IRSN preliminary analysis on statistical methods for NDE performances assessment
Thierry SOLLIER, Christophe BLAIN, Institut de Radioprotection et de Sûreté Nucléaire (IRSN), France

Abstract
French nuclear industry has introduced, through the amendment of the RSEM code in 2013, the use of statistical methods for the assessment of NDT performances for PWRs. The previous approach was deterministic and conservative, considering the worst case for all influential parameters of the inspection.

IRSN carries out research activities on these statistical methods in order to develop its skills on this scientific and technical topic with the objective to provide, as the public expert for nuclear safety, an assessment on these tools for the regulatory body ASN. IRSN has conducted a literature review and has benchmarked a Probability of Detection (POD) approach for an inspection with eddy current done in the aeronautic industry. The results obtained with different software (mh1823®, CIVA®) are compared with an in-house Matlab® script using the "Bootstrapping" method for the assessment of the confidence intervals. Benchmark results are in good agreement. A Model Assisted Probability of Detection (MAPOD) example has then been tested for Steam Generator tube wear using a simplified model based on CIVA software. The simplified MAPOD approach gives results consistent with the industry tube repair criteria.

This work presents a preliminary study of statistical tools used to express the performances of NDT in the framework of nuclear pressure equipment. For nuclear pressure components which rupture is considered in design accidents, such as steam generator tubes, the POD approach seems appropriate for the evaluation of the inspection performances.

Introduction
Nondestructive Evaluation (NDE) is an important contributor to nuclear safety. Its role during manufacturing is to insure the highest quality achievable for production and to detect any drift in quality grade with respect to anticipated manufacturing flaws. NDE role in-service is used to periodically check for potential degradation even if none is anticipated. In this respect, NDE responds to the prevention of accident (8th principle of the IAEA fundamental safety principles [1]) and the first level of defense in depth preventing abnormal failures [2].

The French ruling for the surveillance of the main pressure components of PWRs [3] requires NDE techniques to be qualified. EDF has chosen a qualification methodology in agreement with the European Network for Inspection and Qualification (ENIQ) methodology [4] [5]. As other surveillance and maintenance activities for PWRs, EDF uses a coded practice (RSE-M code [6]) which describes the qualification methodology. The NDE performances are guaranteed by EDF qualification body. The performances are usually expressed using a deterministic and conservative method where every influential parameter for the inspection is considered in its range of variation. The influential parameters are the inspection parameters which can potentially influence the outcome of an inspection [7]. This approach is conservative but may be ineffective to evaluate the actual performances of in-service NDE since the probability is very low to observe all the influential parameters at their worst value during an actual in service inspection.

The RSE-M code has been modified in 2013 to introduce statistical tools in the NDE qualification methodology. The modifications are related to the expression of uncertainty in measurement (GUM 1995 [8]). This methodology has been applied to the qualification of the EPR Reactor Pressure Vessel (RPV) flange threads inspection [9] [10]. EDF is also considering a statistical metric more specific to NDE: the evaluation of the Probability of Detection (POD), which may also rely on NDE simulation: Model Assisted Probability of Detection (MAPOD). EDF has published preliminary work on MAPOD applied to the inspection of SG tube wear under Anti-Vibration Bars (AVB) [11] [12].

IRSN is reviewing the relevance of statistical tools applied to the qualification of NDE techniques. The preliminary results of IRSN analysis are presented here.
**Statistical approach in NDE**

