Probability of detection, localization, and sizing: The evolution of reliability metrics in Structural Health Monitoring

Francesco Falcetelli[1,2] ⓘ, **Nan Yue**[2], **Raffaella Di Sante**[1] and **Dimitrios Zarouchas**[2]

Abstract
The successful implementation of Structural Health Monitoring (SHM) systems is confined to the capability of evaluating their performance, reliability, and durability. Although there are many SHM techniques capable of detecting, locating and quantifying damage in several types of structures, their certification process is still limited. Despite the effort of academia and industry in defining methodologies for the performance assessment of such systems in recent years, many challenges remain to be solved. Methodologies used in Non-Destructive Evaluation (NDE) have been taken as a starting point to develop the required metrics for SHM, such as Probability of Detection (POD) curves. However, the transposition of such methodologies to SHM is anything but straightforward because additional factors should be considered. The time dependency of the data, the larger amount of variability sources and the complexity of the structures to be monitored exacerbate/aggravate the existing challenges, suggesting that much work has still to be done in SHM. The article focuses on the current challenges and barriers preventing the development of proper reliability metrics for SHM, analyzing the main differences with respect to POD methodologies for NDE. It was found that the development of POD curves for SHM systems requires a higher level of statistical expertise and their use in the literature is still limited to few studies. Finally, the discussion extends beyond POD curves towards new metrics such as Probability of Localization (POL) and Probability of Sizing (POS) curves, reflecting the diagnosis paradigm of SHM.

Introduction

The concept of *inspection* is fundamentally different from the concept of *monitoring* according to Derriso et al. in three main aspects: the evaluation frequency, the use of previous system outcomes, and the decision range which is possible exploiting the evaluation process results.[1] Therefore, while inspections are conceived to provide a go/no go evaluation related to the health of a structural component, monitoring offers the possibility to take multiple actions thanks to the higher amount of available information.

Farrar and Worden defined SHM as the process of implementing a damage identification strategy for aerospace, civil and mechanical engineering infrastructure.[2] Other definitions are available in the literature,[3–6] all have the common goal of switching from the current scheduled maintenance philosophy to a condition-based maintenance approach.

Condition-based maintenance empowered by SHM can reduce maintenance costs, inspection time[7] and downtime.[4]

The reduced labor requirement of SHM can lead to an increase in safety[4,8] compared to manual inspections, not only for the personnel, but also for the structure itself which may be accidentally damaged during inspections.[9] For difficult-to-reach area, SHM offers the possibility to overcome the accessibility limitations by permanently installed sensors.[10] The military industry sees SHM as an opportunity to increase the combat asset readiness.[11] The examples of other benefits of SHM are the early detection of damage during normal operational conditions and a drastic decrease of the human factor.[7]

[1]Department of Industrial Engineering, University of Bologna, Forlì, Italy
[2]Structural Integrity & Composites Group, Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands

Corresponding author:
Francesco Falcetelli, Department of Industrial Engineering, University of Bologna, Via Fontanelle 40, Forli 47121, Italy.
Email: francesco.falcetelli@unibo.it

In the recent years, the usage of composites has been increasing. However, the complexity of such materials and the presence of a multitude of different possible damage mechanisms still force engineers to use a conservative design approach.[6] The availability of online monitoring data provided by the SHM may enhance the understanding of the new materials and thus leave room to more innovative and closer-to-the-limit design. The reduction of structural design margins can lead to lighter structures. If the structural weight reduction is higher than the additional weight of the monitoring system (sensors, cables, and electronics), lower fuel consumption thus lower CO_2 emissions as well as wider design range are expected.[4]

A Sandia National Labs report written by Roach in 2011 stated that the Technology Readiness Level (TRL) of SHM systems did not go beyond TRL 8 and the majority were concentrated at TRL 4.[12] In 2013 Seaver et al.[13] presented a classification of different sensing technologies based on their TRL. At that time, the TRL was ranging from 3 to 9 depending on the specific application. In recent years, several technologies based on ultrasonic permanently installed sensors such as guided wave monitoring and point thickness measurements became commercially successful.[14] However, there are still barriers preventing a complete transition toward SHM. In 2018 Cawley addressed the main reasons of this unsatisfying rate of transition from NDE towards practical applications of SHM.[14] The lack of specific techniques for performance validation, regarding both damage detection and its corresponding false call rate, was identified as a critical point preventing the widespread

of SHM. The need of performance validation was also outlined in a recent publication of the same author.[15] The MIL-HKBK-1823A[16] allows the assessment of NDE methods exploiting the concept of Probability of Detection (POD) curves. However, there is a lack of specific guidelines and procedures to evaluate the system monitoring capabilities in the field of SHM.[9]

The awareness in the SHM community about the topic of POD curves is still limited. Figure 1 shows the number of publications with the keywords "SHM" and "SHM+POD" since 1995. Despite increasing attention to SHM, only few studies were related to POD curves.

The establishment of common certification criteria is fundamental to the application of SHM technologies,[17–21] and has the potential of improving the design of the system itself.[22] According to Aldrin et al. the qualification of SHM technologies should be based on already present guidelines,[23] such as: cost–benefit analysis (CBA),[24] materials and structure certification, NDE metrics (i.e., POD curves),[16] and procedures for performing a Failure Mode, Effects, and Criticality Analysis (FMECA).[25] In 2011 Aldrin et at. Formulated a protocol,[26] mainly based on the already existing MIL-HKBK-1823A.[16] One year later, this protocol was applied to a real case study, with promising results.[18] Kessler examined three validation standards,[27–29] already utilized in the aeronautical sector, in an attempt of identifying potential relationships with SHM applications.[30]

The scientific question arising from these preliminary considerations is when it is possible to apply standard POD curves for SHM systems. SHM systems can be classified in
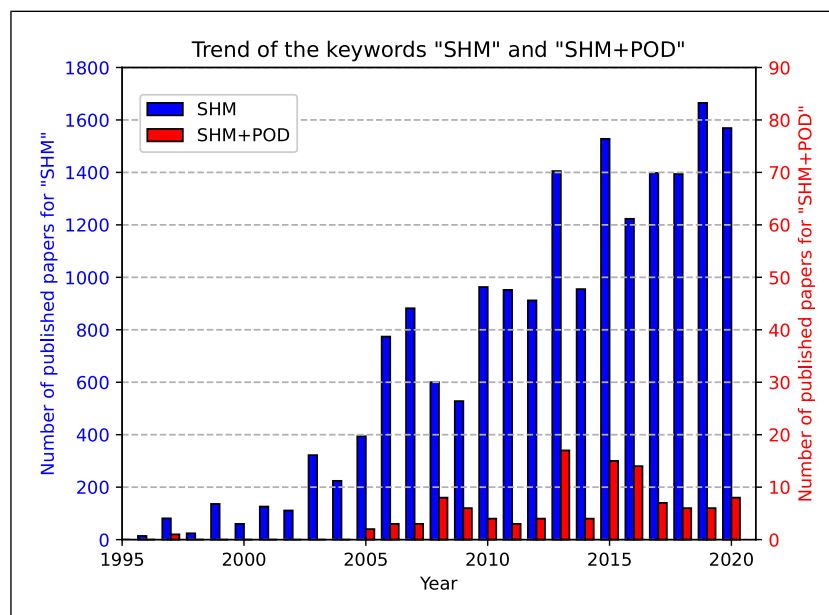

Figure 1. Trend of the keywords "SHM" and "SHM+POD" analyzing the publications selected from Web of Science Core Collection (May 2021).

four categories.[31] First, it is possible to distinguish scheduled SHM (S-SHM) systems, from automatic SHM (A-SHM).[3] In this case the classification is done according to the way sensor data are collected, scheduled time intervals in the former and continuously in the latter. Second, the damage location can be known (KDL) or unknown (UDL), providing another criterion to further classify an SHM system. According to Janapati et al., only the KDL S-SHM could be evaluated using the standard tools of NDE methods such as POD curves.[31] However, the employment of A-SHM has an increasing trend in the SHM community and it is important to being capable of deriving POD curves even for such cases.

Another fundamental aspect is the need of additional metrics to evaluate the reliability of the system also in terms of damage localization and characterization.[18,19,23,26] The Model-Assisted Probabilistic Reliability Assessment (MAPRA) methodology follows this line of reasoning.[19] Kabban and Derriso state that, in the perspective of developing a statistical framework for the certification of SHM systems, the system accuracy and reliability should be assessed with respect to three main points: (i) the capability to determine the presence of the damage (detection problem already common in NDE), (ii) the ability to assess the extent, and (iii) the location of the damage.[22] These additional metrics would find their natural allocation within the paradigm of the SHM phases (detection, localization, assessment, prognosis), initially proposed by Rytter in 1993,[32] and successively chosen as a reference in the field.[33,11,31]

It is interesting to conclude the introduction topic of SHM reliability evaluation with a philosophical question. Could it be necessary to rethink the current regulations and develop new reliability metrics better suited for the advancement of SHM technology? Derriso et at. conceptualized the Cognitive Architecture for State Exploitation (CASE), which resembles the human cognitive behavior and aims to exploit the full potential of a SHM technology, making use of its higher levels (i.e., health management of the full system).[1,34] Despite the CASE approach demonstrated to be more effective in terms of down time costs with respect the Aircraft Structural Integrity Programs (ASIP) philosophy,[35] its full potential cannot be exploited because many of its functionalities should be removed to fulfill the guidelines given in the MIL-HDBK-1823A.

The purpose of this paper is to provide a systematic review of the existing reliability methods in SHM, highlighting the current challenges and areas where further investigation is required. Most of the attempts to quantify the reliability of SHM systems stem from regulations already present in the NDE field. Therefore, it is important to understand the existing guidelines and the basis to transfer the same concept toward SHM.

This paper is organized as follows. *Reliability assessment in non-destructive evaluation* section reviews the statistics behind the POD development in NDE. *Variability sources in structural health monitoring* section examines the variability sources in SHM and their spatial and temporal implications. *Probability of detections for structural health monitoring* section reviews different statistical models to produce POD curves in SHM. *Multivariate-probability of detection* section introduces the concept of Multivariate POD using model assisted methods and metamodels. *Localization and sizing metrics* section discusses a series of localization and sizing metrics used in SHM. *Discussion and perspectives* section summarizes the main findings of the literature review, examining current challenges and areas where further investigation is required. Finally, in Table A1 (see the Appendix at the end of this article) the reader can find the most relevant case studies analyzed in this article.

Reliability assessment in non-destructive evaluation

The detection problem

Table 1 shows the four possible system outcomes for a detection problem.[22,36]

The POD is also often referred to as Positive Predicted Probability (PPP),[36] whereas the PFA is sometimes simply called Probability of False Positive (PFP).[22] In the same manner, the PTN can be named Negative Predicted Probability (NPP).[36] Summing the probability values of each columns in Table 1 always returns the value of one as a direct consequence of set theory.[22,36,37] Exploiting Bayesian conditional probability, it is possible to introduce the concepts of *sensitivity* and *specificity*. Calling P (AD) the probability for the structure to be to be healthy (absence of damage), P(PD) the probability for the structure to be not

Table 1. Possible system outcomes combination in a detection problem.

	Presence of damage	Absence of damage
Detection	True positive (TP) Probability of detection (POD)	False positive (FP) Probability of false alarm (PFA)
No detection	False negative (FN) Probability of false negative (PFN)	True negative (TN) Probability of true negative (PTN)

healthy (presence of damage), P (Det) the probability of the system to report a detection, P(NoDet) the probability of the system to do not report a detection, one has

$$POD = P(Det|PD) = ''sensitivity'' \quad (1)$$

$$PFA = 1 - P(NoDet|AD) = ''1 - specificity'' \quad (2)$$

The POD and PFA are useful in the design phase and to assess the reliability of the measuring system. On the other hand, under operational conditions it can be useful to refer to the other two probabilities: the Positive and Negative Predictive Values (PPV) and (NPV).[22] The engineer can use the PPV and the NPV to determine the conditional probability that a certain damage is present given that it was detected, which is crucial to take the right choice in terms of maintenance

$$PPV = P(PD|Det) = \frac{P(Det|PD) \cdot P(PD)}{P(Det)}$$
$$= \frac{POD \cdot P(PD)}{POD \cdot P(PD) + PFA \cdot P(AD)} \quad (3)$$

$$NPV = P(AD|NoDet) = \frac{P(NoDet|AD) \cdot P(AD)}{P(NoDet)}$$
$$= \frac{(1 - PFA) \cdot P(AD)}{(1 - PFA) \cdot P(AD)) + (1 - POD) \cdot P(PD)} \quad (4)$$

The comparison of equations (1) and (2) with equations (3) and (4) shows that while POD and PFA depends only on the inspection methodology, the PPV, and the NPV depend also on the prevalence. The prevalence is the likelihood of structural damage being present. In a low-prevalence scenario, the PPV can be relatively low even if the POD is high.[22,38]

General considerations on probability of detection curves

In 1989, the studies of Berens[39] were included in the American Society of Metals (ASM) Handbook. Nowadays, the methodology to derive a POD is thoroughly described in the Appendix G of the MIL-HKBK-1823A.[16] The POD is a powerful tool that allows researchers to compare the performance of different monitoring techniques,[40] estimating the sensitivity and reliability of the inspection process.[41] Other definitions are available in the literature.[42–44]

The United States Air Force (USAF), within the ASIP, employs POD curves to assess the reliability of various NDE methods.[45] In the aerospace industry, the POD curve can be exploited to perform risk analyses, to schedule inspections, to estimate the remaining useful life of a certain component, and to develop accept/reject criteria.[46] POD are also becoming attractive for other fields where they were traditionally less popular such as the nuclear industry.[47]

In a POD study it is crucial to have enough data available.[48] A good practice would be to have at least a number of 40 data points.[16,49] Annis et al. investigated what should be the optimum sample size in a POD study and found that beyond 60 samples the improvement on the confidence bounds was less significant.[50] Further details can be found in the paper of Gandossi and Annis.[51] The minimum number of samples also depends on the POD model type. When logistic models are used, at least 60 samples are needed to avoid instabilities.[52] In Koh and Meeker,[53] it is presented a statistical procedure to plan a POD study introducing a dimensionless standardized flaw-size variable. However, since it is often unpractical to produce a statistically significant number of specimens, much effort has been devoted to develop strategies capable to reduce the required data.

â versus a method

The "â vs a" analysis is the most widely accepted method for the POD assessment in the NDE field. The symbol "â" is used to identify the measurement output whereas the "a" parameter denotes the damage size (i.e., crack length) responsible of generating that measurement signal.[16]

From regression to probability of detection curve. Engineers are usually familiar with the Ordinary Least-Squares (OLS) linear regression. However, when dealing with censored data, the OLS would provide non-conservative results. In these situations other techniques such as the Maximum-Likelihood Estimation (MLE) method must be considered. In absence of censored data, the MLE method coincides exactly with the OLS regression analysis.[54] In both cases, the consistency of the models holds if six conditions are verified[52,55].

1. The model must reflect the data.
2. It is required to have a continuous and observable response.
3. The linearity of the parameters must be satisfied.
4. The variance must be homoscedastic (uniform variance) about the regression line.
5. The observations must be uncorrelated (with respect time and/or space).
6. The errors must follow a normal distribution.

