A General Approach To Model Assisted Qualification of Non-Destructive Inspections

Alexander Ballisat

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Engineering Doctorate in the Faculty of Engineering.

May 14, 2019

Word Count: 34789

Declaration of Authorship

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed:

Date:

Abstract

Non-destructive inspections are a cornerstone of a wide range of safety critical industries in which the assurance of the integrity of equipment is essential. Prior to any technique being used in service, it is necessary to qualify the inspection by demonstrating that it is capable of detecting defects of interest. Traditionally, this has been achieved using expensive and time-consuming experimental trials. Within the Ministry of Defence (MOD), this has proved to be a significant barrier and an alternative methodology is needed. This project was instigated by the Defence Science & Technology Laboratory (DSTL) to develop a methodology which will allow this barrier to be overcome.

The use of numerical models to simulate inspections has become increasingly common over the past few decades with the advent of cheaper and more powerful computing resources. This presents the opportunity to replace a significant proportion of experimental trials with faster and cheaper numerical simulations. This thesis presents a general approach to achieving this as well as demonstrating other useful information that a model-based approach can yield. A generalised method of calculating metrics of inspection capability is demonstrated, making no assumptions as to the nature of underlying probability distributions or the response of the inspection. Appropriate sampling and interpolation methodologies provide tools to accurately and rapidly map the inspection's response. Sensitivity analysis is shown to be a suitable tool for quantitatively assessing which parameters can be ignored due to having little impact on the response of the inspection. These methods are applied to a canonical example inspection in the aerospace industry, demonstrating that the reliability of an inspection can be quantified using a range of metrics in a time frame suitable for the MOD.

Acknowledgements

Firstly my thanks to Paul Wilcox for all the help, support and thoughtful discussions about reliability in NDT. I would also like to extend my thanks to Robert Smith and David Hallam for all their help with the industrial application of the research and many interesting discussions on applying this in the real world. Finally, none of this would have been possible without Laura and all her support and encouragement over the years, I can not thank you enough.

For Richard.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Non-Destructive Testing (NDT), also known as Non-Destructive Evaluation (NDE), is the notion of using non-invasive methods, such as ultrasound and x-rays, to inspect the interior of structures without causing damage. This has become a crucial tool for many safety critical industries, such as aerospace, power generation and fossil fuel pipelines. In these sectors the cost of failure can be enormous, with financial, human and environmental impact, therefore there is a need to identify components which are damaged prior to a catastrophic failure.

Significant resources are expended researching and developing novel technologies which can find smaller defects more reliably. However there is always a need, especially in these typically very conservative industries, to demonstrate the capabilities of a technique prior to its introduction to service. This is a problem as old as NDT as a field and the classical approach to achieving this has been to perform empirical trials. Such trials typically involve many samples containing defects and several operators who perform blind trials to demonstrate that defects can be detected reliably. However, in industries where samples are difficult to obtain from in-service equipment or costly to manufacture, such as military aerospace, this can become a prohibitively expensive process. Furthermore, the cost of using qualified operators and attempting to perform the trials in an environment representative of the in-service inspection further increase the outlay required to perform these trials. This is especially true within the Ministry of Defence (MOD), with this barrier being highlighted as the primary obstacle to the introduction of new techniques [1]. The result of this within the MOD is that no new NDT modalities, such as phased array ultrasonics, have been introduced into service in the last 15 years [1]. Given the clear desire and need for increasingly advanced and capable technology, a

method is required that can overcome this barrier.

The increase in accuracy and sophistication of numerical models of inspections coupled with the massive increase in computing power over the last few decades presents the opportunity to replace a significant proportion of the experimental trials with numerical model evaluations, significantly reducing the number of samples required and the time required to perform experiments. This potential has been further enhanced with the advent of massively parallelised calculations utilising graphics processing units (GPUs). These coupled together potentially allow for sophisticated models which can accurately represent the inspection to be evaluated rapidly. This presents the opportunity to assess the impact of variations that may be present in an inspection on the outcome through modelling. Given that this will require every possible outcome of the inspection to be evaluated to fully quantify the effect of variations, optimisation of the inspection technique should be possible as a by-product of the qualification methodology. This is a significant benefit of this process as it should be able to improve the inspection as well as demonstrating its reliability.

