in crack characteristics, sample microstructure and possible measurement system issues are not included.

Inspections conducted by human inspectors are known to exhibit variability in inspector judgement. The variability is seen between different inspectors and between different inspections carried out by the same inspector (see e.g. [17, 18]). This variability resulting from inspector variance is often referred to as "human factors" effect. As shown in Table 1, in the hit/miss approach, inspector judgement is directly included in the results and thus any possible human factors that are present during the exercise are reflected in the POD results. However, inspection conditions during a POD exercise are seldom identical to the real inspection conditions, even when care is taken to make the exercise as representative as possible. Most notably, the number of flaw findings in a POD exercise is typically much higher than in normal inspections, which may affect inspector expectations.

For the â versus a approach, the human factors are not, in general, included. When the â values are sourced from modelling or automated inspection systems, the human factors are not included in the analysis and need to be addressed separately, if the results are to be used in conditions, where human inspectors report â or make flaw decisions. If the â values are sourced from human inspectors, variation in human judgement may affect the results [19].

Over the years, several modifications to the traditional â versus a and hit/miss methodologies have been proposed to

overcome some of its deficiencies. The statistical methodology and, in particular the computation of confidence bounds have been evaluated and alternate methods proposed [11–13, 16, 20–22]. In particular, the focus has been to obtain more robust confidence bounds, in case of small dataset using, e.g. the Bayesian approach.

One of the key assumptions of the hit/miss approach is, that the probability of detection increases monotonically with increasing crack size a. Furthermore, it is typically assumed, that the POD reaches 100% at some crack size a. In reality, there may be error sources, that do not follow such dependence on crack size. Generazio [23–25] proposed alternate formulation based on design of experiment (the design of experiment probability of detection, DOEPOD). The DOEPOD model is based on extending the binomial view of hit/miss data. The main motivation for the DOEPOD model is, that using model-based POD estimation (e.g. ASTM E2862) assumes POD as a function of flaw size follows certain model. In particular, the POD is continuous, monotonically increasing function of flaw size a. This assumption may not always be justifiable, e.g. when the method sensitivity varies for different flaw sizes due to different probes, beam focusing or for some other reason. The DOEPOD model does not assume functional relationship between POD and flaw size. Instead, the inspection results are grouped and analysed, simply stated, as groups to make sure that the binomial 90/95% condition is fulfilled for certain flaw size ranges.

Despite more recent formulations, the MIL-HDBK/ASTM methodologies [1, 8, 9] are still widely used and thus it is of interest to still study and better understand their limitations.

In this paper, a comparison between â versus a and hit/miss analysis was completed for two different data sets. The first data set describes a manual aerospace eddy-current inspection. In this data-set the signal strength is recorded by manual process and thus there is potential for inconsistencies in the â versus a relation. At the same time, the inspection could also be analysed with â versus a methodology, since the procedure does provide signal strength and process calibration and procedure definition are expected to minimize any such operator effects. The data set was arranged and gathered by Patria Aviation and The Finnish Defence Forces in conditions resembling true inspection as closely as possible by introducing test samples into representative locations on airframes and testing them in proper locations.

The second data-set represents a nuclear industry phased array ultrasonic weld inspection and was collected using a simplified online tool. The underlying data has limited real flaws. However, unlike traditional inspection records, the online analysis methodology provided direct opportunity to study the relationship between signal strength and hit/miss judgement with large number of artificially generated flawed data images. These results were compared with model-assisted POD results generated for the same inspection case.

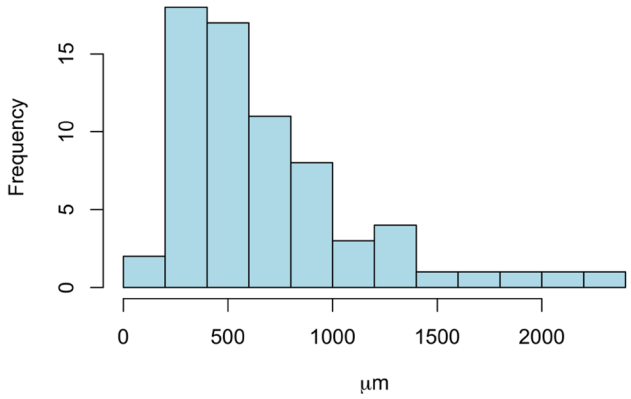


Fig. 1 Crack size distribution

2 Materials and Methods

Two data sets were applied for this study, designated “EC” and “UT”. The EC data set was collected using manual eddy current representing aircraft body inspection. The inspectors were EN 4179 certified level 2 or level 3 inspectors. Each inspector completed the inspector using the normal equipment in his/her use (GE Mentor, Olympus Nortec 600, GE Phasec 3 or equivalent). A rotating probe was used and systems were calibrated to aluminium reference standard with 0.5 mm artificial defect corresponding to 100% of display. The used frequency was 500 kHz and scan rotation 1000 rpm.

A set of cracked samples representing typical rivet hole configuration were prepared using mechanical fatigue loading. The existing cracks were characterized using microscopy to define the true state of the samples. The small plate samples with inspection targets (the rivet holes) were gathered to larger cassettes, which were attached to representative aircraft body location for inspection. Altogether 5 inspectors completed the inspection and reported both signal strength for each inspected hole and their judgement (crack/no crack). Thus the data provided input information for both â versus a and hit/miss analysis. The data set contained altogether 68 cracked locations, 480 inspection locations and 3360 inspection results for 7 inspectors. The crack size distribution is shown in Fig. 3. More detailed description of the inspection set-up is given in [26] (Fig. 1).

