How to Implement a PoD into a Highly Effective Inspection Strategy

Michael WRIGHT

Eclipse Scientific; Waterloo, Ontario, Canada; Phone: 1-(519)-886-6717, Fax: 1-(519)-886-1102; e-mail: mwright@eclipsescientific.com

Abstract

The probability of detection is a mathematical model that associates the likelihood, or frequency an NDT system, technique and/or inspector will detect a specific sized indication. For most engineering models it is very important to know what can be detected with a 90% probability. Also, what defect size yields a confidence level of 90-95%. The confidence level is very similar to the concept of repeatability.

Inspection strategies can be optimized several ways using PoD model values. Evaluating the PoD can be used as a way to quantify inspection capabilities of a technique, inspection system/equipment and the capabilities of the inspection personnel.

There are several types of PoD models. The most common are the à vs. a, binary or hit-miss, Model Assisted (MAPoD) and several advanced models.

Model Assisted PoD (MAPoD) is a newer model building technique that can be done using several methods. Data can be collected and analyzed statistically to build the final PoD model. Data that has been obtained using modelling software can be evaluated and compared with true inspection samples to create the final PoD model. Also, a mathematical simulation based evaluation can be performed to evaluate the final PoD.

There are also several advanced PoD models available for deeper mathematical evaluation or for rare inspection situations.

Keywords: Probability of Detection (PoD), Confidence Level (CL), Model Assisted PoD (MAPoD), optimization, qualification, technical justification.

1. Introduction

There are many levels to an inspection development and several experiments that are performed to gather supporting evidence. As the experiments are performed the analysis helps optimize the inspection itself. All experimentation begins by taking a single measurement. Experiments are designed to help quantify how repeatable measurements are. This helps reveal all of the sources of variation. These sources are called essential and influential parameters. The PoD model, Probability of False Call (PoFC) evaluation, and the noise study are directly linked to each other. Once all three studies are complete, an inspection can be optimized to help produce the highest PoD and the lowest PoFC.

The inspection development process is very similar to the steps in the scientific method.

2. Inspection Developments

Inspections start out as simple techniques. A technique is a method that produces a repeatable measurement or test. Techniques are incorporated into written instructions and are the building blocks of procedures. An instruction outlines the steps required to perform a technique and to produce the same values regardless of the inspection personnel performing the tests. Instructions are created to provide a solution to an inspection situation. As more evidence is gathered an instruction can be turned into a procedure. Procedures are much more detailed than an instruction since they address code and specification requirements. A good procedure has evidence that shows the instructions and techniques meet the requirements of the inspection. Figure 1 shows the standard process flow of a typical inspection development.

![Figure 1 – The inspection development road map](image-url)

With all the statements of a procedure meeting code and specification requirements, a procedure can be approved. Once a procedure is approved, it can be further reviewed and qualified by an independent qualification board.
2.1 The Qualification Roadmap

Qualification traditionally requires all of the supporting evidence to be gathered into a document called a technical justification (TJ). A TJ traditionally contains very specific NDT studies, evaluations and demonstrations. Figure 2 shows the inspection qualification road map.

The measurement of an NDT system’s detection capabilities is evaluated using a mathematical model called the Probability of Detection (PoD). Due to the time and costs involved in evaluating a PoD model most inspections do not develop a full PoD model. When personnel, equipment, and/or procedure qualification is required a full PoD model is normally required to demonstrate an inspection or part of an inspection is meeting the inspection specification requirements. Qualification is demonstrated and/or captured in a technical justification document. This process has been standardized for nearly all inspection developments.

![Figure 2 – The inspection qualification road map](1)

The TJ is a living document that acts as a canister to gather all the supporting evidence that a procedure meets code and/or specification requirements. Also, all the claims in the procedure are demonstrated and this evidence is included as part of the TJ document.

TJs normally contain several sizing accuracy studies. A sizing accuracy study is traditionally included for all reported variables that have tolerance requirements stated in a specification or by code. Sizing accuracy results are directly linked to calibration curve design.