The aim of this project is to develop a generic methodology to assess the reliability of a technique, primarily through the use of numerical models whilst minimising the number of experimental trials required. This methodology will be applicable to any inspection modality. Given the need to demonstrate the viability of the methodology, ultrasonics was chosen as the demonstrator modality, given its widespread use in the military air domain.

This project was instigated by the Defence Science and Technology Laboratory (Dstl) on behalf of the MOD. The Engineering Doctorate (EngD) is itself a part of a larger project to write and demonstrate a qualification protocol for the military air domain, which has been sub-contracted to The Welding Institute (TWI). The goal of this EngD is to develop methods which allow models of inspections to be used efficiently to demonstrate the capability of a technique. Full scale demonstration, including the performance of systematic experimental trials with several qualified operators, is to be left to TWI. The methods presented in this thesis will feed into this protocol and the author has been playing an active role in the development and writing of the protocol. With this in mind, this thesis also considers the practical implementation of these methods so that it can form a good description of the methods should an organisation desire the use of them. It has been found during the course of this research that the optimisation of the workflow of a qualification can yield significant

reductions in the qualification time as well as minimising the number of model evaluations.

The outline of the thesis is as follows. Contemporary qualification methodologies are reviewed in Chapter 2, highlighting their features and shortcomings as well as definitions of reliability of inspections currently in use. Methods for calculating the metrics of inspection reliability are presented and compared in Chapter 3. The efficient mapping of inspection responses is discussed in Chapter 4, presenting a method which minimises the number of model evaluations required. Chapter 5 applies this methodology to an example inspection of increasing complexity, developing a model and considering several degrees of freedom in the inspection. The results of this project are summarised in the conclusions in Chapter 6 along with a discussion of future work that leads on from this thesis.

Chapter 2

Contemporary Qualification Methodologies

This chapter presents an overview of the literature regarding the qualification of NDT techniques, highlighting some of the limitations of such methods and the assumptions that underlie them. Overall, there is not a large volume of literature on methods of qualification however there are many examples of the application of these methods. It is also important to highlight current practices to understand the requirements of a qualification methodology that could be used in industry and thus maximise the impact of the work.

2.1 Definition of Common Reliability Metrics

This section outlines the general definition of some reliability metrics that are commonly used to assess the capability of a technique in contemporary qualification methodologies. The calculation of these metrics is not trivial and is discussed in greater detail in the next Chapter.

The response of an inspection given a set of parameters $\mathbf{x}$, that is quantities that may vary between instances of the performance of the inspection, is given by the function $R(\mathbf{x})$. A sentencing criterion is applied to this response, often based on some decision threshold T. In practice, this may be anything from a simple threshold on the amplitude of a signal to a more complex analysis methodology involving multiple criteria. A decision threshold α on the parameter of interest x_c, for example the crack length having a certain magnitude, is also required. **The Probability of Detection (PoD) is defined as the**

probability that the response of an inspection is greater than a decision threshold on the response, given that a defect with magnitude of a critical parameter greater than the critical parameter threshold is present, that is

$$\text{PoD} = p(R(\mathbf{x}) \geq T | x_c \geq \alpha), \tag{2.1}$$

where $p(a)$ indicates the probability of the quantity a occurring. This is the simplest and most commonly used definition of PoD. Variations of this decision criteria will in general simplify to a conditional probability of a form similar. **The Probability of False Alarm (PFA) is the probability that the response of an inspection is greater than the decision threshold on the response, given that a defect with magnitude of a critical parameter less than the critical parameter threshold is present**. This covers both the case when a defect is present but the response indicates that it is larger than it truly is, that is $0 < x_c < \alpha$, and when a defect is not present but a response is obtained from the inspection greater than the response threshold. Both cases result in the operator taking some action they would not otherwise take if its true magnitude was known. In this case the PFA can be defined as