The use of statistical tools for NDE qualification is new in the RSE-M code but it is not for NDE neither for nuclear safety where Probabilistic Safety Assessment (PSA) is common practice for nuclear power plants global risk assessment [13]. In NDE, the US Air Force has probably pioneered the statistic approach, starting in the early 60' with the Aircraft Structural Integrity Program (ASIP) based on a "safe-life" evaluation in response to in-flight structural failures [14]. The "safe-life" approach was questioned after subsequent structural failure in-service and in the 70', the "damage tolerance approach" was introduced as the new paradigm for United States Air Force (USAF) maintenance policy. This approach considers potential flaws resulting from manufacturing process or appearing in service. It defines inspection intervals which would prevent potential flaw's growth to threaten the structural integrity. The USAF considers this approach successful since the number of destroyed aircraft due to structural reasons is between one and two per ten million flight hours accumulated in the fleet [14]. The metric for detection which became a standard for USAF service provider is the POD and more specifically the probability to detect a flaw of size "a" with a probability of 90% for a confidence level of 95%, denoted $a_{90/95}$. The seminal papers on POD and confidence bounds evaluation were published by Daytona University [15]. USAF has published guidelines for the assessment of POD curves in the form a Military Handbook MIL-HDBK-1823A [16]. Several software tools implement the MIL-HDB-1823A specifications. MIL-HDBK-1823A has therefore become a reference for a larger audience than the initial military aeronautics sector.

In Europe, ENIQ has proposed a recommendation for POD evaluation [17] in accordance with MIL-HDBK-1823A. The Health and Safety Executive (HSE), the UK regulator, has also published recommendations on the use of POD in NDE and its limitations [18].

The American Society for Testing and Materials (ASTM) has published 2 standards for POD evaluation:

- E2862-12: Standard practice for probability of detection analysis for hit/miss data [19];
- E3023-15: Standard practice for probability of detection analysis for $\hat{a}$ versus a data [20].

Both standards refer to the MIL-HDBK-1823A. The standard E3023 is of special interest for eddy current testing (ET) and ultrasonic testing (UT) because it assumes a relationship between the flaw size (a) and the signal response ($\hat{a}$). The relationship is described using a linear regression model with an error normally distributed with mean zero and constant variance. The standard E3023 recommends non-mandatory tools for POD analysis including a model adequacy assessment, based on an analysis of variance.

For NDE performance assessment, POD is a metric commonly used and a well-established practice in the industry and in particular in the aeronautic and petrochemical sectors. It is also used in the nuclear sector, for example EPRI has developed a MAPOD approach for the quantitative evaluation of NDE performance for SG tube inspection [21].

The 2013 revision of the RSE-M code for NDE qualification is based more on the propagation of uncertainties in measurement (GUM 1995) [8] than on the POD approach described in the MIL-HDBK-1823A. The GUM 1995 classifies the measurement uncertainties in two categories:

- **Type A:** the uncertainty is evaluated from a statistical analysis of measured quantity values;
- **Type B:** the uncertainty is determined by means other than a type A evaluation.

The type B of uncertainty can be estimated, for example, by an expert elicitation.

When the measurement parameters (or influential parameters for an inspection) are uncorrelated, the combined standard uncertainty is the root square of the combined variance. The combined variance can be expressed by a first-order Taylor series expansion. More sophisticated methods for the propagation of uncertainty may be used, such as the Monte-Carlo method [22].

The level of confidence can be adjusted by multiplying the combined standard uncertainty by a coverage factor k, typical values taken for k are equal to 2 or 3 which correspond to a level of confidence of approximately 95% for $k = 2$ and respectively of 99% for $k = 3$, providing the degrees of freedom of the combined uncertainty is sufficiently high (for example superior or equal to 30).
POD evaluation for $\hat{a}$ versus $a$ approach: "Berens" benchmark

Several methods and software tools are available for the evaluation of POD curves. We consider the data provided by Berens in its 1989 publication in Table 1 [15] as a benchmark to discuss the methods and the appropriateness of the models for the $\hat{a}$ versus $a$ approach. The first step is to establish a relationship between $\hat{a}$ and $a$. Following Berens’ proposal, we consider a linear relation between $\ln(\hat{a})$ and $\ln(a)$:

$$\ln(\hat{a}) = \beta_0 + \beta_1 \ln(a) + N(0,\tau)$$  
(1)

where $\tau$ is the predictor variable or discontinuity size (true crack depth in Berens publication); $\hat{a}$ is the measured signal response for a given discontinuity of size $a$; $N(0,\tau^2)$ is a noise function (normal distribution with mean 0 and variance $\tau^2$).