Parametric studies are performed to evaluate how significant a parameter causes variation in the inspection results as well as to learn the parameter’s boundary values for the inspection.

TJs normally contain modelling results as well as real inspection data. A PoD study is normally only composed of real inspection data. With the advancement of computers and the availability of modelling, MAPoD is becoming increasingly important.

The PoD study is considered to be an essential component of a TJ, but it is also the most complex of the studies to evaluate properly. A higher understanding of statistical and experimental techniques is normally required to perform a good PoD study. This knowledge is not commonplace in the NDT industry and it results in many PoD studies being weak or poor. The PoD mathematical model was derived from a binomial PoD model. PoD models have evolved and currently there are several mathematical PoD models available.

3. Evolution of the PoD Model

3.1 The Beginning of the NDT PoD Model

During the days of the space race, evaluating an inspection system’s detection capabilities became very important.

Performing a full PoD was not feasible due to the availability of samples, computation requirements and costing. NASA created the first methodology for demonstrating an inspection system’s capabilities. This was the NASA 29 of 29 test. The NASA 29 of 29 test evaluates a point of the traditional PoD curve using a binomial probability model and the exact Clopper-Pearson confidence interval.
The upper confidence limit is defined as: The upper limit $P_u$ such that there is only a $(1 - \alpha)$ probability that the true proportion is $P' > P_u$.

The lower confidence limit is defined as: The lower limit $P_l$ such that there is only a $(1 - \alpha)$ probability that the true proportion is $P' < P_l$.

Where, $P'$ is the true proportion $P_u$ and $P_l$ are the upper and lower proportion confidence limits.

The upper and lower confidence limits are evaluated using the following steps:

- Calculate the degrees of freedom ($v_1, v_2$)
  - Lower limit: $v_1 = 2(X + 1) \& v_2 = 2X$
  - Upper limit: $v_1 = 2(N - X + 1) \& v_2 = 2(N - X)$
- State/specify the $\alpha$ limit required.
- Evaluate the appropriate value of the F-distribution ($F_{a}(v_1, v_2)$).
- The exact Clopper-Pearson confidence intervals are calculated using the following:
  - For the lower limit:
    $$P_l = \frac{X}{X + (N - X + 1)F_{a}(v_1, v_2)}$$
  - For the upper limit:
    $$P_u = \frac{(X + 1)F_{a}(v_1, v_2)}{(N - X) + (X + 1)F_{a}(v_1, v_2)}$$

### 3.2 Current Binary PoD Models [1]

A binomial PoD model uses 0 for a missed indication and 1 for a hit or detection. This is a 0 or 1 based model called a binary PoD model.

PoD function is related to the binary data using Generalized Linear Models (GLM). Using a GLM, first a linear regression model is used to define the initial estimate ($f(a_i)$) of the binary response of different crack sizes. This estimate is in the form of Equation (1):

$$f(a_i) = \beta_0 + \beta_1 a_i$$

Equation (1)

However, this initial estimation cannot be used to relate the indication sizes to the PoD model. The reason is that using the linear regression model $f(a_i)$ varies continuously with the indication size. This is not accurate since the data nature is binary and bounded. To overcome this problem, the GLM links the estimated values of the binary data to explanatory variables through the probability of detection, which varies continuously from 0 to 1.

Four common binary link functions, Logit, Probit, LogLog and cLogLog [3] are listed below:

- **Logit**
  $$f(a_i) = \ln(p(a_i)/(1 - p(a_i)))$$
  Equation (2)

- **Probit**
  $$f(a_i) = \Phi^{-1}(p(a_i))$$
  Equation (3)

- **LogLog**
  $$f(a_i) = \ln(-\ln(1 - p(a_i)))$$
  Equation (4)

- **cLogLog**
  $$f(a_i) = -\ln(-\ln p(a_i))$$
  Equation (5)

Once any of the above options are used the values of $\beta_0$ and $\beta_1$ should be estimated properly to reach the maximum likelihood of the model results. The idea is to select model parameter estimates such that the likelihood is maximized based on the model, given the actual observed inspection data.