The second data-set represents a nuclear industry phased array ultrasonic weld inspection case. The use of POD methodology is not as common in nuclear industry as it is in aerospace industry. Thus, representative sample sets containing sufficient flaws for hit/miss analysis (or even for â vs. a analysis) are rare. The present data was gathered with simplified online tool described in the following.

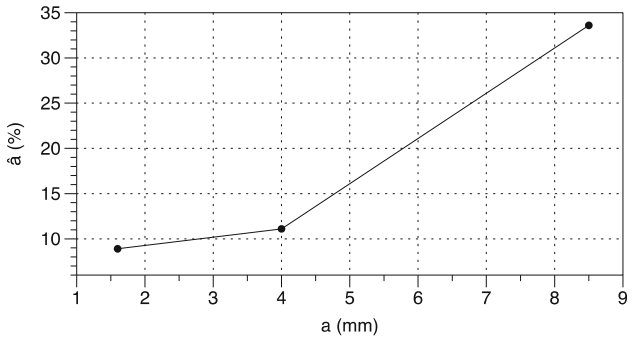


Fig. 2 Measured UT amplitude as a function of crack size. The cracks show significant variation of amplitude even with the small number of real flaws available

For the nuclear inspection case, an austenitic stainless steel butt-weld mock-up representing primary circuit piping was available. The mock-up had three cracks (which is obviously far too small number for direct analysis). The mock-up was scanned with mechanized ultrasonic system and collected A-scans recorded in a data file for later analysis. To compensate for the insufficient number of real cracks in the mock-up, the data file was modified to include additional flaws. Also, for easy collection of hit/miss data, an online tool was created, that provided a simplified set of UT analysis tools necessary for this inspection case and provided tools for crack identification and data gathering. The tool can be accessed online at (http://www.trueflaw.com/utpod/).

The data file provided by the UT equipment was read and the raw UT-data extracted. The locations of the known cracks were compared with known un-cracked locations, and pure flaw signal was extracted by comparison. The flaw signal was then removed from the original measurement data resulting in apparent clean mock-up data. This pre-processed data, with extracted flaw signals and cleaned, flawless, UT-signal provided the source data for the online tool. The data extraction and details on the used data are provided in [1].

For this methodology, the â versus a relation is predetermined for each flaw. Furthermore, the number of different cracks is too small to allow analysis of variability of real â versus a relation was not available. Thus, a direct comparison of â versus a and hit/miss analyses was not possible with this data set. However, with this set-up (and the postulated â vs. a relation), it was possible to study directly the effect of the â to the POD of these inspections. For conventional â versus a analysis, this effect is implicitly assumed to be negligible and thus this allows direct estimation of the possible error caused by this assumption.

Even with the very limited number of cracks available in the data set, the natural cracks exhibited significant variation in the maximum amplitude, as compared to the flaw depth. Figure 2 shows the measured UT amplitude as function of crack size.

For comparison, similar inspection case was modelled using commercially available CIVA software. Artificial reflectors of different height, tilt, skew and location were introduced to the model and resulting expected signal strengths computed. This provided model-assisted â versus a data for the same inspection case. Data were used to compute a corresponding â versus a MA-POD-curve for the case. Where applicable, the â versus a and hit/miss analyses were completed using Military handbook software [2]. Several software packages are available for performing these computations, but the MH1823-package is perhaps the most widely applied and thus it was chosen for this study. Where MH1823-package was not sufficient, in-house developed code was used to augment the analysis using the same mathematical methodology.

3 Results

For the data set EC, the actual hit/miss performance was very good and the inspectors detected very small cracks bordering on the resolution of the microscopy used to confirm the real state of the samples. Some of the inspectors found all the cracks. Consequently, the requirement of having separate regions and transition in the data was not fulfilled. To alleviate this, and to obtain hit/miss results, a single virtual miss was added to small crack size (10 μm) that, if present, would probably be missed. This addition has little effect in cases where real misses exist, because the real misses have much larger crack size and thus dominate the POD fit. However, for the cases with no misses, this allows the maximum likelihood fit to converge and provides sensible POD values and confidence bounds. In practice, the best estimate $a_{90/95}$ is near the average between the first hit and the virtual miss and the 95% confidence bound is near the first hit.