In the regression analysis, it is crucial to identify the right â vs a plot. Four possible combinations can be used: (i) â vs a, (ii) â vs log(a), (iii) log(â) vs a, and (iv) log(â) vs log(a).[56]

A good practice would be to plot all the four possible graphs and choose the one with the best fit.[52] Considering the â vs a case (the same procedure can be extended to the remaining three cases), the two variables are related to each other with the following regression equation

$$\hat{a} = \beta_0 + \beta_1 a + \varepsilon \tag{5}$$

where β_0 and β_1 represent the regression coefficients of the model and $\varepsilon \sim N(0, \tau)$ is the corresponding error term, which follows a normal distribution having zero mean and a standard deviation equals τ.[16,40,49,57] Fitting the model represented in equation (5) one gets

$$E(\hat{a}) = \hat{y} = \widehat{\beta_0} + \widehat{\beta_1} a \tag{6}$$

where E is the expectation operator and $\widehat{\beta_0}$, $\widehat{\beta_1}$ are the parameter predictions with their uncertainty (the true parameters β_0, β_1 of equation (1) are unknown). Once the regression model parameters are computed, the POD curve, whose value corresponds to the shaded area in Figure 2, is derived as follows

$$\text{POD}(a) = \Pr\left(\hat{a} > \widehat{a}_{th}\right) = 1 - \Pr\left(\hat{a} < \widehat{a}_{th}\right) = 1 - \Phi_{Norm}(z) \tag{7}$$

Here, $\widehat{a}_{th}$ is an arbitrary threshold value selected by the engineer and z is simply given by

$$z = \frac{\widehat{a}_{th} - \hat{y}}{\tau} = \frac{\widehat{a}_{th} - \left(\widehat{\beta_0} + \widehat{\beta_1} a\right)}{\tau} \tag{8}$$

Another way to express the POD is given by equation (9)

$$\text{POD}(a) = \Phi_{Norm}\left(\frac{a - \widehat{\mu}}{\widehat{\sigma}}\right) \tag{9}$$

where $\widehat{\mu}$ and $\widehat{\sigma}$ are referred as the location and shape parameters, respectively

$$\widehat{\mu} = \frac{\widehat{a}_{th} - \widehat{\beta_0}}{\widehat{\beta_1}}; \quad \widehat{\sigma} = \frac{\widehat{\tau}}{\widehat{\beta_1}} \tag{10}$$

It is possible to identify two points of interest: a_{90} and $a_{90/95}$ They are the estimated crack lengths at which the POD, and its corresponding 95% lower bound, equal 90% respectively (see Figure 3).

Choosing the right threshold. Every POD depends on the detection threshold. Arbitrarily lowering $\widehat{a}_{th}$ would improve the POD curve but inevitably at the cost of increasing the PFA.[46] Therefore, it is possible to draw a comparison between POD curves of different inspection methodologies if and only if their PFA is the same.[49] The PFA can be computed as

$$\text{PFA}(a) = \Pr\left(\widehat{a}_{noise} > \widehat{a}_{th}\right) \tag{11}$$

One could use the so-called Receiver Operating Characteristics (ROC) graph,[58,59] as a tool to determine the best threshold. $\widehat{a}_{th}$ [60,61] The ROC curve is typically used to assess the performance of a classifier and a comprehensive explanation can be found in the papers of Fawcett.[59,62] In the NDE context, the POD depends on both the PFA and the flaw dimensions. Therefore, it is possible to obtain a family of ROC curves selecting a set of different flaw sizes.[63] Nevertheless, there are alternative approaches to select the right decision threshold. As suggested by the MIL-HKBK-1823A,[16] the best trade-off between POD and PFA can be achieved plotting the critical crack target sizes (such as a_{90}, $a_{90/95}$) and the PFA against the decision threshold. $\widehat{a}_{th}$ [16,52]

Probability of detection curve bounds. Confidence intervals express the statistical uncertainty due to the fact that only a limited amount of data is available.[46] The computation of the POD lower bound can be divided into two steps. First, the confidence and the prediction intervals (the latter differs from the former because they also consider the variability of the observations about the predicted mean) are computed using the Wald method. Second, the so-called Delta Method is applied to transfer these confidence intervals to the POD curve. The Delta method can be regarded as a technique for estimating the moments of functions of random variable,[64] and is applied to compute confidence bounds of non-linear functions.[65] Details about the math behind the Wald and Delta methods can be found in the MIL-HDBK-1823A.[16]

With the aid of synthetic data, Figures 2 and 3 show the result of the â vs a method, and the POD curve with its corresponding lower bound using the Delta Method.

Binary (Hit/Miss) data

Historically, the first method to derive a POD was based on the ratio between the number of defects detected, *n*, usually cracks, and the total number of defects inspected in the

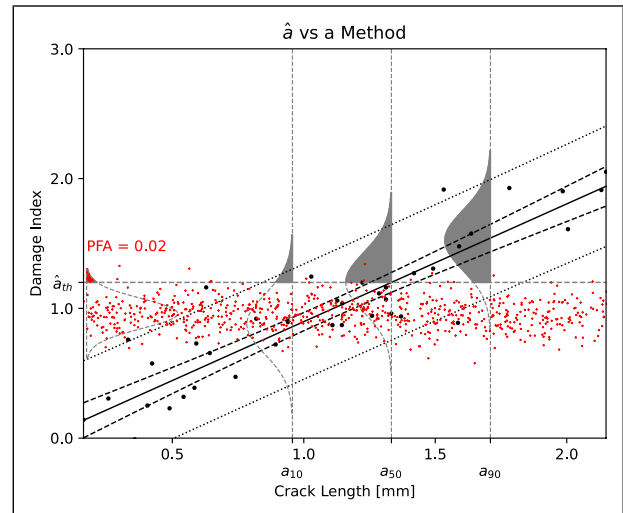**Figure 2.** $\widehat{a}$ vs a linear regression (solid line), its 95% Wald confidence (dashed lines) and prediction intervals (dotted lines). Gaussian noise is represented by red dots. The grey shaded areas of the Gaussian curves represent the POD.
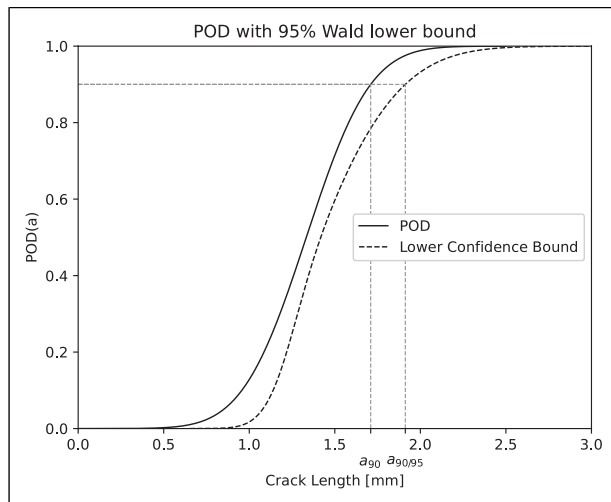
Figure 3. POD curve and its lower 95% Wald confidence bound using the Delta method.

structure, N.[16] Such approach implies an intrinsic tradeoff between the crack length and the POD resolution. Therefore, other methodologies were developed to overcome these statistical deficiencies. When dealing with hit/miss data, the system provides only a qualitative information specifically related to the presence or absence of damage in the structure.[36] The underling statistical models are based on Generalized Linear Models (GLS). The idea is to leverage continuous functions bounded in the interval [0,1], such as the logit, probit, cloglog, and loglog functions,[16] and use the maximum likelihood criterion to compute the model parameters.[66] The interested reader may refer to the following references to delve into POD for hit/miss data and the specific statistical methods to compute the corresponding lower confidence bounds.[16,67–74]

Variability sources in structural health monitoring

In a POD study, it is fundamental to capture all the possible variability sources. Incomplete variability considerations can lead to biased POD curve estimates.[46] Variability is linked to the intrinsic stochastic nature of the phenomenon under examination.[49] Therefore, increasing the amount of data would only shrink the confidence bounds but not the variability. Li et al. proposed the interesting idea of taking into account the inherent population variability using the 0.05 POD quantile estimate and then computed its lower confidence bound to consider also the uncertainty related to the model parameters.[75] Referring to a small quantile of the POD curve can be a more appropriate solution when the engineer is interested to consider the worst possible scenario.[76]

In NDE one possible source of variability comes from the morphology of the cracks itself.[77,78] A portion of variability

could be associated with the sensing device due to the manufacturing process of the instrument. Environmental conditions such as temperature and humidity may change the signal output of a certain measuring technique.[49] Finally, the human factor contribution in NDE systems is often considered the highest within all the variability sources.[79,80]

Structural health monitoring systems inherit all the variability sources of NDE methods. The manufacturing process (sensors, interrogator, specimens, or test structures)[49] and damage morphology are examples of variability sources affecting both NDE and SHM systems.[78] The only exception is the variability associated with the human operator. The SHM system is usually capable of acquiring data automatically even in the areas where typical NDE inspection are unpractical due to complex geometries and accessibility limitations.[10] Nevertheless, stating that SHM systems are not exposed to human variability sources is not completely true. Indeed, the installation of the sensor network on the structure can be regarded as a human related variability source.[79]

Moreover, additional considerations should be considered when performing a POD study in SHM systems. Sensor degradation is one of the main issues. Since the sensor are permanently installed in the structure, they are subjected to degradation over time due to aging and fatigue. Degradation may regard the sensor itself or its coupling with the structure such as the adhesive, welds, and dry couplings, depending on the technique. The system performance can also be affected by changes in the structure itself due to maintenance operations. It has been reported that sensor deterioration over a certain period of time has an impact on the POD curves.[54,81] Environmental and Operational Conditions (EOCs) like temperature, moisture, pressure, and chemical loading, can greatly affect the SHM system response,[82–84] while the same factors are expected to produce minor effects in a NDE system. Other sources of variability can be associated with the loading condition of the structure which can change over time (take-off, cruising, maneuvers, landing, etc.). The reciprocal position between the sensor and the damage is another specific aspect to consider in SHM. When considering POD curves of SHM systems, their relationship with the defect/damage location cannot be neglected because sensor location is a significant source of variability.[31,85] One could produce different PODs depending on the flaw location.[54] Moreover, there could be changes in the recorded signal response due to the on-board SHM device.[49] Mandache et al. suggests that in case of a self-powered sensor, where the recorded data is transmitted with a wireless connection to an on-board memory storage device, electromagnetic interference as well as other possible interference with the avionics are potential sources of variability.[86]

Summarizing, while in NDE variability is mainly attributed to the human factor, in SHM it is equally important

to consider the spatial (location uncertainty)[31,54,85,87] and temporal (environmental effects)[79,80,88] aspects of POD.

Table 2 highlights the main differences between NDE and SHM in terms of variability sources and thus the uncertainties to be considered in the corresponding mathematical models.

The consideration of variability sources in SHM regarding POD curves can be divided into the spatial aspects and the temporal aspects.

Spatial aspects of probability of detection

Optimal sensor placement using probability of detection. The spatial related variability of SHM system has been studied by several authors.[87,89,90] In many cases, the influence of the damage location on the detection performance translates in an Optimal Sensor Placement (OSP) problem.[87] Indeed, the presence of permanently installed sensors suggests that in SHM POD curves may serve not only as a tool to quantify the system performance but also as a tool for the design of the SHM system itself. OSP has been studied for several years and a wealth of literature has been produced. A recent review article written by Tan and Zhang summarizes the main advancement in OSP.[91] An important reference is the study of Flynn and Todd who were among the first to use the concepts of POD and PFA to develop a framework for OSP.[92] Similarly, Azarbayejani et al. demonstrated that OSP can be found maximizing the POD.[93] Markmiller and Chang used the POD as a design constraint for the OSP of an SHM system aiming to monitor the dynamic response of the structure caused by an impact event.[94] Mallardo et al. used POD curves to validate the performance of an artificial neural network whose aim was to detect the location of a certain impact in a composite plate and in a composite

stiffened panel. The OSP was chosen such as the POD curves related to different sensor combinations were maximised.[95,96] Yan et al. used a model-assisted POD approach to validate the performance of different sensor configurations.[97] Chen et al. leveraged POD curves to determine the optimum Lamb wave driving frequency to detect fatigue crack growth in a metallic specimen.[98] Grooteman used POD as an objective function to obtain the OSP for optical fibers applied to a stiffened composite panel.[99] Tabjula et al. used outlier analysis to minimize the number of sensing points in a Guided Lamb Wave (GLW) study and POD were employed to quantify the system performance.[100]

Specimens versus test structures. In NDE, a POD study is carried out testing several specimens. Similarly, in SHM one should employ a certain number of identical structures, which are the equivalent of the specimens in the NDE study. However, this makes an already expensive procedure even more difficult to apply. Identical structures in SHM requires identical sensing systems. Even though this is possible theoretically, Müller et al. showed that this is not feasible in practice.[79] This is the consequence of the amount of variability involved in the manufacturing process and sensors installation. Therefore, the POD curve will apply only for that specific structure and that specific sensing network configuration which was utilized to perform the POD study itself. For these reasons, the procedure becomes tremendously costly and time-consuming. Liu and Chang, in a US patent assigned to Acellent Technologies, propose to mimic the damage by bonding stiff metal or damping patches in the structure to create a POD database for a large structure.[101] However, the authors also stated that using real damages to produce these POD curves may lead to more accurate

Table 2. Variability Sources: NDE versus SHM.

	Variability Sources	NDE	SHM
Aging	Sensor degradation	N	Y
	Coupling degradation	N	Y
Damage-related	Damage morphology	YY	YY
	Reciprocal sensor/Flaw location	N	YY
EOCs	Loading conditions (stress/Strain)	N	YY
	Chemical loading	Y	YY
	Humidity	Y	YY
	Pressure	Y	YY
	Temperature	Y	YY
Human factor	Data interpretation	Y	Y
	Measurement procedure	YY	N
	Installation process	N	Y
Manufacturing process	Sensors	Y	Y
	Interrogator	Y	Y
	Specimen/Test structure	Y	Y
Data communication and storage	Electromagnetic and radio frequency interferences	N	Y

N: Variability source not present; Y: Variability source present; YY: Dominant source of variability.

results. The introduction of real damages implies that the damaged structure might not be reusable, which increases the time and cost.

Decision threshold for structural health monitoring systems. Sometimes the damaged data may not be available or if available it might not be statistically relevant. In such cases, the threshold can be chosen by exploiting algorithms developed for unsupervised learning problems.[11] A multitude of methods are available in the literature within the field of *novelty detection*. Some methodologies require that the feature vector is normally distributed, such as *outlier analysis*.[102] On the other hand, other approaches such as *extreme value statistics*[103] can be used without the normal assumption to determine the best threshold value. A prominent reference is the review of Markou and Singh, which summarized the main statistical[104] and neural network based[105] approaches. Due to these additional challenges in determining the proper threshold, Cobb et al. suggests to use an hit/miss approach when dealing with SHM system.[106] Monaco et al. propose a methodology to evaluate the threshold level on a SHM GLW study based on the statistical analysis of noise.[36] In their study, the Kolmogorov–Smirnov test is used to reject the null hypothesis, being the non-Gaussian distribution of the experimental data. The same approach for the selection of the damage threshold has been used by Memmolo et al. in a study concerning the damage detection in a composite plate using a tomography technique based on GLW.[107] In Yue et al.[108] the detection of multiple barely visible impact damage (BVID) in large composite aircraft panels is achieved by outlier analysis using a reference pristine database gathered from simple coupons and mono-stringer panels under a wide range of temperature variation.

Temporal aspects of probability of detection

It is often stated that the major difference between NDE and SHM regarding POD development is that in the first case subsequent inspections are independent whereas in the second case they are correlated. Such statement is not entirely true. It would be more accurate to state that the degree of statistical independence between subsequent measurements is greater in NDE than in SHM. Statistical independence is a property that holds only for random events.

For example, Forsyth showed that the hypothesis of statistical independence is not completely true for repeated inspection related to simulated penetrants and eddy current testing.[41] This does not mean that one should jettison all the theory developed in MIL-HKBK-1823A, which assumes independent inspections. Even though several statistical independent degrees can be present, the hypothesis of independence might be *true enough* to lead to consistent results.

However, this hypothesis is not valid for SHM systems, where a continuous stream of data is expected to be recorded from the structure. The degree of correlation between measurements separated by a small-time interval cannot be ignored and must be properly handled. This intrinsic dependency of SHM measurements hinders the application of traditional statistical methods to produce POD curves, which is considered the most significant barrier preventing the widespread of the SHM technology.[34] In 2008 Shook et al. recognized this problem and developed a mathematical model to derive POD curves in the presence of repeated dependent data.[109] Discarding specific chunks of information might restore data independence but at the expense of compromising the effectiveness of the SHM methodology itself .[57] For this reason, several studies have been conducted to determine whether it is feasible to generalize the assessment methodology of the POD metric to SHM systems.[49]

In the following section Sequential Data Analysis is introduced, showing how it can allow to deal with slowly evolving spurious signal changes due to EOCs, defect morphology, sensors drift and other kind of variability sources.