$$PFA = p(R(\mathbf{x}) \geq T | x_c < \alpha). \tag{2.2}$$

The PoD is typically plotted as a function of the defect parameter of interest, a plot known as the PoD curve. It may also be plotted as a function of the PFA as T is varied, a plot known as a Receiver Operator Characteristic (ROC) curve. Other reliability metrics can also be defined using conditional probabilities in the same manner, however these two are the most commonly used and are the focus of contemporary qualification methodologies.

2.2 Qualification Protocols

2.2.1 Inspection Qualification

The industries which use NDT are typically very conservative, that is they are slow to implement change in the methods they use. This has been highlighted as a major barrier to the introduction of new inspection techniques [1] and it is also a significant barrier to the introduction of new qualification methodologies. There is therefore only a small number of guidance documents

covering technique qualification across the various sectors, such as [2, 3, 4, 5, 6, 7, 8, 9], and anecdotal evidence suggests that many organisations have their own internal qualification procedures. All of these methodologies constitute guidance of some sort as to the type of outputs that are required for qualification; however none provide clear, rigorous work instructions for performing a qualification campaign. Presently, the majority of these methodologies focus solely on the use of experimental trials to demonstrate capability with only a handful describing model assisted approaches. The primary reason for this is the inertia to introducing new methodologies present in industry qualification methodologies has resulted in them failing to catch up with the massive increase in desktop computing capability of the past few decades. The Military Aircraft Structural Airworthiness Advisory Group (MASAAG) Paper 119 [1] highlights the need for a novel methodology that can allow techniques to be qualified quicker and at a lower cost. Over the course of this EngD project, conversations with a range of organisations about qualification procedures have also highlighted the need for a methodology that is more specific than the guidance documents and presents a clear series of steps that have to be performed in order to demonstrate qualification. This is of particular concern to Small and Medium Enterprises (SMEs) who do not have the resources to perform large scale experiment-based qualification trials of their techniques. It would therefore be of significant benefit to these organisations to have a general methodology that is widely accepted that would allow them to demonstrate the capability of their techniques to end users efficiently.

Within the aerospace industry, the most cited source for technique qualification is the United States Department of Defence Military Handbook 1823A Nondestructive Evaluation System Reliability Assessment [4] (henceforth referred to as 1823A for brevity). 1823A provides detailed methodologies for the experimental trial qualification of both hit/miss inspections (a response is recorded as either above or below a threshold) and inspections based on the quantitative assessment of the response (sentencing is based on the magnitude of the response). As a measure of the quality of an inspection, it uses the 90/95 metric which is the value of the parameter of interest which is detected in 90% of measurements in which it is present (a PoD of 90%) with a 95% confidence in that measurement. Crucially, for hit/miss data it recommends a minimum of 60 specimens of defects with a range of values for the parameter of interest and for quantitative response inspections at least 40 specimens with variations of the defect. As well as this, to obtain an accurate measure of the false calls,

when a positive indication is determined when no defect of the critical size is present, three times these respective numbers of defect free specimens need to be inspected. It should be noted that the analysis method used assumes that these are independent measurements thus multiple inspections of the same defect specimen cannot be used as part of the 60 or 40 defect specimens. This is also true of the defect free specimens required for a quantification of the false calls. Thus in total, for hit/miss data this requires a total of 240 specimens and 160 for quantitative response inspections. Whilst it is highlighted that these need not all be distinct samples, a single specimen may contain multiple defects, this is still a significant burden, especially as it is emphasised that these are the minimum values that should be used. It also highlights the need for there to be an assessment of all the variables that will affect an inspection and emphasises the need for experiments to be performed to assess their effects on the inspection. This further increases the burden of qualification by significantly expanding the number of specimens that need to be acquired and the number of inspections to be performed. Discussions with organisations such as Rolls Royce and Airbus reveal that in practice a much smaller number of specimens are used, such as detecting 29 defects in 29 specimens with no false calls which is the minimum number of specimens required to achieve the 90/95 metric. The derivation of this number is not trivial and is the result of the specific method used for calculating the PoD in 1823A. For a detailed derivation the reader is directed to [10] and [11].