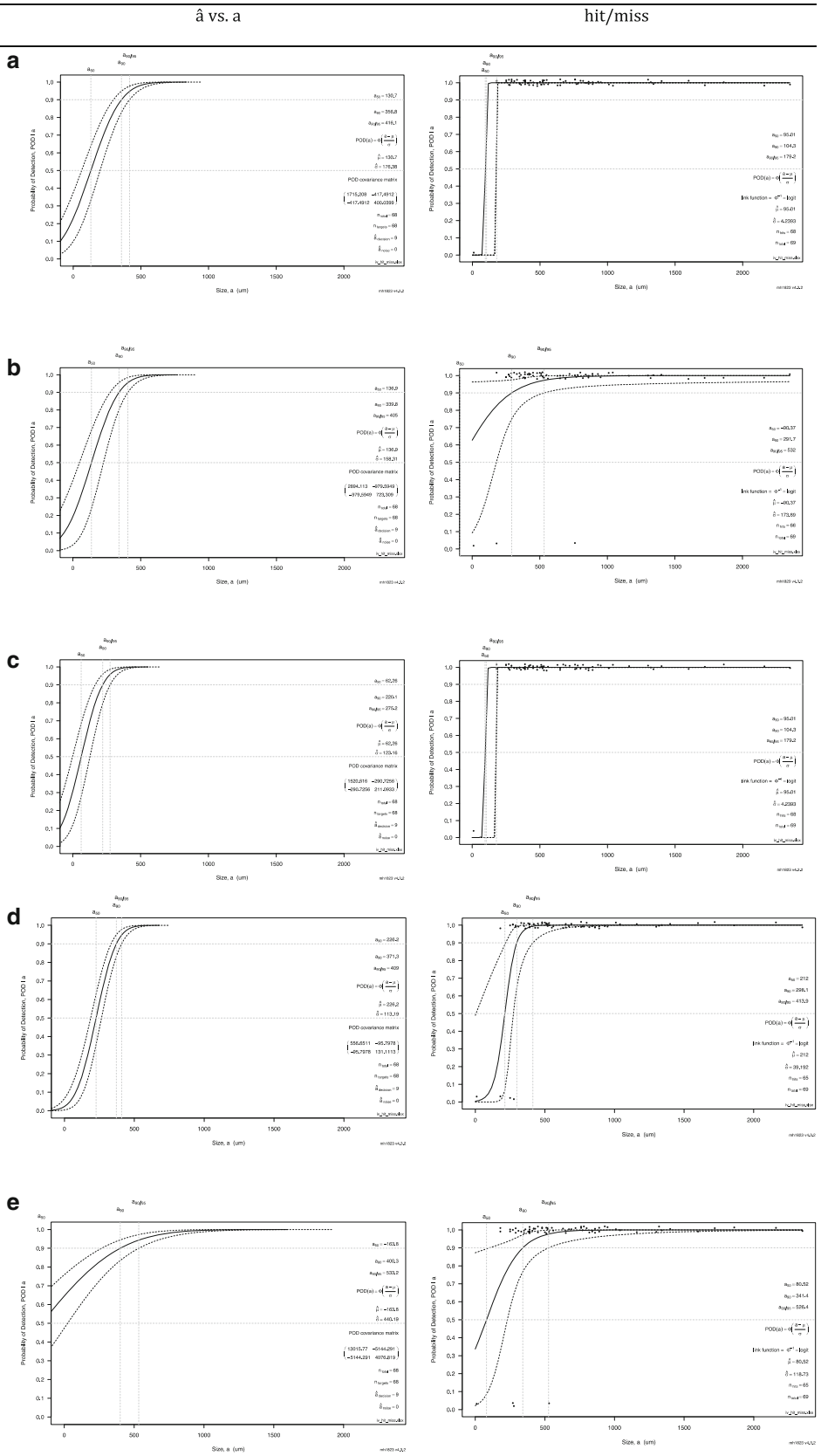
For inspector B, the data contained an outlying miss of a large crack (760 μm) much greater than the otherwise estimated $a_{90/95}$. This significantly increased the computed $a_{90/95}$ values and yet gave $a_{90/95}$ estimates lower than the largest missed cracks. The $a_{90/95}$ value for the â versus a analysis was not affected as much.

The reported hit/miss data and corresponding â values were studied to see if the inspectors followed a consistent â threshold in their hit/miss assessment. All of the inspectors reported hits for lower â values than the highest â value reported for no-flaw and thus none of the inspectors based their hit/miss judgement on the â alone.

Figure 3 shows the computed hit/miss POD curves and corresponding â versus a POD curves. The linear fits used to obtain the â versus a curves are shown in "Appendix".

Many of the â versus a curves show artificially high POD at zero crack length. In reality, cracks very near to size zero cannot be found with the studied methods, and the true POD