Sequential data analysis. As reported in Table 2, EOCs are predominant sources of variability in SHM. They can affect the detection performance of the system and their effect must be considered. In NDE it is possible to make measurements for a certain damage at varying EOCs. However, it is not possible to do the same in SHM. The output of the SHM detection system depends on the whole history of the EOCs. Moreover, this is coupled with the damage evolution, which leads to the need of studying a tremendous number of structures.

For instance, in ultrasonics SHM studies temperature has been reported to be the predominant effect in EOCs.[84,110–112] There are two main methodologies for temperature compensation: the Optimal Baseline Selection (OBS) and the Baseline Signal Stretch (BSS). A good description of the OBS method can be found in the paper of Lu and Michaels[113] whereas the BSS methodology is applied in several references such as Croxford et al.,[114] Michaels,[115] Clarke et al.,[112] Harley and Moura.[116] Recently, data-driven methods have been developed for effective temperature compensation of large temperature variation up to 70°C,[117] and for anisotropic materials.[118]

Liu et al. proposed a hybrid approach to tackle the problem of slowly evolving spurious signal changes due to EOCs.[88] The authors evaluated the signal response of a pipe monitoring system under varying EOCs for the undamaged structure. Then, the damage effect was synthetically superimposed to the undamaged signal. A BSS algorithm was used for temperature compensation, and the baseline subtraction, Singular Value Decomposition (SVD) and Independent Component Analysis (ICA) damage feature extraction methods

were compared. The ICA approach proved to the in general the most efficient to produce reliable ROC curves.

In a recent article, Mariani and Cawley summarize other temperature compensation techniques developed in the last decade.[119] Among them, the location-specific temperature compensation (LSTC) showed promising results for torsional guided wave signals in pipe monitoring,[120,121] resulting in a patent.[122] The same authors proposed a change detection algorithm based on the Generalized Likelihood Ratio (GLR)[123] applied to data obtained through the LSTC or the OBS methods. Their method proved to be sensitive to departures from the pristine state of the structure. However, the methodology only applies if the no sensor drift is present, which is one of the underlining assumptions of the change detection scheme. As a matter of fact, sensor aging and degradation remain an open challenge in the SHM field. Mariani et al. proposed a new methodology to address sensor drift on a thick copper block specimen.[124] They exploited the back wall echo ratio to reduce influence of PZT sensors drift. Another recent article of Mariani et al. leveraged causal dilated convolutional neural networks to both compensate EOCs and sensor drift.[125] Their algorithm, which is an adaptation of WaveNet (a deep neural network for audio waveforms),[126] outperformed the OBS and BSS approaches.

Probability of detection for structural health monitoring

In this section, three POD methods developed for SHM are presented, the Length at Detection (LaD) method, the Linear Mixed-effect Model (LMM) and the Random Effects Model (REM). These methods do not aim to address time dependent data in the way that a sequential analysis does. However, they provide different frameworks to handle the statistical dependence of the measurements being collected.

The length at detection method

The LaD model solves the dependency of sensor data, which is characteristic of SHM systems, by taking the measurement when the crack/damage is detected the first time. Therefore, there are not repeated measures because only the first crack detection is considered.[49] Roach et al. in 2007,[127] and Roach in 2009,[128] studied the reliability of Comparative Vacuum Monitoring (CVM) in aluminum and steel specimens using the LaD method. Sbarufatti et al. in 2016[129] and Sbarufatti and Giglio in 2017[130] applied the LaD methodology for fatigue crack monitoring on the tail boom stringers and fuselage panels of a helicopter using Fiber Bragg Grating (FBG) sensors within a model assisted framework. The crack size recorded in each test corresponds to the one for which a clear and stable detection signal is produced. Therefore, it all boils down to the task of characterizing the probability distribution of the crack lengths at detection and its cumulative distribution function represent the corresponding POD curve.[131] Assuming that the crack population shows a Gaussian distribution

$$POD(a) = \Pr(X < a) = \Phi_{Norm}\left(\frac{a - \bar{x}}{s}\right) \qquad (12)$$

Similarly, if the crack population has a lognormal distribution, the crack length a in equation (12) is replaced by $\ln a$.

The variables $\bar{x}$ and s represent the sample mean and standard deviation, respectively. The assumption of a normal or lognormal distribution of the cracks at detection is not always easy to verify and therefore can be considered a limiting factor for this approach. One possibility is to use the so-called Anderson–Darling test.[132] The assumption of normal or lognormal distributed crack lengths at detection can be rejected if the p-value provided by the test is lower than 0.05, which represents the chosen significance level.[131] Probability plots are another useful tool to test above-mentioned assumptions. In this case, data (crack lengths at detection) are plotted against the theoretical normal (or lognormal) distribution. If the data lie approximately in a straight line, then it is possible to state that the population follows that probability distribution. The LaD methodology holds even for other kind of statistical distributions such as the smallest extreme value and Weibull, and the largest extreme value and Fréchet distributions.[49] Further information about these distributions are available in the Appendix C of reference.[133] Figure 4 uses synthetic data to simulate and visualize the working principle of the LaD method.

The confidence bound can be computed exploiting statistical methods relying on the non-central t distribution,[49,133] or applying the One-Sided Tolerance Interval (OSTI) approach. This methodology was firstly proposed by Roach in the detection of fatigue cracks using CVM. It provides an estimation of the upper bound containing a certain fraction of all measurements in the population with a given confidence level.[9] The percentage of all measurements and the confidence level are the main factors affecting the result. The former is usually taken equal to 90% whereas the standard for the degree of confidence is 95%. The OSTI approach can provide a reliable analysis with only eight flaws with respect the 51 required in a classic binary data POD.[134] Using the same symbols found in Roach,[128] the upper bound for the tolerance interval is given by

$$T = \bar{x} + K_{n,\gamma,\alpha} \cdot s \qquad (13)$$

where T represents the tolerance interval, $\bar{x}$ denotes the mean of the detection lengths, s is the standard deviation of the detection lengths, and K is the probability factor which
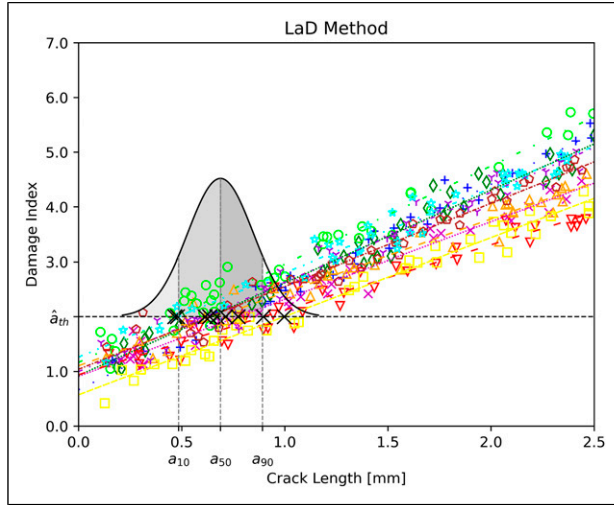
Figure 4. LaD method using 10 specimens (shown by different markers). Each specimen has 40 measurements.

takes into account three parameters.[135] The first parameter is the sample size n, the second is the confidence level γ, and the third one is the detection level α. The K value can be found in specific tables available in several statistic books, see for example, the work of Krishnamoorthy and Mathew[136] or Meeker, Hahn, and Escobar.[133] The probability factor decreases as the sample size increases, which is consistent with the fact that limited number of measurements is associated with a higher level of uncertainty of the sample mean and variance.[9] Increasing the desired level of confidence leads to higher K values, which is reasonable. Finally, the higher is the detection level and the higher is K because an increase in the detection level must correspond to higher crack lengths.

Linear mixed-effect model

Kabban et al. developed a statistical model in 2015 to produce POD curve for time-dependent data extending the classical â vs a methodology.[57] Since the observations are no longer uncorrelated, it is not possible to use OLS or MLE.[55] One possibility is to rely on generalized least square models (originally developed by Aitken in 1936[137]), which are capable to handle such time dependency .[57] The second possibility is to use a Linear Mixed-effect Model (LMM). This kind of approach extends classical linear models and are particularly suited for datasets where data are not truly independent. The LMM acronym comes from the fact that the "model" to be fitted is "linear," and that there is the presence of a "mixed effect": a random effect (this could be the intercept or the slope) and fixed one which generally describes the expected trend of the data. The synthetic data which were considered in the study of Kabban et al. suggested to apply a LMM with a random intercept.[57]

Therefore, using the same terminology employed by the authors, there is a random intercept term for each experimental unit (EU), which is regarded as the basic primary experimental item used to collect data. Equation (14) explains how this approach translates in mathematical terms

$$\hat{a}_{ij} = \beta_0 + \beta_{0i} + \beta_1 a_{ij} + \varepsilon_{ij}; \ i = 1,\ldots,n; \ j = 1,\ldots,m \quad (14)$$

The term $\hat{a}_{ij}$ denotes the jth measurement taken from the ith EU. The two fixed regression coefficients are represented by β_0 and β_1, while a_{ij} it is the actual crack dimension. The random intercept term is expressed by $\beta_{0i} \sim N(0,\omega^2)$, which follows a normal distribution with zero mean and ω^2 variance. Finally, similarly to the conventional â vs a methodology, $\varepsilon_{ij} \sim N(0,\tau^2)$ is the error term. From a statistical point of view this makes a relevant difference with respect to the classical model presented in equation (5). Indeed, in the case of equation (14), the variance of the response depends both on the error and the random effect variances. Hence, this model assumes that there is a correlation in the measurement made within the same EU and that there is independence between measurements taken from different EUs. The parameters estimates can be derived from the marginal model which averages all the random effects to return an average response expected value. Such methodology allows to incorporate data correlation into the variance of the marginal model error terms.

Random effects model

This model, sometimes referred to as Repeated Measures Random Effects Model (REM²),[138] tries to generalize the â vs a analysis described in MIL-HKBK-1823A for SHM systems.[49] Moreover, it takes a step forward with respect to the LMM developed by Kabban et al.,[57] in the sense that it considers the possibility to have at the same time a random intercept and a random slope. Every crack-sensor combination will produce a series of data which can be fitted using a line with its own slope and intercept. Therefore, the method computes the joint distribution of these parameters[49]

$$\hat{a}_{ij} = \beta_{0i} + \beta_{1i}\left(a - \overline{a}\right) + \varepsilon_{ij}; \ i = 1,\ldots,n; \ j = 1,\ldots,m \quad (15)$$

The subscript "i" denotes a certain crack-sensor combination, and "j" indicates a specific reading coming from that crack-sensor pair. Therefore, $\hat{a}_{ij}$ represents the jth measurement response (for example, it could be a scalar value representing a certain damage index) of the ith crack-sensor combination. The regression coefficient β_0 and β_1 utilized in the conventional â vs a analysis are substituted by β_{0i} and β_{1i}, highlighting once again that they are unique for every sensor-damage pair. Analogously, the error term ε, becomes ε_{ij}, which also depends on the specific sensor-damage reading. Moreover, the slope coefficient is not

directly multiplied by the flaw length, but it is multiplied by the difference between the crack length and the sample mean of the crack lengths, $\overline{a}$, related to the entire dataset. The POD formula is given by equation (16)

$$POD(a) = \Pr(a > a_{th}) = 1 - \Phi_{Norm}(z) \qquad (16)$$

Where the z variable is described by the following equation

$$z = \frac{a_{th} - \left[\mu_{\beta_0} + \mu_{\beta_1}\left(a - \overline{a}\right)\right]}{\left[\sigma_{\beta_0}^2 + \left(a - \overline{a}\right)^2 \sigma_{\beta_1}^2 + 2\left(a - \overline{a}\right)\sigma_{\beta_0}\sigma_{\beta_1}\rho + \sigma_{\varepsilon}^2\right]^{\frac{1}{2}}} \qquad (17)$$

At the numerator a_{th} is the detection threshold value in the response, while μ_{β_0} and μ_{β_1} are the mean of the intercepts (referred at the crack size equal to $\overline{a}$ for all the crack/sensor pairs in the dataset) and of the slopes, respectively. At the denominator instead of simply having the standard deviation τ, we an expression containing several terms. Specifically, σ_{β_0}, σ_{β_1}, σ_{ε} represent the standard deviations of the intercepts (referred at the crack size equal to $\overline{a}$), of the slopes, and of the error term (for every flaw/sensor combination), respectively. Finally, ρ is the measure of the correlation existing between the slopes and the intercepts. Figure 5 shows an example of such method with the aid of synthetic data.

The computation of the corresponding lower bounds can be achieved with a MLE approach but Bayesian methods with weekly informative priors are also an option. The interested reader can delve into the work of Meeker et al. for further details about these procedures.[49]

Comparison between length at detection and random effects model methods

In this section are discussed the main differences between the LaD and the REM methods. Since the REM method is a generalization of the LMM approach, the latter is not considered in this analysis.

Both LaD and REM model are valid statistical tools to evaluate POD curves for SHM applications. The LaD approach has been particularly used to evaluate the CVM performance in aerospace structures. Thanks to the Federal Aviation Administration (FAA) research program in SHM started in 2011, recently in the U.S. the use of this approach started to be accepted from the major original equipment manufacturer (OEM) and airline operators such as Boeing and Delta.[139] Despite the LaD offers a relatively simple approach, it discards some information, thus it does not exploit the full potential of the specific SHM application. It also requires an assumption about the crack length at detection distributions (not always easy to verify) and different distribution choices can lead to significantly different results. On the other hand, the REM uses the whole dataset,
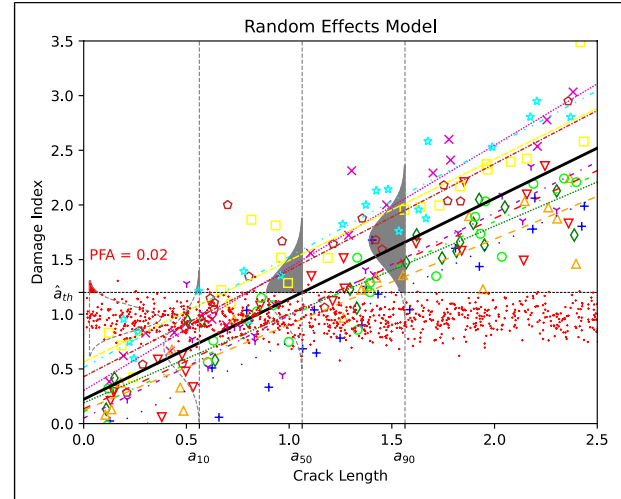

Figure 5. REM method using 10 specimens, each one with 20 measurements. The red dots indicate noise measurements.

and this also implies the model to be more robust against departures from the model assumptions. More important, it can be compatible with a model-assisted approach, which makes it very attractive for future applications. A study made by O'Connor tried to quantify the difference between these two statistical methods.[140] From a qualitative point of view, it is pointed out again the fact that it is difficult to justify the use of a certain distribution (normal or lognormal) in the LaD method. Nevertheless, when few observations are available (less than 10) the LaD method seems to be more appropriate since it may be not possible to fit a 5 parameter REM. Therefore, since the LaD it is also lighter from a computational point of view it might be preferable in certain engineering applications. In reference[140] the two methods are also compared quantitatively using as reference the a_{90} values computed for different datasets. The LaD seemed to overestimate the a_{90} when the normal distribution approximation of the crossing lengths was not appropriate. However, this results in a conservative prediction, which can be acceptable from an engineering point of view. The two methods showed comparable results except for situations where the σ_{β_1} value was high. The LaD and REM models were derived for single parameter describing the damage, which is in many cases the crack length. However, in real applications more quantities could affect the signal response and hence a vector rather than a scalar value should be considered. In this case, a more complex formulations for these models should be developed.