For hit/miss testing, the likelihood of $P$, based on a set of observations is:

$$L(\theta) = \prod_{i=1}^{n} P_i^{x_i} (1 - P_i)^{1-x_i}$$

Equation (6)

Where $\theta = \{a_1, a_2, a_3, ..., a_n\}$ is a set of data, $P_i$, PoD(a), is the probability of detection of crack size $a_i$, and $x_i$ is the inspection outcome, 0 for miss, 1 for hit.

### 3.3 Other PoD Models [1] [3]

There are other PoD models that use the inspection amplitude instead of a binary hit/miss detection. Amplitude response data has more information that can be applied to the PoD model development process. This was the initial reason for the development of an amplitude response PoD model, or as it is more commonly known, the $\tilde{a}$ vs $a$ PoD model [1]. Amplitude based PoDs are very powerful models due to the increase in information.
The relationship between \( \hat{a} \) vs. \( a \) can be expressed:

\[
y = \beta_0 + \beta_1 x + \varepsilon
\]

\[
z = \frac{y - (\beta_0 + \beta_1 x)}{\delta}
\]

Let the standard normal PDF (Probability Density Function) be represented as:

\[
\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}
\]

Equation (7)

Let the standard normal survivor function be represented as:

\[
Q(z) = \int_{z}^{\infty} \phi(\varphi) d\varphi
\]

Equation (8)

The PoD function simplifies to:

\[
PoD(a) = 1 - Q \left[ \frac{\log a - \mu}{\sigma} \right]
\]

Equation (9)

The likelihood function has the following form when accounting for censored data (misses and saturated data samples):

\[
L(\theta) = \prod_{i=1}^{m} (1 - Q(z_{th})) \prod_{i=m+1}^{m+r} \frac{1}{\delta} \phi(z) \prod_{i=m+r+1}^{n} Q(z_{sat})
\]

Equation (10)

Both the binary and \( \hat{a} \) vs. \( a \) models require a large set of sample points. With the availability of modelling software, there is a large desire to supplement the required number of samples needed to evaluate a PoD model.

### 3.4 Model Assisted Probability of Detection (MAPoD)

There are several methods available to join modelling with the PoD mathematical models. This enables building a PoD model using less real physical data. The two most common methods involve statistical and/or finite element methods. [1]

A statistical build MAPoD relies almost entirely on real samples that are scanned with repeated trials.

Building a statistically designed MAPoD begins with a statistical investigation phase. The first test to perform is the evaluation of each target type to determine if they are Independent and Identically Distributed (IID). The IID test is the most common step that is skipped or missed during a good PoD development. The IID test is not unique for MAPoD, but it should be applied for all PoDs. If different defect mechanisms are not identically distributed they will require separation during the evaluation of the model parameters. All targets that are not IID should have a separate qualification that includes a PoD model for each mechanism. If different target types have the same distribution response, that is if they are IID, then they can be grouped into the same PoD.

Most initial development has a small set of targets available. This is a set of 5 targets as a minimum, but 9+ targets are still recommended. This set of targets can be a simple feasibility set that shows the effect of a parameter, or enables building a calibration curve. This set of indications is scanned with repeated trials 30 to 35 times. This enables building response distribution models for each target in the small set. Once the distribution parameters for each target is available a function can be fit to the model parameters.

With the distribution parameters across the range of targets a random sample of points can be generated. Each set of points is generated using the parameters from the function that was generated using the statistical evaluation of the real target response data. All generated points should be weighted with regards to their distribution type, as well as the distribution parameters. This ensures a generated set of points will represent the possible inspection data.

Each set of points is inserted into the PoD parameter generating algorithm and the final model parameters are saved. The generation of points is repeated and the PoD parameters are solved. This point generating and PoD solving is repeated for many trials. The number of trials can be in excess of 10000+ iterations.

As the iterations are completed, each model parameter is plotted with regards to the other model parameters. Each pair of parameters has a likelihood from the generated set of points. The likelihoods are compared and the one model with the highest likelihood is used to represent all the possible points. This point is the generated set of data that represents all the possible data with the highest likelihood and this single set of data represents the best fit for the entire data range that is possible.
Figure 3 shows a MAPoD set of generated points and Figure 4 shows the $\beta_0\beta_1$ solution field and the generated set with the maximum likelihood.