Fig. 3 â versus a and hit/miss POD curves for the EC data set



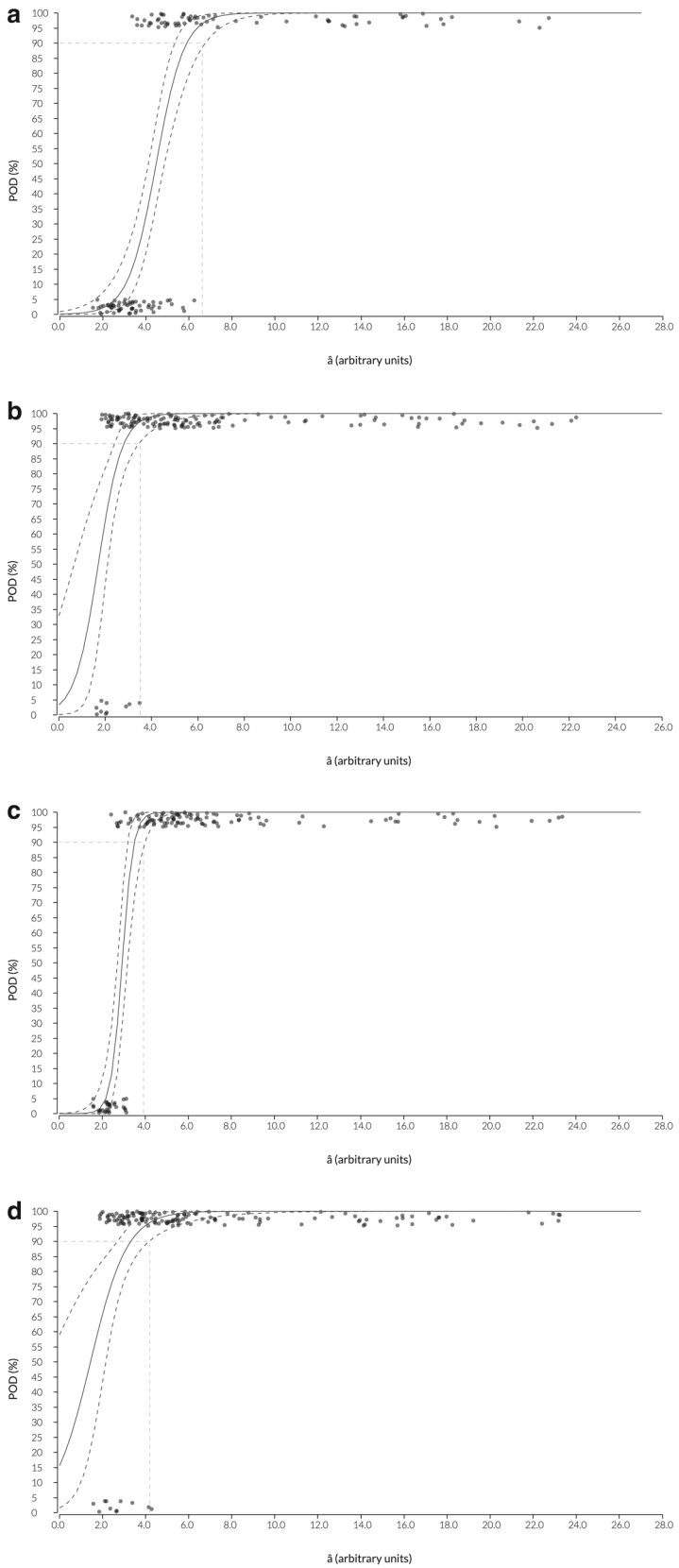
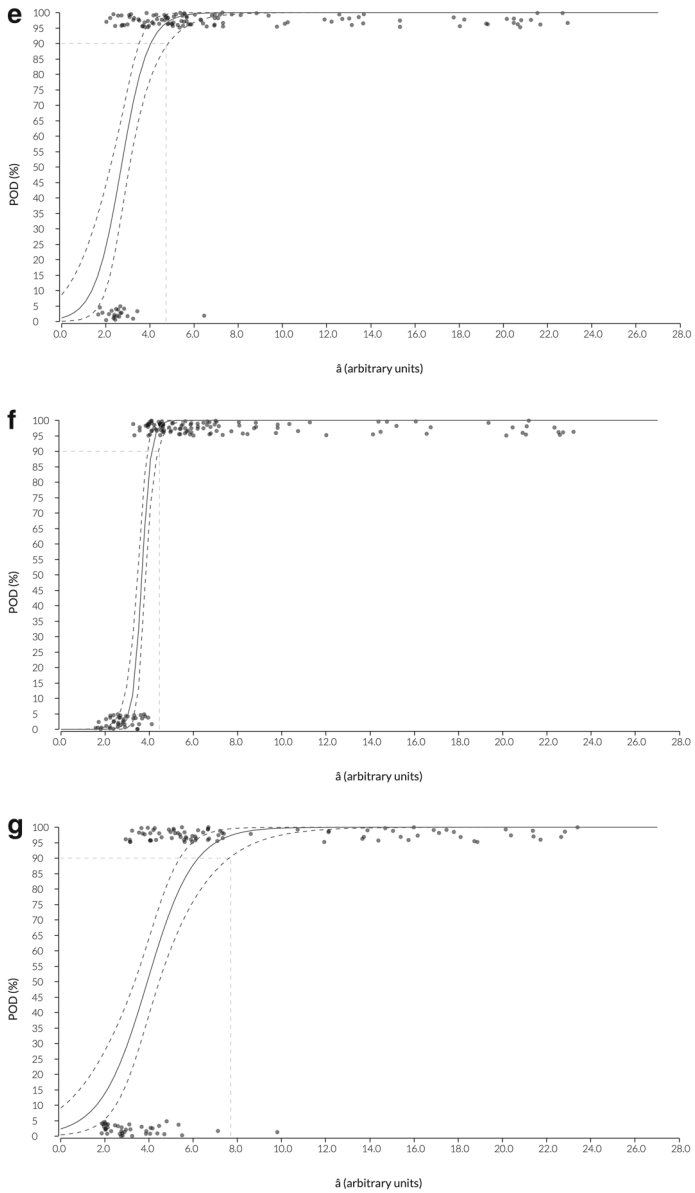


Fig. 4 POD curves for the 7 inspectors as function maximum amplitude (â)

Fig. 4 continued



curve is expected to show zero probability of detection for crack size zero. Thus, some of the â versus a curves show unrealistically high POD at small crack sizes. This discrepancy is related to the reporting uncertainties of the â values, which result in somewhat unrealistic extrapolated â values at zero crack length. The effect is particularly notable for inspector E, where the reported â versus a values showed marked nonlinearity due to reporting discrepancies and consequent unrealistic â versus a fit.