Previous SHM studies tended to neglect data dependency but now the literature highlights that this is not the right way to proceed. Despite these statistical methods are relatively new in the SHM field, there are already several case studies where they are leveraged. Recently Kessler et al. made use of both LaD and REM approaches to develop

POD curves.[138] The authors used a 4-point bend test to obtain a crack growth starting from an Electrical Discharge Machined (EDM) notch on Aluminum bars. They monitored the crack evolution with a carbon nanotube (CNT) sensor, which have a great potential for aerospace applications.[141] A recent study, using the methodology proposed by Meeker,[49] made use of a Bayesian approach to derive POD curves for different case studies.[142] Typically, in the majority of the SHM systems based on ultrasound techniques, the relationship between the damage index and the damage size is not linear, which goes against the assumption made by Meeker.[49] However, linearity can be restored applying a logit transformation to the damage index.

Multivariate-probability of detection

One single parameter may be not sufficient to describe the defect characteristics satisfactorily. In evaluating the corrosion present in aircraft structures, Bode et al. derived POD curves as a function of both defect size and percent corrosion.[143] Lee et al. created a M-POD surface using a multivariate log-logistic regression model based on hit/miss detection, where both the defect length and depth are considered as parameters in a ECT application.[144] In a similar ECT study, Hoppe developed a M-POD as a function of crack length, l, and depth, d, but this time extending the classical â vs a method.[145] In 2012 Aldrin et al., along the lines of the previous work of Hoppe,[145] found that including the crack depth in addition to the crack length was reducing the model uncertainty of about 20%.[146] The same authors leveraged a physic-based model (VIC-3D©) to consider several parameters and thus to reduce the variability and the required number of samples. Another case study regarding the ECT of fastener sites for fatigue cracks,[147] revealed that the calibrated physic-based model, which considered multiple parameters rather than the simple crack length, performed better with respect the classic â vs a method.[148] Pavlović et al.[149] developed a M-POD for an ultrasonic inspection of a cast iron component. With this approach it was possible to compute several POD curves as a function of the desired variable holding the other parameter values. Yusa and Knopp pointed out that the M-POD in Pavlović et al.[149] was based on 12 coefficients which were not easy to compute, and that it is unlikely to have a uniform variance.[150] Therefore, they proposed a multi-parameter approach where even the variance is a function of the parameters instead of being constant. Alternatively, in Gao et al. can be found a linear mixed effect model describing the response of a vibrotermography test as a function of the vibration amplitude, pulse length, trigger force and crack length.[151]

Only recently, M-POD related to SHM systems with permanently mounted sensors have been studied. M-POD models are particularly attractive for SHM applications systems because they provide frameworks to include the extra variability sources typical of SHM systems. To bring this approach into the SHM world, the aid of numerical simulations seems to be unavoidable. For SHM, M-POD often requires the use of a model assisted approach.

Model-assisted probability of detection

Model-Assisted Probability of Detection (MAPOD) curves have their roots in NDE but at the same time provide a framework suitable for SHM studies.[49] An extensive review of MAPOD studies can be found in the Pacific Northwest report written by Meyer et al. in 2014.[152] This research field was pioneered by Thompson, who guided from 2003 to 2010 the Model-Assisted Probability of Detection (MA-POD) Working Group at Iowa State University.[153] One objective of a MAPOD study is to reduce the amount of experimental data, required to generate a reliable POD, gathering additional information through a physics based model.[154] There are two main MAPOD variants: the transfer function approach (XFN) and full model-assisted methodology (FMA).[46,154]

The XFN exploits the relationship existing between the output signal of real flaws and synthetic produced flaws which are easier and less costly to realize.[155] Using the XFN approach, starting from an existing fully empirical POD curve for a certain technique, it is possible to transfer these results to another similar configuration. The underling transfer function may be computed exploiting a physic based model or by specific laboratory tests.[156]

The FMA approach aims to predict the signal strength of a certain NDE/SHM technique as a function of several parameters and flaw properties, capturing all the variability sources, combining the information provided by physics-based models with empirical knowledge[156] such as experimental noise.[157] The first attempts were made on ultrasonic testing methods but the FMA concept is general and can be applied to other sensing techniques.[46]

Thompson concluded that the XFN and the FMA approaches were just two sides of the same coin.[156] In 2008 a unified approach for MAPOD was proposed in the form of a protocol,[154] and later on the methodology was included in the MIL-HKBK-1823A.[16]

Gianneo et al., taking the work of Pavlović et al.[149] as reference, leveraged MAPOD in a study concerning GLW for SHM systems with lightweight material.[83,85] From the M-POD curve (called in the paper "master" POD) the authors derived several conventional POD curves as a function of single parameters like the flaw size, the angle with respect the PZT sensors and the Lamb wave mode (A_0 or S_0). On the other hand, the remaining parameters were treated as random variables. When numerical models are employed, their success in capturing all the variability sources strongly depends on the *known unknowns*.[158]

Previous knowledge about the important variables is indeed important to obtain variability data sources from experiments and to integrate them into the numerical models using noise signals as an example. For instance, Memmolo et al. decided to use a MAPOD approach for a GLW based SHM technique.[159] The variability sources were considered by adding a random noise to the FEM model output and randomly choosing parameters that are related to the damage such as its morphology and its position in the structure. Tschoke et al. studied the feasibility of MAPOD to produce POD maps in an automotive component made of Carbon Fiber Reinforced Polymers (CFRP) obtaining promising results.[90] This can be considered a M-POD since the additional parameter of the damage location is considered. Similarly, Leung and Corcoran evaluated the POD spatial distribution and combined this information with the probability of defect location.[87]

Metamodels

In general, the computational effort increases with the number of variability-related parameters considered in the M-POD. High dimensionality problems require many model evaluations that need to be run, which can be a prohibitive process. A possible solution to lessen the computational burden is to leverage metamodels. Metamodels, sometimes referred to as surrogate models, are basically a simplified model of the original physics based model.[160] For example, CIVA[161,162] is a software that allows to use metamodels for Model Assisted POD studies both in NDE and SHM applications. Moreover, they can be used for other purposes such as sensitivity analysis evaluating the Sobol indices or to derive nonparametric POD curves.[160,163]

Miorelli et al. show that in the CIVA software metamodels can be derived using the Output Space Filling Criterion or the Support Vector Regression algorithm. This kind of solution is particularly important in MAPOD studies. Engineers are often making assumptions about the probability distributions of the parameters related to variability source. These assumptions are difficult to verify, and a huge number of simulations is required to explore all the possible parameter combinations. Relying on conventional physics-based models leads to unfeasible computational times for practical applications. Dominguez et al. developed an algorithm to generate beams of POD curves and derive confidence bounds.[164] They generated a database to develop a surrogate model. The process is computational expensive but if beams of POD must be produced, the procedure becomes soon convenient.

Bayesian methods

Bayesian statistics can be regarded as another useful tool to manage the high amount of requested data.[165] It provides a mathematical framework to take advantage of prior knowledge for inference and decision making.[166] In this case the prior regards the quantity and type of damages present in the structure. Prior belief can then be updated once new experimental evidence is available. One may be wondering why this approach that seems to fit so nicely into this problem has not been considered in the past. The answer is that it can be very time consuming from a computational point of view. Nevertheless, recent progresses in terms of high-speed computing, made possible the easy implementation of Markow Chain and Monte Carlo techniques. These algorithms, combined with the use of physic-based models, allow the derivation of the likelihood required in the famous Bayes' formula. Even in the case of having a poor informative prior or not having it at all, this approach can be applied simply considering a uniform distribution of the prior. In this way the posterior will not be affected by the prior belief and relies entirely on the likelihood. The likelihood could be derived experimentally but also exploiting physics-based models information.[167] Despite Bayesian statistics has already been leveraged in the field of NDE to develop POD curves,[168–171] to the best of the authors knowledge its use in SHM reliability studies is limited. The Bayesian approach has the benefit of exploit the full response of the measuring system in contrast with conventional methods where the only information that is considered is whether or not a certain threshold is exceeded.[172] Therefore, this is a promising field of research for SHM that should be further investigated.

Fusion of probability of detection curves

An SHM system is made of several sensor-damage combinations. Moreover, different types of sensors may coexist in the same structure with the aim of providing complementary information. Therefore, for the same structure several POD curves are expected to be produced, each one related to a certain measuring technique. Ameyaw et al. applied the concept of POD curves to vibration based fault detection and isolation (FDI).[56,173,174] It was found that, depending on the sensor type, position and damage location, different POD curves are generated. Therefore, it was considered reasonable to develop a strategy to combine different POD curves related to different sensors. In this way all the available information is utilized, and the system reliability is expected to improve. Ameyaw et al. proposed a methodology in which, rather than fusing all the POD curves into a single POD curves, several belief values are computed using the Bayesian Combination Rule (BCR).[56,173,174] In this approach, all the possible sensor combinations are considered. For example, a certain damage could be detected (i.e., signal higher than the threshold value) only by certain sensors (each sensor with its own POD curve). By applying the BCR for each

possible detection/missed detection combination, it is possible to derive a corresponding number of belief curve as a function of the damage size.

Nevertheless, it is often not desirable to fuse the curves as it dilutes the available information. It would be more appropriate to apply sensor fusion at lower level and derive a single POD curve using a single damage index combining features extracted from different sensors. Several fusion algorithms exist in the literature[11,175–177] and therefore it is reasonable to think that more than one strategy may be applied.

Localization and sizing metrics

Probability of localization

The second phase of a SHM system aims to localize damage and this section analyses the main progresses made to quantify the localization performance. Localization is only related to unknown damage location (UDL) SHM systems. Known damage location (KDL) SHM systems, sometimes also referred as hot-spot monitoring, do not require any localization as the damage position known. However, must be clarified that hot-spot monitoring is not NDE. Even if they share the fact that damage location is known, they cannot be treated using the conventional NDE methods because all the considerations regarding data correlation discussed in the previous sections. Nevertheless, damage localization remains an essential component of the SHM paradigm because (1) it is not always possible to identify hot spots where damage is likely to occur in the structure, and (2) unexpected events such as impacts, or unknown failure mechanism are always possible. Moreover, SHM systems can produce information beyond mere damage existence. Therefore, additional metrics are required to reach adequate reliability standards.

Aldrin et al. in their study claim that such metric should consist of an error with its corresponding uncertainty related confidence bounds.[23] They gave an example with a potential candidate, the so-called Normalized Localization Accuracy (NLA)

$$NLA = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} \left(\frac{\varepsilon_i^p}{p_i'}\right)^2} \qquad (18)$$

The terms in equation (18) are the number of location estimates, N_p, the error of the ith estimation, ε_i^p, related to the location, p, and a normalizing length factor, p_i'. The resulting NLA is then used to derive the confidence bounds (with a specified level of confidence, for instance 95%) around the damage location estimate. Gagar et al. investigated the location accuracy of Hsu-Nielson and fatigue crack AE sources using broad band piezoelectric sensors. They authors considered the cumulative frequency of error margin as measurand and plotted it against the error to obtain a

probability curve capable to reflect the system performance.[178] In 2011 Flynn et al. proposed a novel damage localization algorithm, in the field of Guided-Wave (GW) propagation, based on the Rayleigh Maximum-Likelihood Estimate (RMLE).[179] The authors recognized the need of a statistical tool to compare the performance of their algorithm against other state of the art methodologies. After clarifying that the peak sharpness around the damage location in an image cannot be considered a reliable metric, two approaches were proposed. The first idea was to produce a density map of the localization probability density function (LPDF). Despite this method provides useful insights about the localization performance, its qualitative nature makes it not suitable to SHM, where quantitative metrics are required to make decisions. The second approach has its foundations in the ROC curve. Similarly, Flynn et al. introduced the localizer operating characteristic (LOC) curve, whose points are a measure of the likelihood of predicting a damage location position inside a certain area around the true location.[179] Therefore, every damage location is expected to show its own LOC and the global algorithm performance is assessed averaging multiple LOCs. Mallardo et al. in 2012 employed a genetic algorithm (GA) to solve an optimization problem regarding the OSP for impact localization in smart composite panels.[96] The evaluation of the fitness function, crucial in every GA, was achieved through the computation of the cumulative probabilistic distribution function (CDF), which is obtained integrating the probability distribution function (PDF). The PDF is the probability density function of locating a certain damage with respect different values of the error distance (distance between the true and computed locations). The CDF turned out to be a reliable metric, capable of being used for the computation of the fitness function.[96] Moriot et al. retrieved the LOC and the CDF methodologies and developed a Probability of Localization (POL) curve,[180,181] defined as the probability of locating the damage inside a tolerance circle of radius ε. With that in mind, equation (20) represents the POL mathematical formulation as a function of ε.

$$POL(\epsilon) = \frac{1}{K} \sum_{j=1}^{K} H\left(\epsilon - AEL_j\right) \qquad (19)$$

where K is the number of experiments and AEL is the absolute error of localization. In other words, it is the Euclidean distance between the computed location and the actual location of the flaw. In equation (21) (x_a, y_a) are the real damage location coordinates whereas $(\widehat{x}_a, \widehat{y}_a)$ are their corresponding estimation

$$AEL = \sqrt{\left(\widehat{x}_a - x_a\right)^2 + \left(\widehat{y}_a - y_a\right)^2} \qquad (20)$$

The symbol H represents the Heaviside-step function. In this way, are counted only the cases where $\epsilon > AEL_j$, which means only the cases where the estimated locations fall inside a circle with radius ϵ and center the point of coordinates $(x_a, \ y_a)$. Figure 6 illustrates the results using a synthetic dataset.

The same authors introduced the concept of Model-Assisted Probability of Localization (MAPOL) as a tool to generate synthetic data and build POL curves. Despite this methodology represents a step forward to derive a reliable localization metric in analogy with common POD curves, it has the drawback of having confidence bounds without any meaning because the POL is not the result of any regression task. This lack of uncertainty evaluation capability makes this approach not appropriate for many applications where decisions are made according to an acceptable risk.

Yue and Aliabadi studied a hierarchical approach for determining the reliability of SHM systems using guided waves.[182] The third level of such methodology regards damage localization and its performance metrics. They proposed to use the concepts of trueness and precision[183,184] to quantify the accuracy of the damage location estimations. The trueness, typically associated with a systematic error, is defined similarly to the AEL

$$Trueness = \sqrt{\left(\bar{x} - x_d\right)^2 + \left(\bar{y} - y_d\right)^2} \qquad (21)$$

where $(\bar{x},\bar{y})$ denote the mean coordinates of the estimated locations whereas (x_a,y_a) are simply the true damage position coordinates. On the other hand, precision, which is usually compromised by random errors, is computed through the area of the ellipse linked to the covariance matrix of the estimated damage locations.[182]

$$Precision = \pi ab \sqrt{\chi^2_{2, \ 95\%}} \qquad (22)$$

where a and b are the smallest and highest eigenvalues of the covariance matrix, and $\chi^2_{2, \ 95\%}$ is a two degree of freedom chi-square distribution at the 95% of confidence. An example is provided in Figure 7, where a synthetic dataset was employed.