The relationship between $\ln(\hat{a})$ and $\ln(a)$ is plotted in Figure 1. The saturation limit in counts is 4095 and the recording level threshold is 70. In Figure 1 (a), the software mh1823 edited by Charles Harris [23] is used as a reference and is compared in Figure 1 (b) to a Matlab® script for the evaluation of the model parameters ($\beta_0$, $\beta_1$, $\tau$). The two evaluations are perfectly matched. In the Figure 1 (b), the linear regression based on the maximum likelihood methodology is plotted in red and is compared to the ordinary least squares technique plotted in blue. The difference is clearly apparent and is due to censored data (marked with (a) and (b) in Table 1). The number of iterations for the evaluation of the parameters is 7 with mh1823 software using Newton-Raphson method and is 5 with Matlab® using "fsolve" function with the default "trust-region-dogleg" algorithms. For both cases, the number of iterations is clearly below the limit value of 20 recommended in the ASTM-E3023 standard, which proves the model and the regression algorithms are satisfactory.

<table>
<thead>
<tr>
<th>Crack depth $a$</th>
<th>Probe A peak voltage in counts $\hat{a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>1052</td>
</tr>
<tr>
<td>1.4</td>
<td>4095 (a)</td>
</tr>
<tr>
<td>0.38</td>
<td>1480</td>
</tr>
<tr>
<td>0.25</td>
<td>723</td>
</tr>
<tr>
<td>0.74</td>
<td>4095 (a)</td>
</tr>
<tr>
<td>0.46</td>
<td>2621</td>
</tr>
<tr>
<td>0.3</td>
<td>377</td>
</tr>
<tr>
<td>0.23</td>
<td>223</td>
</tr>
<tr>
<td>0.56</td>
<td>1654</td>
</tr>
<tr>
<td>1.65</td>
<td>4095 (a)</td>
</tr>
<tr>
<td>0.08</td>
<td>(b)</td>
</tr>
<tr>
<td>0.25</td>
<td>669</td>
</tr>
<tr>
<td>0.18</td>
<td>374</td>
</tr>
<tr>
<td>0.03</td>
<td>(b)</td>
</tr>
<tr>
<td>0.18</td>
<td>409 (a)</td>
</tr>
<tr>
<td>0.28</td>
<td>895</td>
</tr>
<tr>
<td>0.2</td>
<td>374</td>
</tr>
<tr>
<td>0.79</td>
<td>4095 (a)</td>
</tr>
<tr>
<td>0.23</td>
<td>638</td>
</tr>
<tr>
<td>0.15</td>
<td>533</td>
</tr>
<tr>
<td>0.08</td>
<td>150</td>
</tr>
<tr>
<td>0.28</td>
<td>749</td>
</tr>
<tr>
<td>0.2</td>
<td>433</td>
</tr>
<tr>
<td>0.36</td>
<td>879</td>
</tr>
<tr>
<td>0.23</td>
<td>286</td>
</tr>
<tr>
<td>0.23</td>
<td>298</td>
</tr>
<tr>
<td>0.41</td>
<td>1171</td>
</tr>
<tr>
<td>2.54</td>
<td>4095 (a)</td>
</tr>
</tbody>
</table>

Table 1: Berens data published in 1989 for the $\hat{a}$ versus $a$ approach [15]

(a) peak voltage above the saturation limit (b) peak voltage below the recording level threshold
Figure 1: Relationship between ln(\(\hat{a}\)) and ln(a) and evaluation of the parameters according to equation (1)

The residual plot is given in Figure 2, on the left side. The residuals are approximately spread evenly over and below the horizontal zero line which is in agreement with the recommendations of the ASTM standard. The histogram of the residuals is plotted in Figure 2 on the right hand side. The ASTM E3023 standard recommends the histogram of residual should be roughly bell-shaped and symmetrical. In this case, the histogram deviates from a perfect Gaussian curve and is not truly symmetrical. This deviation might stem from a limited number of data (29 values). The small size of the dataset is relevant for industrial qualification practice since the measurement of the true depth of flaws (a) by metallography is time consuming and expensive. From a statistical point of view, the limited number of values is responsible for a larger confidence bound.