For the hit/miss results, inspectors B and E show similarly unrealistic POD at zero crack length. In the case of hit/miss, this is caused by the insufficient amount of misses in combination with hits at greater crack sizes. The pattern displayed by these inspectors is somewhat inconsistent with the precondition of the used hit/miss methodology, that POD increases with increasing crack size. However, for the hit/miss curves, the inconsistency is also reflected in the rapidly widening confidence bounds.

For the dataset UT, results from 7 inspectors were available. For these data, the hit/miss POD curve was computed similarly to the EC data set. In this case, clear region of "unlikely to find" and "likely to find" equivalent crack sizes were available and there was a clear transition in between. Thus, no additional conditioning was necessary and the hit/miss analyses were completed with the data "as-is".

For this data the number of real cracks was insufficient to establish traditional â versus a POD estimate. For the generated UT-images, a linear â versus a relation was postulated

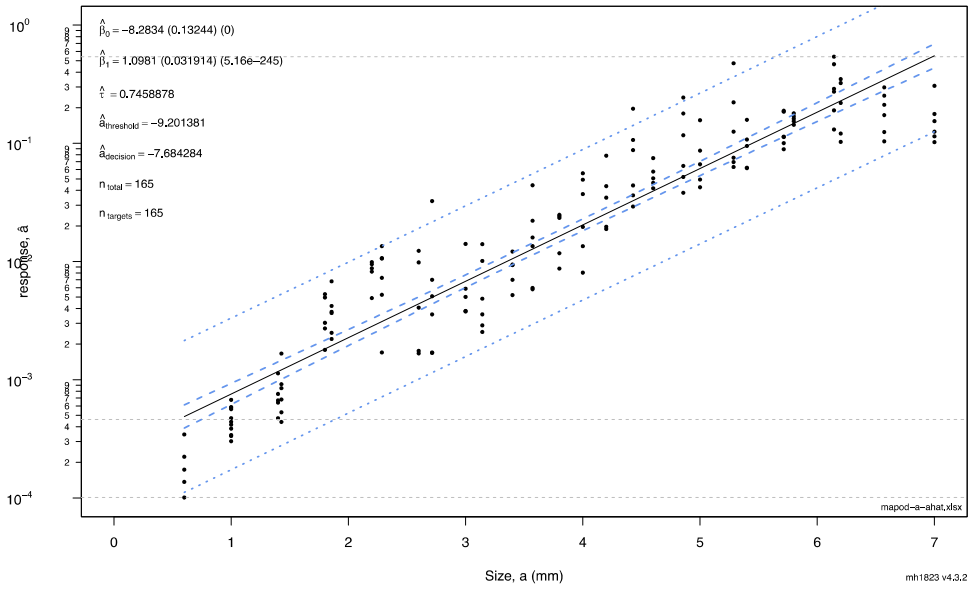


Fig. 5 Model assisted â versus a linear model fit obtained for the same inspection case

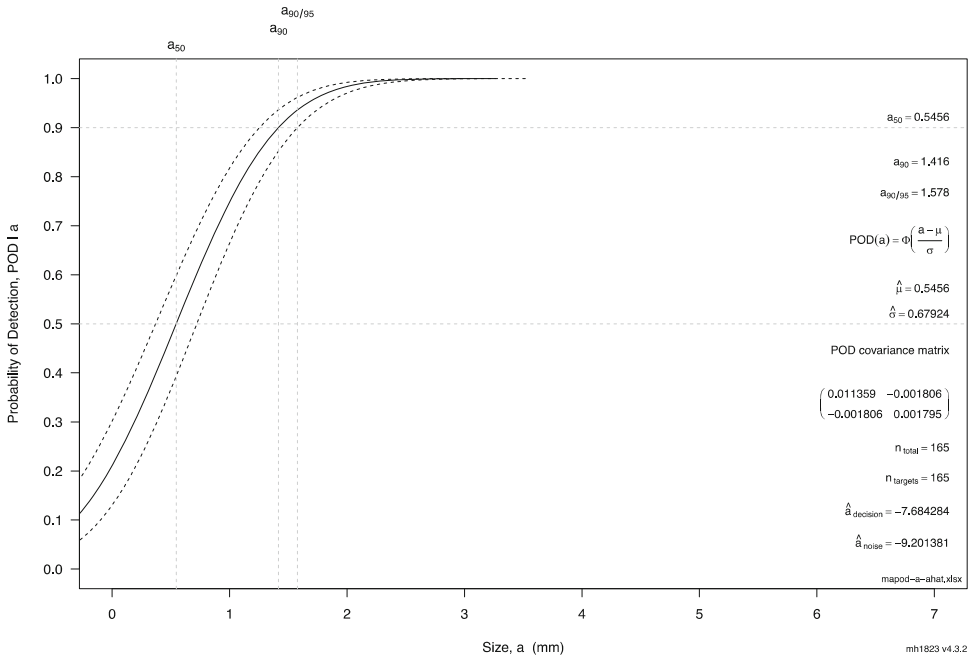


Fig. 6 Model assisted â versus a POD curve obtained for the UT inspection case

for each crack and the images generated accordingly. Consequently, the data gives unique opportunity to study hit/miss in terms of â. The variation in hit/miss judgement of the inspectors as function of â represents the missing variation unaccounted for in the â versus a POD analysis. Figure 4 show hit/miss POD curves as function of â computed from the data.