The same authors developed a probabilistic framework based on the Bayes' law to quantify the probability of determining in the correct manner the damage location inside a selected area.[182] Leung and Corcoran developed the interesting concept of Probability of Damage Location (PDL) maps[87] that can be mathematically expressed as

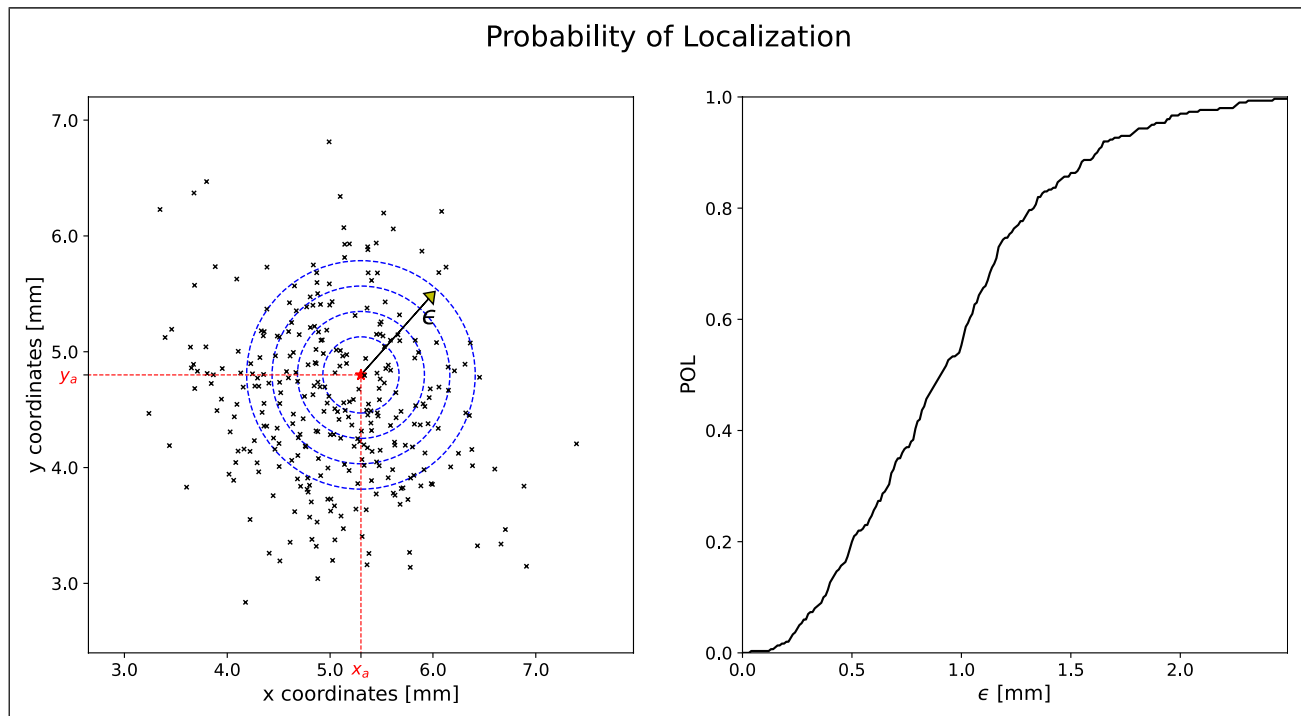$$PDL(i) = \frac{P_f(i)}{\sum_{k=1}^{n} P_f(k)} \qquad (23)$$

Figure 6. POL computed according to the methodology of Moriot et al.[181]

where the numerator represents the probability of the damage being present at the ith location, and the denominator is the sum of all these probabilities among the n number of discretized locations considered in the analysis.

There has been an increasing interest in the development of many different localization metrics. In the future much effort should be posed in defining a common methodology to lay the foundations for an accepted standard.

Probability of sizing

In the third phase of a SHM system, the main objective is to characterize the damage being previously detected and localized. Although the previous two phases do not leave room for any misinterpretation, damage characterization or identification can be confusing. Depending on the specific application there could be the necessity to quantify the damage size, classify the different damage shapes or different damage types. Taking composite material as an example, being able to classify matrix cracking, delamination, fiber breakage, and fiber pull-out is as important as determining the actual damage size.

The Probability of Sizing (POS) could be regarded as the probability of correct sizing a damage or a defect. In other words it describes the accuracy of estimating the size of a defect.[185] Attempts to evaluate the sizing accuracy of a certain measurement technique have already been made in the past.[186] For example, in Automated Ultrasonic Testing (AUT) researchers evaluate the sizing performance using the so-called safety Limit against Under Sizing (LUS). The LUS metric, also known as 95% LUS, can be thought as the lower 95% uncertainty bound of the linear regression model where the true size of a certain flaw (usually evaluated with destructive testing) is plotted against the value given by AUT.[187,188]

Annis et al. recommend being particularly cautious in the use of the LUS metric. Indeed, the assumptions upon the LUS is based, such as the linearity in the response and the homoscedasticity of the variance are not necessarily true.[55] More in general some authors, in analogy with the "â vs a" analysis, perform a regression task between the measured versus actual damage sizes.[189–192] For example, Lee et al. quantify the reliability of sizing results for axial outside diameter stress corrosion cracks, spotted near the top of tube sheet in steam generator tubes.[144] They compute the coefficient of determination r^2 related to the linear regression analysis between the true size, estimated by destructive examination, and the measured size, obtained by eddy current test (ECT). The r^2 score is then used as a reference to estimate the sizing performance of the ECT technique. Ginzel et al. retrieved the equations originally developed by Ermolov in 1972[193] in order to predict the size of flaws given by ultrasonic methods.[194] The authors
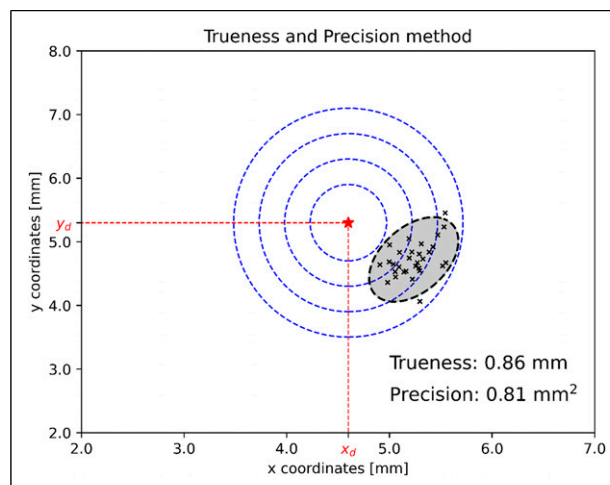

Figure 7. Location accuracy estimation with trueness and precision according to Yue and Aliabadi.[182]

pointed out that the sizing accuracy depends on many parameters, depending on the measuring technique, the material, the structure layout, the defect orientation, etc. Nath et al. proposed to assess the reliability of the Time-of-Flight Diffraction (TOFD) inspection method in terms of POD and POS curves.[189,190,195] Specifically, POS curves are developed similarly to POD curves using the â vs a method but replacing the signal response with the measured defect size in the â value. Then, the decision threshold was set arbitrarily to a certain value or equal to the maximum difference between the measured flaw size â (the depth in that specific case), and the actual flaw size. Despite the fact Nath et al. claimed to have developed POS curves, the way such curves were built is not in compliance with the definition of POS.[185] Indeed, the curves built in such way represent the probability that the estimated defect size is greater than a certain size and not the probability of correctly sizing the defect itself. Alternatively, performing several inspections on a series of representative damages, it is possible to develop a probability density function relative to the damage severity which is used to obtain an upper bound on the damage size.[196,197]

Aldrin et al. claim that a sizing metric should consist of an error with its corresponding uncertainty related confidence bounds.[23] They gave an example with a potential candidate, the so-called Normalized Quantification Accuracy (NQA), which is a metric analogous to the NLA (see Probability of localization) but related to sizing. In 2014, in the attempt of better formalizing the current sizing and localization metrics, was introduced the Characterization Error (CE), ê, which is the difference between the estimated damage characteristic (location, size, depth, width, etc.), â, and the actual damage state, a.[45] This new metric it is likely to be developed upon the mathematical framework provided

in the MIL-HKBK-1823A, but is expected to be more complicated than traditional POD studies, requiring both engineering and statistical expertise to be applied. Poor characterization results could be attributed to a low signal to noise ratio, measurements close to the saturation level, ill-posed inversion problems, failure mechanisms which are independent with respect to the defect size.[198]

Despite defect characterization, and hence sizing, represents the third fundamental level of SHM,[32] to the best of the authors knowledge there are not specific case studies which have attempted to develop such metric for SHM systems.

Discussion and perspectives

In this article, the evolution of POD and the development of localization and sizing metrics has been described. This section summarizes the most relevant studies in the field (as shown in Table A1 in Appendix) and discusses future perspectives in SHM reliability metrics.

The first observation arising from the "Field" column of Table A1 is the progressive shifting from NDE toward SHM studies, confirming the growth of the SHM field.

The "Metric" column shows that most of the studies are related to POD curves and only few of them focus on localization and sizing. There is still a high heterogeneity in metrics for localization and sizing and there are not unique well-established definitions with an accepted standard. Moreover, the relationship between these metrics should be analyzed because SHM systems may have a hierarchical approach where damage detection, locating and sizing are performed in a subsequent order. The final decision regarding the reliability of a SHM system could be based considering all these metrics.

Sequential data analysis proved to be efficient to address the important challenge of dealing with serially correlated time series data in SHM due to varying EOCs, defect morphology and sensor drift.

Statistical methodologies such as the LaD, LMM, and REM are becoming the new standards in the field. Although the LaD is a relatively simple approach, the LMM and the REM are more complex and require a sound understanding of advanced statistical tools.

The use of M-POD in SHM requires the aid of numerical models to compute POD curves as a function of several parameters. MAPOD are therefore becoming fundamental tools for the derivation of M-POD. The biggest challenge is being capable of capturing all the variability sources of the system. As the number of parameters considered increases, the computational time rises tremendously due to the curse of dimensionality. Metamodels are good candidates to solve this challenge since they can reduce the simulation time of many orders of magnitudes, allowing the production of beams of POD curves.

When multiple sensing systems are employed, POD fusion methodologies can be used. This topic is still in its infancy and is not clear yet if it makes sense or not to fuse different POD curves since it may dilute the available information. However, fusing different sensors information at a lower level could produce more effective damage indices, thus improving POD curves.

There are no methodologies available in the literature to handle hit/miss data in SHM because all the POD for SHM stem from the â vs a case. When hit/miss data are present in SHM they are erroneously treated with conventional methodologies used in NDE. Therefore, there is the need to further investigate this topic.

The "Sensing" column showcases all the sensing techniques used in reliability assessment studies. A multitude of measurement techniques have been used but at the same time others are unexpectedly rare. For instance, despite acoustic emissions[199] and distributed optical fiber sensors[200] showed to be promising technologies in the field of SHM, reliability studies concerning these techniques in terms of POD curves are still limited to few studies.

The lack of specific reliability metrics is one of the major impediments for the validation and certification of SHM system. The "Objective" column highlights that there are many studies trying to assess the reliability of a specific system, but only few studies focus on the development of new reliability metrics. This discrepancy suggests that more effort should be posed in the derivation of common standard procedures to evaluate the performance of SHM systems in terms of detection, localization and sizing.

Even though a SHM study should be conducted with a real representative structure rather than simple specimens (as is typically done in NDE), this is not entirely reflected in the "Material and Structure" column of Table A1. Despite few exceptions,[88,119] most of the analyzed literature uses simplified structure components or even specimens to develop POD curves.

The situation highlights two requirements for a POD study in SHM that go in opposite directions: the need of reproducing a test structure as similar as possible to reality which is capable to capture all the variability sources, and the need of many such structures to obtain a statistically relevant study. The number of SHM studies leveraging numerical models for the development of POD curves has an increasing trend, as shown in the "Numerical Analysis" column. However, it is often difficult to substitute completely experimental data since the model could not be able to represent accurately the real structure or to capture or the variability source. A hybrid approach where experiments and numerical simulations are both exploited seems to be the most promising strategy.[88] Bayesian statistics could be used to integrate these two kinds of information. To the best of the authors knowledge this has only been done for NDE POD studies and should be object of further investigation

also for SHM systems where the issue of having a large amount of data is even more exacerbated.

The column title "Damage and its Estimation" of Table A1 emphasizes that POD curves are plotted as a function of the *estimated* damage size rather than the *true* damage size. This point has never been addressed because it was implicitly assumed that the measurement technique used to address the true damage size was much more accurate than the value provided by the SHM system. Even if this assumption can hold, in many cases it does not or it is simply impossible to prove it. Therefore, this topic should be further investigated to implement the crack length estimation uncertainty in the current SHM methods to derive POD curves.

The concluding remarks of this review article are summarized below:

- The SHM field is increasing faster than the development of its reliability metrics. There is the need of further research to develop statistical methods capable of quantifying the performance of SHM systems.
- Is not possible to use conventional NDE POD curves in SHM.
- POD curves are usually developed assuming zero uncertainty in the damage size which is not always acceptable.
- Sequential data analysis can be used to study SHM systems under varying EOCs and to address problems where sensor drift is present. It is a promising research field which deserves further research.
- Statistical models to generate POD in presence of dependent data are available (LaD and REM) but few studies use them to produce POD curves.
- Multiple parameters can be considered using M-POD. In M-POD studies the use of MAPOD is fundamental. The curse of dimensionality can be addressed using metamodels.
- Fusing data from several sensor sources may improve the POD and ROC curves.
- Capturing all the variability sources is the key to obtain reliable numerical models and meaningful experiments.
- Bayesian methods can be used to combine experimental and numerical data in POD studies for SHM. Recent studies showed that this approach is promising.
- Despite recent studies developed metrics for localization and sizing, there is not a common well-accepted standard as for POD curves. Research is needed to produce a protocol capable of unifying these scattered efforts, using statistical tools capable to produce confidence intervals.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Francesco Falcetelli ⓘD https://orcid.org/0000-0003-2820-776X

References

1. Derriso MM, DeSimio MP, McCurry CD, et al. Industrial Age non-destructive evaluation to Information Age structural health monitoring. *Struct Health Monit* 2014; 13: 591–600.
2. Farrar CR and Worden K. An introduction to structural health monitoring. *Philosophical Trans R Soc A: Math Phys Eng Sci* 2007; 365: 303–315.
3. SAE International. *Aerospace Recommended Practice: Guidelines for Implementation of Structural Health Monitoring on Fixed Wing Aircraft*, ARP6461, 2013.
4. Balageas D, Fritzen C-P and Güemes A (eds) *Structural health monitoring*. London; Newport Beach, CA: ISTE; 2006.
5. Boller C, Chang F-K and Fujino Y (eds) *Encyclopedia of structural health monitoring*. Chichester, West Sussex, UK: John Wiley; 2009.
6. Aliabadi MH and Sharif-Khodaei Z (eds) *Structural health monitoring for advanced composite structures*. New Jersey: World Scientific; 2017.
7. Roach D. *Validation and Verification Processes to Certify SHM Solutions for Commercial Aircraft Applications*. SAND2013-7867C. Albuquerque, New Mexico: Sandia National Labs, 2013.
8. Achenbach JD. Structural health monitoring - What is the prescription? *Mech Res Commun* 2009; 36: 137–142.
9. Roach D, Rice T and Swindell P. Convergence of Multiple Statistical Methods for Calculating the Probability of Detection from SHM Sensor Networks. In: *Structural Health Monitoring 2017*. Lancaster, PA: DEStech Publications, Inc.; 2017. Epub ahead of print 28 September 2017. DOI: 10.12783/shm2017/14209.
10. Roach D and Rice T. *Wide Area Monitoring of Aircraft Structures Using Acousto-Ultrasonic Sensor Networks*. China: Xiamen, https://www.osti.gov/servlets/purl/1398324 (2016, accessed 12 October 2020).
11. Farrar CR and Worden K. *Structural Health Monitoring: A Machine Learning Perspective*. Chichester, West Sussex-Hoboken, NJ: Wiley, 2012.
12. Roach D. Does the Maturity of Structural Health Monitoring Technology match user Readiness? In: *In Proceedings of the International Workshop on SHM*. Stanford, CA, USA, 2011.
13. Seaver M, Chattopadhyay A, Papandreou-Suppapola A, et al. Workshop on transitioning structural health monitoring technology to military platforms. *J Intell Mater Syst Structures* 2013; 24: 2063–2073.

14. Cawley P. Structural health monitoring: closing the gap between research and industrial deployment. *Struct Health Monit* 2018; 17: 1225–1244.

15. Cawley P. A development strategy for structural health monitoring applications. *J Nondestructive Eval Diagn Prognostics Eng Syst* 2021; 4: 041012.

16. Department of Defense Handbook. *Nondestructive Evaluation System Reliability Assessment*, 2009.

17. Lindgren EA, Buynak CF, Aldrin JC, et al. Model-assisted Methods for Validation of Structural Health Monitoring Systems. In: *Proceedings of the 7th International Workshop on Structural Health Monitoring*. Stanford, CA; 2009, pp. 2188–2195.

18. Aldrin JC, Medina EA, Santiago J, et al. *Demonstration Study for Reliability Assessment of SHM Systems Incorporating Model-Assisted Probability of Detection Approach*. Burlington, VT, pp. 1543–1550.

19. Lindgren EA and Buynak CF. The need and requirements for validating damage detection capability. In: *Proceedings of the 8th International Workshop on SHM*. DEStech Publications, Inc.; 2011, pp. 2444–2451.