Figure 2: Residuals plot (left) and histogram of the residuals (right)
The POD curve was evaluated using Berens’ methodology and a Matlab® script. The POD curve is plotted in blue in the Figure 3. The 95% confidence bound is plotted in green. The decision threshold for the peak voltage is 250 in counts. The estimate $a_{50}$ of the crack depth for a 50% POD, the estimate $a_{90/50}$ of the crack depth for a 90% POD and $a_{90/95}$ its upper 95% confidence bound are given in Table 2 and compared to the original values given in [15]. There is an excellent agreement for $a_{50}$ and $a_{90}$ values. The agreement for the $a_{90/95}$ value seems not as good and the authors suspect a potential glitch in the in-house Matlab® implementation of the Berens methodology for the evaluation of the confidence bound. The methodology proposed by Berens is based on the work of Cheng [24] and involves several computational steps:

- Computation of the information matrix (Fisher information matrix);
- Inversion of the information matrix;
- Computation of the variance-covariance matrix;
- Inversion of the variance-covariance matrix.

An alternative option for the assessment of the confidence bounds is to use a Bootstrapping method [24]. This method is very easy to implement and is very robust. It is based on the random sampling of the data with replacement. For instance, we have sampled 1000 times the original Berens dataset and we have evaluated the 1000 corresponding POD curves. The POD curves obtained by a random resampling of Berens dataset are plotted in the Figure 4. The confidence bounds are estimated with a percentile analysis. For example, the 95% upper bound in green is given directly by the curve including 95% of the POD curves.

A comparison of various software and approaches for the evaluation of $a_{90/50}$ and $a_{90/95}$ is given in Table 3. There is a perfect agreement on the evaluation of $a_{90/50}$. Considering the different approaches for the evaluation of the confidence bound on the POD curve, some differences can be seen for the evaluation of $a_{90/95}$. We can see a very good agreement between the value given in Berens publication [15] and CIVA® evaluation. On the other hand, and using a different methodology for the evaluation of the confidence bounds, we observe an excellent agreement between the $a_{90/95}$ value evaluated with mh1823® software and the in-house bootstrapping method based on a Matlab® script.

<table>
<thead>
<tr>
<th></th>
<th>Berens Probe A data [15]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{50}$ : estimate of crack depth for a 50% POD</td>
<td>0.15 mm</td>
<td>0.152 mm</td>
</tr>
<tr>
<td>$a_{90}$ : estimate of crack depth for a 90% POD</td>
<td>0.21 mm</td>
<td>0.207 mm</td>
</tr>
<tr>
<td>$a_{90/95}$ : upper 95% confidence bound on the estimate of $a_{90}$</td>
<td>0.25 mm</td>
<td>0.258 mm</td>
</tr>
</tbody>
</table>

Table 2 : Values of crack depth for a POD of 50% and 90%
Figure 3: POD curve with 95% confidence bound estimate using Berens methodology [15]
MAPOD example: simplified eddy current inspection for steam generator tubes

POD curves may be assessed with simulation software, this approach is generally referred as Model-Assisted Probability of Detection (MAPOD) [25]. This example addresses a simplified inspection configuration for the detection of tube wear in steam generator (SG). The inspection technique is based on the bobbin coil. In this example, we will consider only a differential channel and a frequency of 280 kHz. Current in-service inspection procedures require differential and absolute modes combined with several frequencies. Frequencies mixing may also be used producing additional analysis channels and associated detection criteria. Therefore the performances for detection may be underestimated in this example limited to one frequency and a differential mode.

The inspection geometry is given in Figure 5. The model addresses the free span area of a tube with an outside diameter of 19.05 mm (3/4") and a thickness of 1.09 mm (0.043"). The tube wear is modelled with a rectangular cross-section (in red). The axial length of the wear is 1 mm. The predictor
variable \( a \) (or discontinuity size \([20]\)) is the depth of the flaw. It ranges from 10\% to 60\% of the tube wall. The measured signal \( \hat{a} \) is the peak-to-peak voltage on the vertical axis of the display, the noise signal from the probe wobble during the bobbin coil motion inside the tube being observed on the horizontal axis.