The POD variation as function of â shows the effect of the inspector judgement that uses features other than the amplitude to assess crack presence (e.g. signal as compared to local variation etc.). It is of interest to know, how this adaptive judgement compares with the simplistic amplitude threshold used in the traditional â versus a analysis. This can be obtained by selecting a threshold slightly above the highest noise peak in the data file. For the present data, this corresponds to amplitude of 4.9. As can be seen from Fig. 4, most of the inspectors show somewhat better performance than would be obtained with the simplified threshold. However, some of the inspectors missed flaws with amplitude significantly above the noise.

To compare, the same UT inspection case was modelled and model-assisted POD curves completed. These are shown

Table 2 Comparison of $a_{90/95}$ values obtained from different sources

Source	a90/95
Inspector hit/miss	
a	3.7
b	1.1
c	1.9
d	1.6
e	2.4
f	2.5
g	3.7
Average	2.4
Simple threshold hit/miss	3.7
CIVA modelled â versus a	1.6

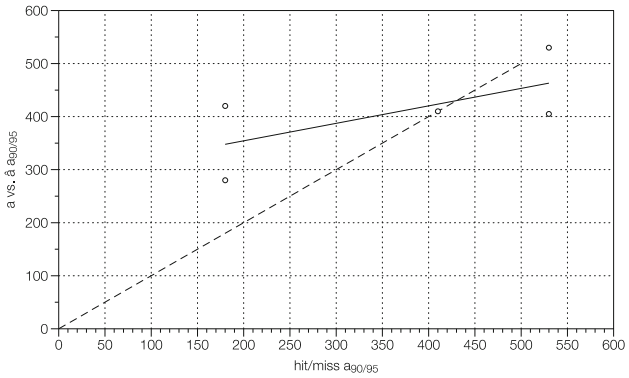


Fig. 7 Comparison of hit/miss and â versus a POD $a_{90/95}$ values for the EC data-set. Dashed line shows the expected line, where results from both methodologies concur. The solid line shows regression line, which shows poor correlation ($R^2 = 0.36$)

in Figs. 5 and 6. The â threshold was set to correspond to average simulated amplitude for 1 mm crack. This is to be compared to the noise amplitude in the measured data at the flaw locations, which were equivalent to expected signal from crack sizes 0.4–1.0 mm.

Finally, with $a_{90/95}$ values computed from three sources, i.e. inspector hit/miss results from amplitude-varied data, simple threshold hit/miss from amplitude-varied data and â versus a for simulated data, the different method can be compared directly. This comparison is shown in Table 2.

4 Discussion

Based on the result of this study, both the â versus a and the hit/miss methodologies present a valid and standardized way to estimate POD curves and the $a_{90/95}$ values. Nevertheless, the results display some discrepancies. The correlation between â versus a and hit/miss for the EC dataset is shown in Fig. 7. There is overall correlation, but also significant variation. For the very small $a_{90/95}$ sizes, where inspectors found all or almost all 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 cited. Conversely, for the larger $a_{90/95}$ values, the â versus a results show smaller $a_{90/95}$ values. This can be attributed to the â versus a methodology failing to account sufficiently to the larger missed cracks in the data. Furthermore, the â versus a is sensitive to variation in the â versus 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 versus a and thus measured performance. On one occasion, the reported â versus a relation showed significant non-linearity and the reliability of the â versus 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 versus a for such cases. In addition, it seems that the â overestimates the $a_{90/95}$ for small crack sizes and underestimates it with larger $a_{90/95}$ values, as compared to the hit/miss.

For the UT dataset, direct observation on the effect of â to inspector judgement was obtainable. Here, the interest is mainly in establishing whether the inspector judgement provides superior results to the simplistic amplitude threshold. As revealed by comparison in Table 2, none of the inspectors performed worse than a simple amplitude threshold and some inspectors showed quite significantly better performance. It is also noteworthy, that even with the small number of inspectors, the variance between inspectors is significant. Thus, the results indicate, that a simple amplitude threshold, properly applied, will underestimate the expected performance. Also, an amplitude threshold will not represent the true variance to be expected from inspections because it fails to capture the inspector variability in hit/miss assessment.

The considerable overlap of hits and misses Fig. 4, as well as variation between inspectors indicates, that inspector judgement may be a significant source of uncertainty even when the variation in signal amplitude is accounted for. This uncertainty would not be addressed by an â versus a analysis even when the variation of measured signal strengths due to inspection conditions were taken into account as, e.g. by Bato et al. [19].

The issues seen in the hit/miss analysis, i.e. convergence problems with data-sets with insufficient misses and unrealistic POD curves for data-sets with unclear separation of misses, primarily stem from the data not fitting the assumption of increasing POD with increasing flaw size. Thus,

using alternate methodology, such as the DOEPOD ([23–25]) would solve these issues, albeit with an increased number of samples required.

The modelled â versus a shows significantly better $a_{90/95}$ than would be obtained with simple amplitude threshold and hit/miss analysis. The $a_{90/95}$ obtained from modelling is highly dependent on the variance provided by the modelled cases and on the chosen threshold amplitude. As noted before, the modelled â response can only incorporate variance from directly modelled flaw characteristics, such as orientation and thus fails to include variation in natural flaw characteristics and/or microstructural changes. This may make the MAPOD values overly optimistic. More importantly, the amplitude threshold in present case was chosen to represent the perceived noise level as is typical for such analysis. Further comparison with the experimental data revealed, that the selection was overly optimistic. Furthermore, the choice of amplitude threshold is critical determinant of the resulting $a_{90/95}$ values and thus, proper choice of threshold is critical to the reliability of the whole assessment. Unfortunately, no guideline can be given for proper threshold selection for the present case, since the inspectors do not follow an amplitude threshold consistently.