20. Chang FK. *The Need of SHM Quantification for Implementation*, https://hal.inria.fr/hal-01010064 (2014, accessed 27 April 2021).

21. Güemes A, Fernandez-Lopez A, Pozo AR, et al. Structural Health Monitoring for Advanced Composite Structures: A Review. *J Compos Sci* 2020; 4: 13.

22. Kabban CM and Derriso MM. Certification in structural health monitoring systems. In: *Proceedings of the 8th International Workshop on SHM*. Stanford, CA, USA, pp. 2429–2436.

23. Aldrin JC, Medina EA, Lindgren EA, et al. *Model-Assisted Probabilistic Reliability Assessmnt for Structural Health Monitoring Systems*. AIP, 1965–1972.

24. Medina EA and Aldrin JC. Value Assessment Approaches for Structural Life Management. In: Boller C, Chang F-K and Fujino Y (eds). *Encyclopedia of Structural Health Monitoring*. Chichester, UK: John Wiley & Sons, p. shm189.

25. Department of Defense. *Procedures for Performing a Failure Mode, Effects, and Criticality Analysis*. Washington, DC: MIL-STD-1629A, 1980.

26. Aldrin JC, Medina EA, Lindgren EA, et al. Protocol for Reliability Assessment of Structural Health Monitoring Systems Incorporating Model-assisted Probability of Detection (MAPOD) Approach. In: *Proceedings of the 8th international workshop on structural health monitoring*. Stanford; 2011, pp. 2452–2459.

27. *Environmental Conditions and Test Procedures for Airborne Equipment*. RTCA Paper No. 111-04/SC135-645, Washington, DC, 2005.

28. *Department of Defense Test Method Standard for Environmental Engineering Considerations*. 2000.

29. *Department of Defense Interface Standard Requirements for the Control of Electromagnetic Interference Characteristics of Subsystems and Equipment*. 1999.

30. Kessler SS. Certifying a Structural Health Monitoring System: Characterizing Durability, Reliability and Longevity. In: *Proceedings of the 1st International Forum on Integrated Systems Health Engineering and Management in Aerospace*. Napa, CA; 2005.

31. Janapati V., Kopsaftopoulos F., Li F., et al. Damage detection sensitivity characterization of acousto-ultrasound-based structural health monitoring techniques. *Struct Health Monit* 2016; 15: 143–161.

32. Rytter A. Vibrational Based Inspection of Civil Engineering Structures. *PhD Thesis, Dept. Of Building Technology and Structural Engineering*. Aalborg , Denmark: Aalborg University, 1993.

33. Doebling SW, Farrar CR, Prime MB, et al. *Damage Identification and Health Monitoring of Structural and Mechanical Systems from Changes in Their Vibration Characteristics: A Literature Review*. LA–13070-MS, p. 249299.

34. Derriso MM, McCurry CD and Schubert Kabban CM. A novel approach for implementing structural health monitoring systems for aerospace structures. In: *Structural Health Monitoring (SHM) in Aerospace Structures*. Elsevier, pp. 33–56.

35. Department of Defense Standard Practice. Aircraft Structural Integrity Program (ASIP). *MIL-STD-1530c*, United States Air Force, 2005.

36. Monaco E, Memmolo V, Ricci F, et al. Guided waves based SHM systems for composites structural elements: statistical analyses finalized at probability of detection definition and assessment. In: Kundu T (ed) San Diego. CA, p. 94380M.

37. Casella G and Berger RL. *Statistical Inference*. 2nd ed. Australia; Pacific Grove, CA: Thomson Learning, 2002.

38. Annis C. *False Positives*, https://statistical-engineering.com/false-positives/(accessed 22 December 2020).

39. Berens AP. NDE reliability data analysis. In: *Metals handbook Volume 17: Nondestructive evaluation and quality control*. Metals Park, Oh: American Society for Metals; 1989, pp. 689–701.

40. Kanzler D and Müller C. *How Much Information Do We Need? A Reflection of the Correct Use of Real Defects in POD-Evaluations*. Minneapolis, MN, 2008.

41. Forsyth DS. On the independence of multiple inspections and the resulting probability of detection. In: AIP Conference Proceedings. Montreal, Canada. AIP, pp. 2159–2166.

42. Wright M. *How to Implement a PoD into a Highly Effective Inspection Strategy*. Burlington, ON: Canada: NDT.net, https://www.ndt.net/events/NDTCanada2016/app/content/Paper/44_Wright.pdf (2016, accessed 29 December 2020).

43. Petrin C, Annis C and Vukelich S. A recommended methodology for quantifying NDE/NDI based on aircraft engine experience. *AGARD LECTURE SERIES* 1993; 190.

44. Achenbach JD. Quantitative nondestructive evaluation. *Int J Sol Structures* 2000; 37: 13–27.

45. Aldrin JC, Annis C, Sabbagh HA, et al. *Assessing the Reliability of Nondestructive Evaluation Methods for Damage Characterization*. Baltimore, MD, 2078, p. 2071.

46. Meeker WQRB. *Thompson's Contributions to Model Assisted Probability of Detection*. Burlington, VT, pp. 83–94.

47. Gandossi L and Simola K. Derivation and use of probability of detection curves in the nuclear industry. *Insight - Non-Destructive Test Condition Monit* 2010; 52: 657–663.

48. Kanzler D, Müller C and Pitkänen J. *Combining of Different Data Pools for Calculating a Reliable POD for Real Defects*. Idaho: Boise, 1924–1932.

49. Meeker WQ, Roach D and Kessler SS. Statistical Methods for Probability of Detection in Structural Health Monitoring. In: *Structural Health Monitoring 2019*. DEStech Publications, Inc.; 2019. Epub ahead of print 15 November 2019. DOI: 10.12783/shm2019/32095.

50. Annis C, Gandossi L and Martin O. Optimal sample size for probability of detection curves. *Nucl Eng Des* 2013; 262: 98–105.

51. Gandossi L and Annis C. *ENIQ Report No. 47*. Luxembourg: Publications Office, 2012.Influence of sample size and other factors on Hit/Miss Probability of Detection Curves

52. Annis C and Gandossi L. Probability of Detection Curves: Statistical Best-Practices. *ENIQ Report No 41*. Luxembourg: Publications Office of the European Union, https://publications.jrc.ec.europa.eu/repository/bitstream/111111111/15527/1/reqno_jrc56672_jrc56672_eniq_report_41.pdf%5B1%5D.pdf (2010, accessed 12 September 2020).

53. Koh Y-M and Meeker WQ. *Methods for Planning a Statistical POD Study*. Denver, CO, pp. 1725–1732.

54. Aldrin JC, Annis C, Sabbagh HA, et al. Best practices for evaluating the capability of nondestructive evaluation (NDE) and structural health monitoring (SHM) techniques for damage characterization. In: AIP Conference Proceeding, Minneapolis, MN, 2016. Epub ahead of print 2016. DOI: 10.1063/1.4940646.

55. Annis C, Aldrin JC and Sabbagh HA. What is Missing in Nondestructive Testing Capability Evaluation?. *Mater Eval*; 73.

56. Ameyaw DA, Rothe S and Söffker D. A novel feature-based probability of detection assessment and fusion approach for reliability evaluation of vibration-based diagnosis systems. *Struct Health Monit* 2020; 19: 649–660.

57. Schubert Kabban CM, Greenwell BM, DeSimio MP, et al. The probability of detection for structural health monitoring systems: Repeated measures data. *Struct Health Monit* 2015; 14: 252–264.

58. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; 240: 1285–1293.

59. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Lett* 2006; 27: 861–874.

60. Rouhan A and Schoefs F. Probabilistic modeling of inspection results for offshore structures. *Struct Saf* 2003; 25: 379–399.

61. Schoefs F, Clément A and Nouy A. Assessment of ROC curves for inspection of random fields. *Struct Saf* 2009; 31: 409–419.

62. Fawcett T. *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*. Technical Report HPL-2003–4. Palo Alto, CA: HP Laboratories, 2003.

63. Meeker WQ, Jeng S-L, Chiou C-P, et al. Improved Methodology for Predicting POD of Detecting Synthetic Hard Alpha Inclusions in Titanium. In: Thompson DO and Chimenti DE (eds) *Review of Progress in Quantitative Nondestructive Evaluation*. Boston, MA: Springer US, pp. 2021–2028.

64. Oehlert GW. A Note on the Delta Method. *The Am Statistician* 1992; 46: 27–29.

65. Cox C. Delta Method. In: Armitage P and Colton T (eds). *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons, p. b2a15029.

66. Annis C. *How Hit/miss Models Work*, https://statistical-engineering.com/how-hit-miss-models-work/(accessed 14 December 2020).

67. Virkkunen I, Koskinen T, Papula S, et al. Comparison of â Versus a and Hit/Miss POD-Estimation Methods: A European Viewpoint. *J Nondestructive Eval* 2019; 38: 89.

68. Knopp J, Grandhi R, Zeng L, et al. Considerations for statistical analysis of nondestructive evaluation data: hit/miss analysis. *E-Journal Adv Maintenance* 2012; 4: 105–115.

69. *Standard Practice for Probability of Detection Analysis for Hit/Miss Data*. ASTM E2862-18.

70. Harding CA. Statistical Analysis of Probability of Detection Hit/Miss Data for Small Data Sets. In: AIP Conference Proceedings. *Bellingham*. Washington (USA): AIP, pp. 1838–1845.

71. Annis C and Knopp J. Comparing the Effectiveness of a90/95 Calculations. In: *AIP Conference Proceedings*. Portland, OR: AIP, pp. 1767–1774.

72. Spencer FW. The Calculation and Use of Confidence Bounds in POD Models. In: *AIP Conference Proceedings*. Portland, OR: AIP; 1791-1798.

73. Cheng RCH and Iles TC. Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables. *Technometrics* 1983; 25: 77–86.

74. Cheng RCH. and lies TC. One-Sided Confidence Bands for Cumulative Distribution Functions. *Technometrics* 1988; 30: 155–159.

75. Li M, Spencer FW and Meeker WQ. *Distinguishing between Uncertainty and Variability in Nondestructive Evaluation*. Burlington, VT, pp. 1725–1732.

76. Li M, Meeker WQ, Thompson RB, et al. *PHYSICAL MODEL ASSISTED PROBABILITY OF DETECTION IN NONDE-STRUCTIVE EVALUATION*. San Diego, CA, pp. 1541–1548.

77. Berens A and Hovey P. Statistical Methods for Estimating Crack Detection Probabilities. In: Bloom J and Ekvall J (eds) Probabilistic Fracture Mechanics and Fatigue Methods: Applications for Structural Design and Maintenance. *100 Barr Harbor Drive, PO Box C700*. West Conshohocken, PA 19428-2959: ASTM International, pp. 79–94.

78. Forsyth DS. *Structural Health Monitoring and Probability of Detection Estimation*. Minneapolis, MN, 2004.

79. Müller I, Janapati V, Banerjee S, et al. On the performance quantification of active sensing SHM systems using model-assisted POD methods. In: *Proceedings of the 8th international workshop on structural health monitoring*. Stanford, CA: DEStech Publications, Inc.; 2011, pp. 2417–2428.

80. Janapati V, Kopsaftopoulos F, Roy S, et al. Sensor network configuration effect on detection sensitivity of an acousto-ultrasound- based Active SHM system. In: *Proceedings of the 9th International Workshop on Structural Health Monitoring*. Stanford, CA: DEStech Publ; 2013, pp. 2147–2156.

81. Pado LE, Ihn J-B and Dunne JP. Understanding Probability of Detection (POD) in Structure Health Monitoring systems. In: *Proceedings of the 9th International Workshop on Structural Health Monitoring*. Stanford, CA: DEStech Publications, Inc.; 2013, pp. 2107–2114.

82. Hayo T, Frankenstein B, Boller C, et al. Approach to the Technical Qualification of a SHM System in terms of Damage Detection in Aerospace Industry. In: *Proceedings of the International Workshop Smart Materials, Structures & NDT in Aerospace*, https://www.ndt.net/article/ndtcanada2011/papers/76_Hayo_Rev2.pdf (2011, accessed 10 April 2021).

83. Gianneo A, Carboni M and Giglio M. *A Preliminary Study of Multi-Parameter POD Curves for a Guided Waves Based SHM Approach to Lightweight Materials*. Minneapolis, MN, p. 030018.

84. Sohn H. Effects of environmental and operational variability on structural health monitoring. *Philosophical Trans R Soc A: Math Phys Eng Sci* 2007; 365: 539–560.

85. Gianneo A, Carboni M and Giglio M. Feasibility study of a multi-parameter probability of detection formulation for a Lamb waves-based structural health monitoring approach to light alloy aeronautical plates. *Struct Health Monit* 2017; 16: 225–249.

86. Mandache C, Genest M, Khan M, et al. Considerations on structural health monitoring reliability. In: *Proceedings of the International Workshop Smart Materials, Structures & NDT in Aerospace*. Montreal, Quebec https://www.ndt.net/article/ndtcanada2011/papers/17_Mandache_Rev1.pdf (2011, accessed 26 April 2021).

87. Leung MSH and Corcoran J. Evaluating the probability of detection capability of permanently installed sensors using a structural integrity informed approach. *J Nondestructive Eval* 2021; 40: 82.

88. Liu C, Dobson J and Cawley P. Efficient generation of receiver operating characteristics for the evaluation of damage detection in practical structural health monitoring applications. *Proc R Soc A: Math Phys Eng Sci* 2017; 473: 20160736.

89. Howard R and Cegla F. Detectability of corrosion damage with circumferential guided waves in reflection and transmission. *NDT E Int* 2017; 91: 108–119.

90. Tschoke K, Mueller I, Memmolo V, et al. Feasibility of model-assisted probability of detection principles for structural health monitoring systems based on guided waves for fibre-reinforced composites. *IEEE Trans Ultrason Ferroelect, Freq Contr* 2021: 1.

91. Tan Y and Zhang L. Computational methodologies for optimal sensor placement in structural health monitoring: A review. *Struct Health Monit* 2020; 19: 1287–1308.

92. Flynn EB and Todd MD. A Bayesian approach to optimal sensor placement for structural health monitoring with application to active sensing. *Mech Syst Signal Process* 2010; 24: 891–903.

93. Azarbayejani M, El-Osery AI, Choi KK, et al. A probabilistic approach for optimal sensor allocation in structural health monitoring. *Smart Mater Structures* 2008; 17: 055019.

94. Markmiller JFC and Chang F-K. Sensor network optimization for a passive sensing impact detection technique. *Struct Health Monit* 2010; 9: 25–39.

95. Mallardo V, Sharif Khodaei Z and Aliabadi F. A bayesian approach for sensor optimisation in impact identification. *Materials* 2016; 9: 946.

96. Mallardo V, Aliabadi MH and Khodaei ZS. Optimal sensor positioning for impact localization in smart composite panels. *J Intell Mater Syst Structures* 2013; 24: 559–573.

97. Yan J, Laflamme S and Leifsson L. Computational framework for dense sensor network evaluation based on model-assisted probability of detection. *Mater Eval* 2020; 78: 573–583.

98. Chen C-D, Chiu Y-C, Huang Y-H, et al. Assessments of structural health monitoring for fatigue cracks in metallic structures by using lamb waves driven by piezoelectric transducers. *J Aerospace Eng* 2021; 34: 04020091.

99. Grooteman FP. Damage detection and probability of detection for a SHM system based on optical fibres applied to a stiffened composite panel. In: *Proceedings of the 25th International Conference on Noise and Vibration engineering, ISMA2012 in conjunction with the 4th International Conference on Uncertainty in Structural Dynamics*. Leuven, Belgium: USD-Katholieke Universiteit Leuven; 2012, pp. 3317–3330.

100. Tabjula JL, Kanakambaran S, Kalyani S, et al. Outlier analysis for defect detection using sparse sampling in guided wave structural health monitoring. *Struct Control Health Monit* 2021; 28. DOI: 10.1002/stc.2690. Epub ahead of print March 2021.

101. Liu B and Chang F-K. *Generating Damage Probability-Of-Detection Curves in Structural Health Monitoring Transducer Networks*. US8069011B2.