**4. Building a PoD Model**

**4.1 PoD Model Fit Evaluation** [1]

In statistics, many tests are available to measure how well a model fits the data it was constructed from. This is called measures of goodness of fit. The most common measures of model fit for PoD models are the AIC and/or the BIC.

AIC value is calculated using the following formula:

$$AIC = 2k - \ln(L)$$  \hspace{1cm} \text{Equation (11)}

Where $k$ is the number of estimated parameters and $L$ is the maximum likelihood of the model. AIC value indicates an estimation of the amount of information lost using a given statistical model. Therefore, lower AIC values indicate a more accurate model for a given data set.

BIC is another measure of the quality of a model to be used for a data set. BIC is very similar to AIC, with the difference of consideration of the number of parameters used for building the model into account, in addition to the data used for the model. Commonly, the likelihood of the model can be increased if more parameters are added to the model relative to the number of available data points, but this can have drawbacks such as over fitting. Therefore, increasing the relative number of parameters to the number of data points available can have penalties, in terms of measuring the quality of a model.

This penalty is considered in BIC as the following formula:

$$BIC = -2 \ln(L) + k \ln(n)$$  \hspace{1cm} \text{Equation (12)}

**4.2 PoD Model Selection**

The goal of selecting the PoD model should be based on the model representing the inspection situation to the highest accuracy.

First, the model selection process is started by choosing the main model type. If the data is amplitude response based, the $a$ vs. $a$ model should be used. If the data is not amplitude based then a binary or hit/miss model should be used. Whenever there is any uncertainty if the data is amplitude based, the model should be selected based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC and BIC values represent how well a set of data fits a model. When two models have the same AIC/BIC values, the logic of Occam’s razor should be applied. Occam’s razor states that the simplest explanation is likely the most correct explanation. Finally, if the data has a strong amplitude response, always default to the $a$ vs. $a$ model.

For an $a$ vs. $a$, if the target response is normally distributed, use the data type that produces the lowest AIC/BIC values excluding $\ln(a)$ & $\ln(\hat{a})$. If two data types have the same AIC/BIC values, apply the logic of Occam’s razor. If the target response is not normally distributed, the best data type to use is $\ln(a)$ & $\ln(\hat{a})$. Whenever there is any uncertainty if the target response is normally distributed the data type should be selected using the lowest AIC/BIC values. When two models have the same AIC/BIC values, the logic of Occam’s razor should be applied.
For a binary or hit/miss model, if the target response is normally distributed, the Logit or Probit model should be selected based on the model that produces the lowest AIC/BIC values. As stated above, when two models have the same AIC/BIC values the logic of Occam’s razor should be applied. Whenever there is any uncertainty about the distribution type, the binary model should be selected using the AIC/BIC that produces the lowest value and also respecting Occam’s razor for models with the same AIC/BIC values.

If the target set is not normally distributed, right skewed data (biased with too many large targets) should use the LogLog model. Left skewed data should use the cLogLog model. Whenever unsure how the data is skewed the LogLog and cLogLog models should be selected based on the model that produces the lowest AIC/BIC values. As with all model selection, when two models have the same AIC/BIC value, the model should be selected using the logic of Occam’s razor (LogLog model).

4.3 PoD Model Confidence Level Evaluation [1]

There are several mathematical methods available to calculate a confidence interval. The most common are the method of Chen & Iles, Wald, Wilson-Score & Clopper-Pearson methods. Not every confidence interval calculation is appropriate for all situations. This is similar to the idea of model fit using the AIC’s & BICs.

With model fitting contexts, the Wald interval about $\hat{\theta}$ can be used because an ideal $\hat{\theta}$ is modelled and the fit is computed from the ideal.

For MAPoD, the Wald test should be used, while for real PoD data, the Wilson interval will handle the variation of the data better. The Wilson-Score method has less impact from outliers as well. When plotting observed $\hat{\theta}_s$, the Wilson interval should be applied.