5 Conclusions

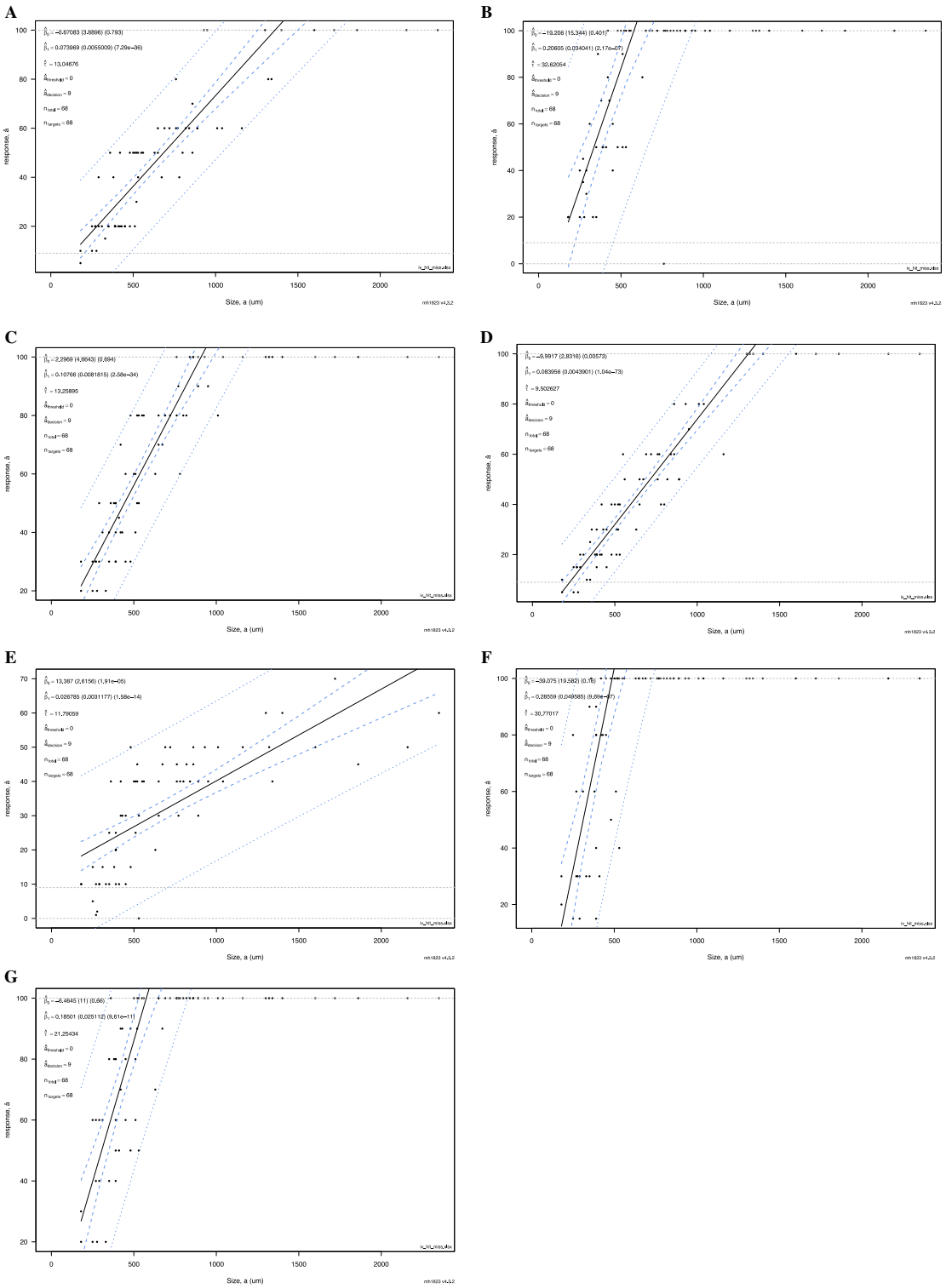
The following conclusions can be drawn from the study:

- the two standard methodologies (â vs. a and hit/miss) may give significantly different results, if true hit/miss decision is to be based on inspector judgement (and not automated signal threshold),
- true inspector hit/miss performance shows significant variance that is not attributable to signal amplitude
- MAPOD, as performed for the present study, is not able to model the inspector performance due to lack of representative amplitude threshold and difficulties in capturing true signal variance.

Consequently, the â versus a approach can only be recommended for inspections, where a consistent signal threshold is enforced, e.g. by an automated system. Similarly, MAPOD can be recommended only where, in addition to the enforced signal threshold, the modelled flaw variance can be well justified. In general, hit/miss approach is seen to be more robust and thus preferable, albeit may also exhibit issues for insufficient data.

Acknowledgements Open access funding provided by Aalto University. The EC data set was arranged and gathered by Patria Aviation (Jouni Pirtola) and The Finnish Defence Forces (Ari Kivistö). Their support is gratefully acknowledged.