102. Worden K, Manson G and Fieller NRJ. Damage detection using outlier analysis. *J Sound Vibration* 2000; 229: 647–667.

103. Roberts SJ. Novelty detection using extreme value statistics. *IEE Proc - Vis Image, Signal Process* 1999; 146: 124–129.

104. Markou M and Singh S. Novelty detection: a review-part 1: statistical approaches. *Signal Process* 2003; 83: 2481–2497.

105. Markou M and Singh S. Novelty detection: a review-part 2:. *Signal Process* 2003; 83: 2499–2521.

106. Cobb AC, Fisher J and Michaels JE. Model-assisted probability of detection for ultrasonic structural health monitoring.

In: *European-American Workshop on Reliability of NDE*. Berlin, Germany, https://www.ndt.net/article/reliability2009/Inhalt/th2a2.pdf (2009, accessed 16 April 2021).

107. Memmolo V, Maio L, Boffa ND, et al. Damage detection tomography based on guided waves in composite structures using a distributed sensor network. *Opt Eng* 2015; 55: 011007.

108. Yue N, Khodaei ZS and Aliabadi MH. Damage detection in large composite stiffened panels based on a novel SHM building block philosophy. *Smart Mater Structures* 2021; 30: 045004.

109. Shook BD, Millwater HR, Enright MP, et al. Simulation of Recurring Automated Inspections on Probability-of-Fracture Estimates. *Struct Health Monit* 2008; 7: 293–307.

110. Mariani S, Heinlein S and Cawley P. Compensation for temperature-dependent phase and velocity of guided wave signals in baseline subtraction for structural health monitoring. *Struct Health Monit* 2020; 19: 26–47.

111. Raghavan A and Cesnik CES. Effects of Elevated Temperature on Guided-wave Structural Health Monitoring. *J Intell Mater Syst Structures* 2008; 19: 1383–1398.

112. Clarke T, Simonetti F and Cawley P. Guided wave health monitoring of complex structures by sparse array systems: Influence of temperature changes on performance. *J Sound Vibration* 2010; 329: 2306–2322.

113. Lu Y and Michaels JE. A methodology for structural health monitoring with diffuse ultrasonic waves in the presence of temperature variations. *Ultrasonics* 2005; 43: 717–731.

114. Croxford AJ, Wilcox PD, Drinkwater BW, et al. Strategies for guided-wave structural health monitoring. *Proc R Soc A: Math Phys Eng Sci* 2007; 463: 2961–2981.

115. Michaels JE. Detection, localization and characterization of damage in plates with anin situarray of spatially distributed ultrasonic sensors. *Smart Mater Structures* 2008; 17: 035035.

116. Harley JB and Moura JM. Scale transform signal processing for optimal ultrasonic temperature compensation. *IEEE Transactions Ultrasonics, Ferroelectrics, Frequency Control* 2012; 59: 2226–2236.

117. Fendzi C, Rébillat M, Mechbal N, et al. A data-driven temperature compensation approach for Structural Health Monitoring using Lamb waves. *Struct Health Monit* 2016; 15: 525–540.

118. Yue N and Aliabadi MH. A scalable data-driven approach to temperature baseline reconstruction for guided wave structural health monitoring of anisotropic carbon-fibre-reinforced polymer structures. *Struct Health Monit* 2020; 19: 1487–1506.

119. Mariani S and Cawley P. Change detection using the generalized likelihood ratio method to improve the sensitivity of guided wave structural health monitoring systems. *Struct Health Monit* 2020: 147592172098183.

120. Mariani S, Heinlein S and Cawley P. Location Specific Temperature Compensation of Guided Wave Signals in Structural Health Monitoring. *IEEE Trans Ultrason Ferroelectrics, Frequency Control* 2020; 67: 146–157.

121. *Guided Ultrasonics Ltd.*, https://www.guided-ultrasonics.com/(accessed 4 October 2021).

122. Mariani S. *Signal Processing*, WO2020058663.

123. Lai TL. Sequential analysis: some classical problems and new challenges. *Stat Sinica* 2001; 11: 303–351.

124. Mariani S, Liu Y and Cawley P. Improving sensitivity and coverage of structural health monitoring using bulk ultrasonic waves. *Struct Health Monit* 2020: 147592172096512.

125. Mariani S, Rendu Q, Urbani M, et al. Causal dilated convolutional neural networks for automatic inspection of ultrasonic signals in non-destructive evaluation and structural health monitoring. *Mech Syst Signal Process* 2021; 157: 107748.

126. Oord AVD, Dieleman S, Zen H, et al. *WaveNet: A Generative Model for Raw Audio*. arXiv:160903499 [cs], http://arxiv.org/abs/1609.03499 (2016. accessed 6 October 2021).

127. Roach D, Rackow K, DeLong W, et al. Use of Composite Materials, Health Monitoring, and Self-Healing Concepts To Refurbish Our Civil and Military Infrastructure. *SAND2007-5547*. Albuquerque: New Mexico 87185 and Livermore, California 94550: Department of Energy - Sandia National Laboratories, https://core.ac.uk/download/pdf/71318799.pdf (2007, accessed 12 September-November 2020).

128. Roach D. Real time crack detection using mountable comparative vacuum monitoring sensors. *Smart Structures Syst* 2009; 5: 317–328.

129. Sbarufatti C, Corbetta M, San Millan J, et al. *Model-Assisted Performance Qualification of a Distributed SHM System For Fatigue Crack Detection on a Helicopter Tail Boom*. Bilbao, Spain, EWSHM. https://www.ndt.net/search/docs.php3?id=19881&msgID=0&rootID=0 (2016, accessed 8 October 2021).

130. Sbarufatti C and Giglio M. Performance qualification of an on-board model-based diagnostic system for fatigue crack monitoring. *J Am Helicopter Soc* 2017; 62: 1–10.

131. Roach D, Rice T, Neidigk S, et al. Establishing the Reliability of SHM Systems Through the Extrapolation of NDI Probability of Detection Principles. In: *Structural Health Monitoring*. Destech Publications; 2015. Epub ahead of print 2015. DOI: 10.12783/SHM2015/330.

132. Anderson TW and Darling DA. A Test of Goodness of Fit. *J Am Stat Assoc* 1954; 49: 765–769.

133. Meeker WQ, Hahn GJ and Escobar LA. *Statistical Intervals: A Guide for Practitioners and Researchers*. Second edition. Hoboken, NJ: Wiley, 2017.

134. Roach D. *Calculating Probability of Detection for SHM Systems Using One-Sided Tolerance Intervals – Applications & Limitations*. SAND2015-3998PESandia National Labs, https://www.osti.gov/servlets/purl/1258205 (2015, accessed 3 April 2021).

135. Roach D and Swindell P. Generating Viable Data to Accurately Quantify the Performance of SHM Systems. In: *Structural Health Monitoring 2019*. DEStech Publications, Inc.; 2019. Epub ahead of print 15 November 2019. DOI: 10.12783/shm2019/32220.

136. Krishnamoorthy K and Mathew T. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Hoboken, NJ: Wiley, 2009.

137. Aitken AC. IV.-on least squares and linear combination of observations. *Proc R Soc Edinb* 1936; 55: 42–48.

138. Kessler SS, Dunn CT, Swindell P, et al. Detection Sensitivity Analysis for a Potential Drop (PD) Structural Health Monitoring (SHM) System. In: *Proceedings of the 12th International Workshop on Structural Health Monitoring*. Stanford, CA, USA: DEStech Publications, Inc.; 2019. Epub ahead of print 10 September 2019. DOI: 10.12783/shm2019/32219.

139. Swindell P, Doyle J and Roach D. *Integration of Structural Health Monitoring Solutions onto Commercial Aircraft via the Federal Aviation Administration Structural Health Monitoring Research Program*. Atlanta, GA, USA, p. 070001.

140. O'Connor E. *Quantifying Method Differences in Predicting the Probability of Detection for Structural Health Monitoring ApplicationsMaster of Science (MS) in Statistics*. Iowa State University, 2019, https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1268&context=creativecomponents (accessed 1 January 2021).

141. Kessler S., Thomas G., Borgen M., et al. Carbon Nanotube Appliques for Fatigue Crack Diagnostics. In: *Structural Health Monitoring*. Destech Publications; 2015. Epub ahead of print 2015. DOI: 10.12783/SHM2015/207.

142. Mishra S, Yadav SK and Chang FK. Reliability of Probability of Detection (POD) of Fatigue Cracks for Built-in Acousto-Ultrasound Technique as "in-situ" NDE. In: *Structural Health Monitoring 2019*. DEStech Publications, Inc.; 2019. Epub ahead of print 15 November 2019. DOI: 10.12783/shm2019/32506.

143. Bode MD, Ashbaugh DM, Boyce KL, et al. Corrosion structured experiment. In: AIP Conference Proceedings. *Brunswick*. Maine: AIP, pp. 1779–1786.

144. Lee JB, Park JH, Kim HD, et al. Evaluation of ECT reliability for axial ODSCC in steam generator tubes. *Int J Press Vessels Piping* 2010; 87: 46–51.

145. Hoppe WC. *A Parametric Study of Eddy Current Response for Probability of Detection Estimation*. Kingston (Rhode Island), 1895–1902.

146. Aldrin JC, Sabbagh HA, Murphy RK, et al. *Demonstration of Model-Assisted Probability of Detection Evaluation Methodology for Eddy Current Nondestructive Evaluation*. Burlington, VT, pp. 1733–1740.

147. Aldrin JC, Knopp JS, Lindgren EA, et al. Model-assisted probability of detection evaluation for eddy current inspection of fastener sites. In: *AIP Conference Proceedings*. Chicago, IL: AIP, pp. 1784–1791.

148. Aldrin JC, Knopp JS and Sabbagh HA. *Bayesian Methods in Probability of Detection Estimation and Model-Assisted Probability of Detection Evaluation*. Denver, CO, pp. 1733–1740.

149. Pavlović M, Takahashi K and Müller C. Probability of detection as a function of multiple influencing parameters. *Insight* 2012; 54: 606–611.

150. Yusa N and Knopp JS. Evaluation of probability of detection (POD) studies with multiple explanatory variables. *J Nucl Sci Tech* 2016; 53: 574–579.

151. Gao C, Meeker WQ and Mayton D. Detecting cracks in aircraft engine fan blades using vibrothermography nondestructive evaluation. *Reliability Eng Syst Saf* 2014; 131: 229–235.

152. Meyer RM, Crawford SL, Lareau JP, et al. *Review of Literature for Model Assisted Probability of Detection*. PNNL–23714, p. 1183633.

153. *Model-Assisted POD Working Group*, https://static.cnde.iastate.edu/mapod/index.htm (accessed 1 March 2021).

154. Thompson RB, Thompson DO and Chimenti DE. A Unified Approach to the Model-Assisted Determination of Probability of Detection. In: AIP Conference Proceedings. Golden, CO. AIP, pp. 1685–1692.

155. Smith KD. *POD Transfer Function Approach*. Palm Springs, CA: MAPOD Working Group Meeting, https://static.cnde.iastate.edu/mapod/2005%20February/POD%20Transfer%20Function%20Approach_Smith.pdf (accessed February-June 2005-2021).

156. Thompson RB, Brasche LJ, Forsyth D, et al. *Recent Advances in Model-Assisted Probability of Detection*. Berlin, Germany: NDT.net, https://www.ndt.net/article/reliability2009/Inhalt/we1a1.pdf (2009, accessed 1 March 2021).

157. Gallina A, Paćko P and Ambroziński Ł. Model assisted probability of detection in structural health monitoring. In: *Advanced Structural Damage Detection: From Theory to Engineering Applications*. Wiley, pp. 57–72.

158. Austin R, Ziehl P, Yu J, et al. Development and validation of Acoustic Emission structural health monitoring for aerospace structures. In: *Proceedings of the 9th International Workshop on Structural Health Monitoring*. Stanford, CA, USA: DEStech Publications, Inc.; 2013, pp. 2123–2129.

159. Memmolo V, Ricci F, Maio L, et al. Model assisted probability of detection for a guided waves based SHM technique. In: Kundu T Las Vegas. NV, p. 980504.

160. Foucher F, Fernandez R, Leberre S, et al. *New Tools in CIVA for Model Assisted Probability of Detection (MAPOD) to Support NDE Reliability Studies*, 2018, p. 12.

161. *CIVA*, https://www.extende.com/ (accessed 6 October 2021).

162. Foucher MF, Lonne S, Toullelan G, et al. *An overview of validation campaigns of the CIVA simulation software*, 9.

163. Spencer FW, Thompson DO and Chimenti DE. *Nonparametric POD Estimation for Hit/Miss Data: A Goodness of Fit Comparison for Parametric Models*. San Diego, CA, pp. 1557–1564.

164. Dominguez N, Reboud C, Dubois A, et al. *A New Approach of Confidence in POD Determination Using Simulation*. Denver, CO, pp. 1749–1756.

165. Mahadevan S and Rebba R. Validation of reliability computational models using Bayes networks. *Reliability Eng Syst Saf* 2005; 87: 223–232.

166. Meeker WQ and Escobar LA. Introduction to the Use of Bayesian Methods for Reliability Data. In: *Statistical methods for reliability data*. New York, NY: Wiley; 1998, pp. 343–368.

167. Thompson RB, Wu HF, Diaz AA, Shull PJ, et al. (eds). *NDE simulations: critical tools in the integration of NDE and SHM*. USA: San DiegoCalifornia, p. 729402.

168. Jenson F, Dominguez N, Willaume P, et al. *A Bayesian Approach for the Determination of POD Curves from Empirical Data Merged with Simulation Results*. Denver, CO, pp. 1741–1748.

169. Syed Akbar Ali MS, Kumar A, Rao PB, et al. Bayesian synthesis for simulation-based generation of probability of detection (PoD) curves. *Ultrasonics* 2018; 84: 210–222.

170. Forsyth D and Leemans DV. *Bayesian Approaches to Using FIeld Inspection Data in Determining the Probability of Detection*, 62.

171. Kanzler D, Mueller C, Ewert U, et al. Bayesian Approach for the Evaluation of the Reliability of Non-Destructive Testing Methods. In: *18th World Conference on Nondestructive Testing*. Durban, South Africa: NDT.net, p. 6.

172. Thompson RB, Thompson DO and Chimenti DE. *A Bayesian Approach to the Inversion of NDE and SHM Data*. Kingston (Rhode Island), pp. 679–686.

173. Ameyaw DA, Rothe S and Söffker D. Probability of Detection (POD)-Oriented View to Fault Diagnosis for Reliability Assessment of FDI Approaches. 30th Conference on Mechanical Vibration and Noise. Quebec City, Volume 8. Quebec, Canada: American Society of Mechanical Engineers (ASME), p. V008T10A041.. In:

174. Ameyaw DA, Rothe S and Söffker D. Adaptation and Implementation of Probability of Detection (POD)-based Fault diagnosis in elastic structures through vibration-based SHM approach. In: *9th European Workshop on Structural Health Monitoring*. Manchester, UK: NDT.net, https://www.ndt.net/article/ewshm2018/papers/0029-Ameyaw.pdf (2018, accessed 12 September 2020).

175. Hall DL and Llinas J. An introduction to multisensor data fusion. *Proc IEEE* 1997; 85: 6–23.

176. Kralovec C and Schagerl M. Review of structural health monitoring methods regarding a multi-sensor approach for damage assessment of metal and composite structures. *Sensors* 2020; 20: 826.

177. Eleftheroglou N, Zarouchas D, Loutas T, et al. Structural health monitoring data fusion for in-situ life prognosis of composite structures. *Reliability Eng Syst Saf* 2018; 178: 40–54.

178. Gagar D, Irving PE, Jennions IK, et al. Development of probability of detection data for structural health monitoring damage detection techniques based on acoustic emission. In: *Proceedings of the 8th International Workshop on Structural Heath Monitoring*. Stanford, CA; 2011, pp. 1391–1398.

179. Flynn EB, Todd MD, Wilcox PD, et al. Maximum-likelihood estimation of damage location in guided-wave structural health monitoring. *Proc R Soc A: Math Phys Eng Sci* 2011; 467: 2575–2596.