For the binary PoD models, the Wald is the most common method due to its ease in calculation, but this is not always the method that produces the highest accuracy. The method that represents the data with the highest accuracy, with regards to the inspection requirements should be the method chosen.

With less than 100 samples, the Wald method can be either very liberal or very conservative in its approximation. For larger sample sets and probabilities between 10% and 90%, the Wald method will work with a high degree of accuracy.

For sample sets less than 100 samples, the Wilson-Score method can be used with a much higher degree of accuracy. This does not mean that the Wilson-Score method should be used for all cases.

For small sample sets ($\leq 25$ samples) and probabilities $<10\%$ and $>90\%$ the exact method (Clopper-Pearson) will produce the highest accuracy. For larger sample sets ($>100$ samples) and probabilities $<1\%$ and $>99\%$, the exact method will produce $<2\%$ error. The exact method also works well with extremely small or large probabilities.

For situations where no error can be tolerated, the exact method should always be used. This requires the proper number of samples. The exact method is guaranteed to never fall below a nominal level. As the accuracy is increased, the required number of samples is increased as well.

5. Optimization of a PoD Model

A PoD model is directly linked to the noise study and the Probability of False Call (PoFC) study.

For amplitude and binary based PoD models, there is a threshold when a signal is classified as a defect or not. The threshold is the link value between the noise study, PoD and the PoFC.

By varying the threshold an inspection procedure can have its PoD maximized, while keeping the PoFC as low as possible. As the threshold is lowered, more signals are classified as defects. Since the noise response is fixed relative to the threshold, signals that are noise will be falsely reported as defect signals. This is one of the ideal ways a PoD can be used to help optimize an inspection technique.

Any indication response at or above 40% SH will be reported as detected or as hits (1 for a binary PoD).
An example PoD model is shown in Figure 5. This PoD model was constructed using a threshold that was set to 40% SH.

Figure 6 illustrates the effect on the PoD curve when varying the threshold from 15% to 80%. Each color in the graph represents a new PoD model that was generated using a specific threshold value as listed in the graph's legend.

Figure 7 shows the size of indication that meets the required a90/95 qualification level. As the threshold is reduced, the size of the target that meets the a90/95 level also reduces.

Figure 8 shows the final PoD model built using 30% SH. It is common for UT procedures to set a threshold at or around 40% Screen Height (SH). All signals at or above 40% SH will be reported as defect signals. Amplitude based PoD models can have greatly different threshold values and levels. For example, ET inspections can have voltage as the threshold measurement. 1.0 Volts peak-to-peak can be an example threshold level.

A threshold should be kept as high as possible and not reduced unless there are specific gains from the effect on the PoD levels. This always keeps the PoFC as low as possible, but it enables an inspection to possibly meet an inspection requirement a higher threshold might not be capable of reaching.

Most procedures do not have this optimization performed for various reasons. As the inspection industry is being driven to detect smaller indications, the advantages for optimizing the PoD and PoFC are becoming increasingly important. New inspection developments constantly demand the ability to detect smaller indications or to detect indications with higher confidence and repeatability.
6. What makes a good PoD Study?

A good PoD study has the following characteristics:

- Uses proper inspection techniques that represent the inspection being evaluated as close as possible. If the technique being evaluated is not consistently used, the resulting PoD model will not represent the inspection it is being constructed for.
- Model selection is based on model fit parameters, such as AIC or BIC values.
- It is constructed using a fair statistical set of samples that cover the entire variable range.
- It is constructed using IID samples only.
- The influence and/or significance of variables is known. If more than one factor directly influences the signal response, each factor should be covered using the appropriate range of samples.

A good PoD will be constructed fairly. Fair PoD evaluations can be compared with good accuracy. Comparing fair PoDs makes sense, while comparing two randomly generated PoDs does not make any sense at all.

Comparing the final values from the PoD model, such as the a50, a90 and/or the a90/95, can reveal important improvements during a high level inspection development. Comparing a PoD level is a great method to help quantify a change to a technique. This evaluation can enable a quantification for any inspection improvements.

7. References