The noise in ET amplitude and phase measurements is modelled by the probe wobble in the tube (simplified approach). The parameter associated to the noise in this CIVA MAPOD model is the bobbin offset from the tube axis (displacement along the Z-axis according to Figure 5 coordinates). The offset variation is randomly sampled from a uniform distribution with a maximum value of 0.3 mm. The ET signal is calibrated against 4 through-wall \( \varnothing 1 \) mm holes 90\° apart of the same tube section, accordingly to EDF procedure. The decision threshold is also set after EDF procedure.

The MAPOD result for this setup is given in Figure 6. The fit between \( \ln(\hat{a}) \) and \( \ln(a) \) is shown in the top-left plot. The residuals are plotted at the top-right. The residuals are well balanced around the zero line. The postulated linear model between \( \ln(\hat{a}) \) and \( \ln(a) \) is therefore appropriate. The POD curves for the decision threshold considered are given on the lower plot. The POD curve is plotted in blue and the 95\% lower confidence bound is plotted in red. The \( a_{90\%} \) value for a simplified fretting wear with an axial length of 1 mm is equal to 24\% of the tube thickness. This value gives only an order of magnitude of the performances that could be achieved during in-service inspection and the confidence bound is a pure statistical indicator. In this case, it does not provide a metric of the reliability of the inspection performances. The noise model is very simplified and only one channel is considered. A more realistic model for the noise and a more comprehensive review of all the influential parameters should be used for a quantitative assessment of the performances.

![Figure 5: ET bobbin coil CIVA model for SG tube inspection (fretting wear, in red, modelled with a rectangular cross-section)](image-url)
Using CIVA simplified simulation results dataset, the Bootstrapping technique has been applied to estimate the confidence bounds on the POD curve. The result is presented in Figure 7 and the values are compared to CIVA in Table 4. There is an excellent agreement between the estimated values of $a_{90/50}$ and $a_{90/95}$. In this simplified example, the value of the depth for a reliable detection of a tube wear (24%) is consistent with a repair criteria of 40% when the fretting root cause is a loose part identified and subsequently removed.
Discussion

This work presents a preliminary study of statistical tools used to express the performances of NDT in the framework of nuclear pressure equipment.

Firstly, the authors consider the POD approach as a suitable metric to express the performances of NDT for components whose rupture is postulated in design accident. The POD approach was initially developed for the US Air Force and is consistent with a damage tolerance approach and a potential failure. This is the case for steam generator tubes. A Steam Generator Tube Rupture (SGTR) is a design accident and several degradation mechanisms are identified and, in some cases, are active such as fretting wear under anti-vibration bars or tube wear correlated to loose parts. Based on structural analysis and anticipated degradation kinetic, utilities have defined repair criteria and in-service inspection intervals with respect to the NDT performances for detection and sizing. In this context and providing that a significant safety factor is defined, the POD approach seems appropriate.

Secondly, the authors consider the GUM standard [8] suitable to express uncertainties in flaw’s location and sizing but less appropriate to express the detection performances. For the authors, the appropriate statistical tool for a quantitative evaluation of the performances in detection is the probability of detection (POD) as defined in ASTM standard E3023 [20] or ENIQ report 41 [17].

Acknowledgements

The authors are grateful to Charles Harris from Statistical Engineering for providing the mh1823 software.
References

[5] Jardet, P., Dib, M., Comparison between NDE qualification standards and practices in different countries, 10th International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurized Components, Cannes, France, 1-3 October 2013
[6] RSE-M In-Service Inspection Rules for Mechanical Components of PWR Nuclear Islands - AFCEN, 2010
[18] Probability of Detection (PoD) curves, derivation, applications and limitations, HSE research Report 454, 2006
[21] Steam Generator Integrity Assessment Guidelines, revision 2, EPRI report 1012987, 2008