180. Moriot J, Quaegebeur N, Le Duff A, et al. Characterization of the robustness of SHM imaging techniques using the absolute error of localization. In: *Proceedings of the 8th European workshop on structural health monitoring*. Bilbao, https://www.ndt.net/events/EWSHM2016/app/content/Paper/279_Moriot.pdf (2016, accessed 12 April 2020).

181. Moriot J, Quaegebeur N, Le Duff A, et al. A model-based approach for statistical assessment of detection and localization performance of guided wave-based imaging techniques. *Struct Health Monit* 2018; 17: 1460–1472.

182. Yue N and Aliabadi MH. Hierarchical approach for uncertainty quantification and reliability assessment of guided wave-based structural health monitoring. *Struct Health Monit* 2020: 147592172094064.

183. ISO 5725-1:1994 Accuracy. *(Trueness and Precision) of Measurement Methods and Results — Part 1: General Principles and Definitions*.

184. ISO 5725-2:2019 Accuracy. *(Trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method*.

185. Visser W. *Great Britain, Health and Safety Executive*. POD/POS Curves for Non-destructive Examination. Sudbury: HSE Books, 2002.

186. Førli O. Nordtest Report. *NT TECHN REPORT 394*. Nordtest Report, 1998.

187. Ducharme P, Rigault S, Strijdonk I, et al. *Automated Ultrasonic Phased Array Inspection of Fatigue Sensitive Riser Girth Welds with a Weld Overlay Layer of Corrosive Resistant Alloy*. NDT.net, 2012, https://www.ndt.net/article/ndtnet/2012/1_Ducharme.pdf.

188. Spencer R, Todorov E, White P, et al. *Advanced Technologies and Methodology for Automated Ultrasonic Testing Systems Quantification*. Washington, DC: EWI Project No. 50454GTHU.S. Department of Transportation Pipeline and Hazardous Materials Safety Administration, https://primis.phmsa.dot.gov/matrix/FilGet.rdm?fil=6729&s=E95FE952E286468B9FCB6FDA258E65E3 accessed 29 April 2011).

189. Nath SK, Balasubramaniam K, Krishnamurthy CV, et al. Reliability assessment of manual ultrasonic time of flight diffraction (TOFD) inspection for complex geometry components. *NDT E Int* 2010; 43: 152–162.

190. Nath SK. Effect of variation in signal amplitude and transit time on reliability analysis of ultrasonic time of flight diffraction characterization of vertical and inclined cracks. *Ultrasonics* 2014; 54: 938–952.

191. Barrett A, Smith R and Modarres M. A multivariate model to assess the probability of detection and sizing of defects in aluminum panels using eddy current inspections. *Eng Fail Anal* 2018; 94: 182–194.

192. Schneider CRA and Rudlin JR. Review of statistical methods used in quantifying NDT reliability. *Insight - Non-Destructive Test Condition Monit* 2004; 46: 77–79.

193. Ermolov IN. The reflection of ultrasonic waves from targets of simple geometry. *Non-Destructive Test* 1972; 5: 87–91.

194. Ginzel E, Ginzel R and Kanters W. Ermolov Sizing Equations Revisited. *NDT.net*; 7, https://www.ndt.net/article/v07n01/ginzel/ginzel.htm (2002, accessed 2 March 2021).

195. Nath SK. Estimates of Probability of Detection and Sizing of Flaws in Ultrasonic Time of Flight Diffraction Inspections for Complex Geometry Components With Grooved Surfaces. *J Nondestructive Eval Diagn Prognostics Eng Syst* 2021; 4: 021003.

196. Beard S, Liu C-C and Chang F-K. . In: Davis LP, Henderson BK and McMickell MB (eds). *Design of a robust SHM system for composite structures*. San Diego, CA, p. 652709.

197. Brennan F and de Leeuw B. The Use of Inspection and Monitoring Reliability Information in Criticality and Defect Assessments of Ship and Offshore Structures. *Structures, Safety and Reliability*, Volume 2. Portugal: EstorilASMEDC, pp. 921–925.. In:

198. Aldrin JC, Annis C, Sabbagh HA, et al. *Case Study on NDE Characterization Metrics for Optimization, Validation and Quality Control*. Idaho: Boise, pp. 845–855.

199. Saeedifar M and Zarouchas D. Damage characterization of laminated composites using acoustic emission: A review. *Composites B: Eng* 2020; 195: 108039.

200. Di Sante R. Fibre optic sensors for structural health monitoring of aircraft composite structures: recent advances and applications. *Sensors* 2015; 15: 18666–18713.

201. Servais P, Ibarra-Castenado C, Maldague X, et al. Probability of detection for in field thermal non destructive testing of aircraft composite structures. In: *Proceedings of the 2010 International Conference on Quantitative InfraRed Thermography.* QIRT Council; 2010. Epub ahead of print 2010. DOI: 10.21611/qirt.2010.124.

202. Duan Y, Servais P, Genest M, et al. ThermoPoD: A reliability study on active infrared thermography for the inspection of composite materials. *J Mech Sci Tech* 2012; 26: 1985–1991.

203. Kurz JH, Jüngert A, Dugan S, et al. Reliability considerations of NDT by probability of detection (POD) determination using ultrasound phased array. *Eng Fail Anal* 2013; 35: 609–617.

204. Junyan L, Yang L, Fei W, et al. Study on probability of detection (POD) determination using lock-in thermography for nondestructive inspection (NDI) of CFRP composite materials. *Infrared Phys Tech* 2015; 71: 448–456.

205. Roach D. Use of Comparative Vacuum Monitoring Sensors for Automated, Wireless Health Monitoring of Bridges and Infrastructure. In: *Proceedings of the 9th International Conference on Bridge Maintenance, Safety and Management.* https://www.osti.gov/servlets/purl/1525921 (2018, accessed 12 October 2020).

206. Heinlein S, Cawley P and Vogt T. Validation of a procedure for the evaluation of the performance of an installed structural health monitoring system. *Struct Health Monit* 2019; 18: 1557–1568.

207. Calmon P, Chapuis B, Jenson F, et al. *The Use of Simulation in POD Curves Estimation: An Overview of the IIW Best Practices Proposal*, 7.

208. Huthwaite P. Accelerated finite element elastodynamic simulations using the GPU. *J Comput Phys* 2014; 257: 687–707.

Appendix

Table A1. Case studies in NDE and SHM reliability and performance quantification assessment.

Field	Metric	Sensing	Objective	Material and structure	Damage and its estimation	Numerical analysis	Ref	Year
SHM	POD (LaD)	CVM	Reliability assessment of CVM	Thin Al specimen and thick steel specimen	FCs (initial crack lengths known from manufacturing and crack growth assessment from known initial distance between the crack and the CVM sensor)	No	[128]	2009
NDE	MAPOD (FMA)	ECT	Analyze the sensitivity of MAPOD (FMA) related to the model factors	Two-layer Al structure	Cracks around titanium fastener holes	VIC-3D®, MC simulations	[147]	2009
NDE	POD (hit/miss)	IFT	Reliability assessment of IRT	CFRP specimen	Impacts (dimension estimated with C-Scan UT)	No	[201]	2010
NDE	POD (â vs a) POS (different from def. Given in [185])	TOFD	Reliability assessment of TOFD	Al specimen	EDM vertical slits (dimension known from manufacturing process)	No	[189]	2010
NDE	M-POD (hit/miss multivariate log-logistic regression) r^2 and RMSE (true size vs measured size)	ECT	Reliability assessment of ECT	Top of tube sheet in steam generator tubes	ODSCC (dimension estimated from destructive testing)	No	[144]	2010

(continued)

Table A1. (continued)

Field	Metric	Sensing	Objective	Material and structure	Damage and its estimation	Numerical analysis	Ref	Year
NDE	M-POD (â vs a generalization)	ECT	Extend traditional â vs a POD for ECT to MP-POD.	5 different sets of flat plate specimens of different materials	Cracks (dimension estimation not specified)	No	[145]	2010
SHM	POD and PFA maps	Piezoelectric patches	Define a methodology for OSP	Different shaped plates	Defect is modeled as a symmetric scatterer	Modeling of the wave propagation	[92]	2010
SHM	POD as a function of the number of sensors	PZT	OSP for impact detection	CFRP stiffened panel	Impacts were simulated applying the load in 281 points	Abaqus model of the structure	[94]	2010
NDE	POD (â vs a) (hit/miss)	PT	Reliability assessment of PT using several PP techniques	CFRP panel	Teflon insertions at various depths simulating delamination (hypothesis: Delamination area = teflon insertions area)	No	[202]	2012
NDE	MAPOD (â vs a)	ECT	Demonstration of MAPOD methodology for ECT	Ti-6Al-4V specimens	FC and EDM notches	VIC-3D®, surrogate model	[146]	2012
SHM	POD (hit/miss)	OFS	Reliability assessment of OFS	Modeled thermoplastic affordable primary aircraft structure stiffened panel	Several damage types were simulated numerically	Only FEM simulations were used for the entire study	[99]	2012
NDE	POD (â vs a)	UPA, SPA	Reliability assessment of UPA and SPA	AUSS specimen, ferritic specimen, dissimilar metal welds, cladded specimen	Inter-granular stress corrosion cracking or FC, EDM notches, welding irregularities. (Dimension assessed with surface crack testing, radiography, creation of microsections)	No	[203]	2013
NDE	M-POD	ECT	Use of bayesian methods to evaluate M-POD and POD with physic-based models	See ref. [146,147]	See ref. [146,147]	MC, MCMC	[148]	2013
NDE	M-POD (mixed effect model)	VT	Find the best VT setting and development of POD for similar applications	Aircraft engine turbine blade (material is not specified, probably is a superalloy)	Cracks (dimension assessment using acetate replication)	No	[151]	2014
SHM	POD (LMM)	AE with PZT	Development of POD method for SHM (independence hypothesis is removed)	Wing attachment lug (material is not specified, probably is al)	FCs (fatigue cycling paused every 1000 cycles to collect PZT data and to record visual crack length information)	Simulated dataset is used	[57]	2015
NDE	POD (â vs a)	LITPI	Reliability assessment of LITPI	CFRP specimen	Flat-Bottom holes (dimension known from manufacturing	No	[204]	2015
SHM	POD (hit/miss)	Nodes in the numerical model	Reliability assessment of GLW based tomography	CFRP plate	Damage is simulated via local stiffness degradation in the FEM model	FEM using abaqus CAE	[159]	2016
SHM	POD (â vs a OSTI)	GLW with smart patch PZT network	Reliability assessment of smart patch PZT network	Beam structure like ones found in rotorcraft cabin frames and bulkheads. (Material is a not specified metal)	FC initiated from starter notches (dimension assessment with not specified NDE methods)	No	[10]	2016
SHM	MAPOD (OSTI)	FBG	Reliability assessment of FBG network	Fuselage panels of a helicopter	FC visually identified and measured with a caliper about 100,000 cycles after artificial notch creation	Finite element model	[130]	2017
SHM	POD (â vs a) POL	GLW with PZT	Reliability assessment of GLW imaging techniques	Al 6061-T6 plate	Viscoelastic disc loaded with a mass of 500 g	MAPOD and MAPOL are produced using COMSOL	[181]	2017
SHM	M-POD	GLW with PZT	Reliability assessment of M-POD for GLW	Al plates	Artificial slot was realized by a blade cut (dimension known from manufacturing)	FEM (abaqus CAE)	[85]	2017

(continued)

Table A1. (continued)

Field	Metric	Sensing	Objective	Material and structure	Damage and its estimation	Numerical analysis	Ref	Year
SHM	ROC	GLW with PZT	Development of a reliability assessment methodology under varying EOCs	Pipe specimen	Growing corrosion is represented by the constant growth rate of the cross-sectional area loss. (Damage is synthetically added to the received signal)	Finite element model	88	2017
SHM	POD (OSTI)	CVM	Reliability assessment of CVM	Metal specimen	FCs	No	205	2018
SHM	POD (LaD, REM^2)	CNT	Reliability assessment of CNT	Al-Li alloy specimen	FCs initiated from EDM notches. Image was captured every 1000 cycles to capture the crack extending)	FEM to simulate CNT resistance as a function of the crack length	138	2019
SHM	POD (â vs a) and POD fusion	SG, LS, ACC	Extend the use of POD to the field of fault detection and isolation	Steel beam	Additive masses in a vibration-based test (modal hammer excitation)	No	56	2019
SHM	POD (LaD, REM^2)	See 10,128,138,141,205	Development of POD methods for SHM (independence hypothesis is removed)	See ref. 10,128,138,141,205	See 10,128,138,141,205	See 10,128,138,141,205	49	2019
SHM	POD (hit/miss) TRN PRC	GLW with PZT	Development of a hierarchical approach for reliability assessment of GLW-SHM systems	Simple flat panels, panels with 4 CFRP Ω stiffeners and 3 Al stringers	BVID produced by low-Velocity weight-Drop impacts. BVID assessment using C-Scan	No	182	2020
SHM	ROC	GLW with guided ultrasonics Ltd. gPIMS sensor ring[121]	Development of a change detection algorithm for SHM under varying EOCs	5 m long, 8-in schedule 40 pipe	Damage was simulated by adding to the experimental signals the reflection expected from a defect. Further details in Heinlein et al[206]	Finite element model	119	2020
SHM	MAPOD (â vs a)	GLW with PZT	A POD map is used to study the effect of damage position on system reliability	Automotive component made of CFRP	Each delamination was modeled as a local change in stiffness	Elastodynamic finite integration. Technique was used to model Lamb waves	90	2021
SHM	MAPOD (as implemented in CIVA[160,207]) POD and PDL maps	bulk-wave ultrasonic sensor and a potential-drop sensor	Develop a probabilistic framework for the reliability assessment permanently installed sensors	Rectangular beam undergoing three-point and four-point fatigue bending	Square defects with side length a = 2–8 mm	Finite element model (Abaqus and COMSOL)	87	2021
SHM	ROC	GLW with PZT	Causal dilated convolutional neural networks for automatic inspection of ultrasonic signals	500 × 500 × 5 mm steel plate	Through thickness hole was introduced at the desired location by removing elements in the numerical model	Finite element model (pogo[208])	125	2021

Abbreviations

Field

NDE	Non-Destructive Evaluation
SHM	Structural Health Monitoring

Metrics

LaD	Length at Detection
LMM	Linear Mixed-effects Models
MAPOD	Model-Assisted Probability of Detection
MAPOD (FMA)	Model-Assisted Probability of Detection (Full-Model Assisted approach)
MAPOD (XFN)	Model-Assisted Probability of Detection (Transfer Function approach)
MAPOL	Model-Assisted Probability of Localization
M-POD	Multivariate Probability of Detection
OSTI	One-Sided Tolerance Interval
POD	Probability of Detection curve
POL	Probability of Localization curve
POS	Probability of Sizing curve
PRC	Precision
r^2	Coefficient of Determination
REM^2	REpeated Measures Random Effects Model
RMSE	Root Mean Square Error
TRN	Trueness

Materials

Al	Aluminum
AUSS	AUstenitic Stainless-Steel
CFRP	Carbon Fiber Reinforced Polymer
Li	Lithium

Damage

EDM	Electrical Discharge Machined
FBH	Flat-Bottom Holes
FC	Fatigue Crack
LVWDI	Low-Velocity Weight-Drop Impact
ODSCC	axial Outside Diameter Stress Corrosion Cracks

Numerical Analysis

FEM	Finite Element Method
MCMC	Markov Chain Monte Carlo
MC	Monte-Carlo

Sensing

ACC	Accelerometers
AE	Acoustic Emissions
CNT	Carbon Nanotube
CVM	Comparative Vacuum Monitoring
ECT	Eddy Current Test
FDI	Fault Detection and Isolation
FBG	Fiber Bragg Grating
GLW	Guided Lamb Waves
IRT	InfraRed Thermography
LIT	Lock-In Thermography
LITPI	Lock-in Thermographic Phase Imaging inspection
LS	Laser Sensors
PT	optical Pulsed Thermography
PZT	Lead Zirconate Titanate
SG	Strain Gauges
SPA	Sampling Phased Array
TOFD	Time-of-Flight Diffraction inspection
UPA	Ultrasonic Phased Array
UT	Ultrasonic Testing
VT	Vibrothermography