Appendix: â Versus a Linear Fit Results for the EC-Data



References

1. Charles Annis, P.E.: Statistical best-practices for building Probability of Detection\n(POD) models. R package mh1823, version 4.3.2 (2016). http://StatisticalEngineering.com/mh1823/
2. Underhill, P.R., Krause, T.W.: Eddy current analysis of mid-bore and corner cracks in bolt holes. NDT&E Int. **44**, 513–518 (2011). https://doi.org/10.1016/j.ndteint.2011.05.007
3. Rummel, W.D.: Nondestructive evaluation—a critical part of structural integrity. Procedia Eng. **86**, 375–383 (2014). https://doi.org/10.1016/j.proeng.2014.11.051
4. Garza, J., Millwater, H.: Sensitivity of the probability of failure to probability of detection curve regions. Int. J. Press. Vessels Pip. **141**, 26–39 (2016). https://doi.org/10.1016/j.ijpvp.2016.03.012
5. Carboni, M., Cantini, S.: A model assisted probability of detection approach for ultrasonic inspection of railway axles. In: 18th World Conference on Nondestructive Testing, 16–20 April 2012 (2012)
6. Carboni, M., Cantini, S.: Advanced ultrasonic "Probability of Detection" curves for designing in-service inspection intervals. Int. J. Fatigue **86**, 77–87 (2016). https://doi.org/10.1016/j.ijfatigue.2015.07.018
7. Gandossi, L., Annis, C.: Probability of Detection Curves: Statistical Best-Practices. ENIQ report nr. 41, vol EUR 24429 EN. European Commission (2010)
8. ASTM: Standard Practice for Probability of Detection Analysis for Hit/Miss Data, vol. E2862-12. West Conshohocken, ASTM International (2012)
9. ASTM: Standard Practice for Probability of Detection Analysis for â Versus a Data, vol. ASTM-E3023. West Conshohocken, ASTM International (2015)
10. Berens, A.: NDE reliability data analysis. In: Lampman, S.R., Zorc, T.B. (eds.) ASM Metals Handbook, 9th edn. ASM, Ohio (1989)
11. Syed Akbar Ali, M., Kumar, A., Rao, P., Tammana, J., Balasubramaniam, K., Rajagopal, P.: Bayesian synthesis for simulation-based generation of probability of detection (PoD) curves. Ultrasonics **84**, 210–222 (2018). https://doi.org/10.1016/j.ultras.2017.11.004
12. Syed Akbar Ali, M.S., Rajagopal, P.: Probability of detection (PoD) curves based on weibull statistics. J. Nondestr. Eval. (2018). https://doi.org/10.1007/s10921-018-0468-2
13. Le Gratiet, L., Iooss, B., Blatman, G., Browne, T., Cordeiro, S., Goursaud, B.: Model assisted probability of detection curves: new statistical tools and progressive methodology. J. Nondestr. Eval. (2017). https://doi.org/10.1007/s10921-016-0387-z
14. Annis, C., Gandossi, L., Martin, O.: Optimal sample size for probability of detection curves. Nucl. Eng. Des. **262**, 98–105 (2013). https://doi.org/10.1016/j.nucengdes.2013.03.059
15. Knopp, J.G., Zeng, L., Aldrin, J.: Considerations for statistical analysis of nondestructive evaluation data: hit/miss analysis. E J. Adv. Maint. **4**(3), 105–115 (2012)
16. Harding, C.A., Hugo, G.R.: Statistical analysis of probability of detection hit/miss data for small data sets. In: AIP Conference Proceedings. AIP, pp 1838–1845 (2003)
17. McGrath, B.: Programme for the assessment of NDT in industry. PANI 3. Serco Assurance, UK (2008)
18. D'Agostino, A., Morrow, S., Franklin, C., Hughes, N.: Review of Human Factors Research in Nondestructive Examination. NRC, Rockville (2017)
19. Bato, M.R., Hor, A., Rautureau, A., Bes, C.: Impact of human and environmental factors on the probability of detection during NDT control by eddy currents. Measurement **133**, 222–232 (2019). https://doi.org/10.1016/j.measurement.2018.10.008
20. Ben Abdessalem, A., Jenson, F., Calmon, P.: Quantifying uncertainty in parameter estimates of ultrasonic inspection system using Bayesian computational framework. Mech. Syst. Signal Process. **109**, 89–110 (2018). https://doi.org/10.1016/j.ymssp.2018.02.037
21. Yusa, N., Chen, W., Hashizume, H.: Demonstration of probability of detection taking consideration of both the length and the depth of a flaw explicitly. NDT E Int. **81**, 1–8 (2016). https://doi.org/10.1016/j.ndteint.2016.03.001
22. Seuaciuc-Osorio, T., Ammirato, F.: Materials Reliability Program: Development of Probability of Detection Curves for Ultrasonic Examination of Dissimilar Metal Welds MRP-262, 3rd edn. EPRI, Charlotte (2017)
23. Generazio, E.R.: Design of experiments for validating probability of detection capability of NDT systems and for qualification of inspectors. Mater. Eval. **67**(6), 730–738 (2009)
24. Generazio, E.R.: Validating design of experiments for determining probability of detection capability for fracture critical applications. Mater. Eval. **69**(12), 1399–1407 (2011)
25. Generazio, E.R.: Directed Design of Experiments for Validating Probability of Detection Capability of NDE Systems (DOEPOD) (2015)
26. Virkkunen, M., Ylitalo, M.: Practical experiences in POD determination for airframe ET inspection. In: 2016/11/3 (2016)