However, within 1823A there is a lack of continuity between this requirement and the description of the number of specimens. It is not made clear how having multiple variables, such as material variations and pre-processing of the specimens, affects the number of specimens required. A full understanding of how these variables affect the inspection would require several times more specimens than the figures quoted above, greatly increasing the cost of the process. This is simply not feasible for many applications therefore it is somewhat unsurprising that qualifications often focus on the effect of a single variable. This document also has an appendix on performing Model Assisted Probability of Detection studies, based on work by Thompson [12], however this is again light on practical implementation details and is therefore very difficult to use in anger. A flow diagram of the methodology is shown in Fig. 2.1. Within the text of the methodology, statements such as “Select best available physics-based models that are applicable for the conditions of interest” and “Calculate flaw signal distribution simulations and noise signal distribution simulations” do not

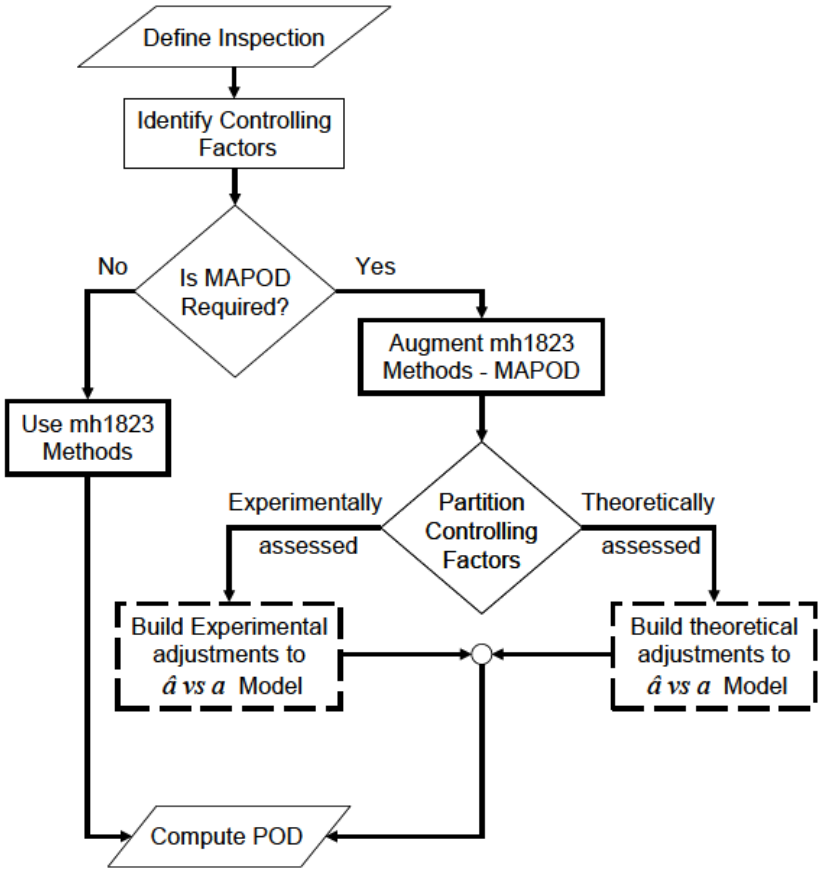

FIGURE 2.1: The flow diagram for the model assisted qualification methodology presented in the United States Department of Defence Military Handbook 1823A Nondestructive Evaluation System Reliability Assessment [4].

provide any insight into how these actions are performed and their various intricacies. No further detail is provided on these actions and the reader is left to their own devices to establish the details of implementing the methodology. The reasoning for the lack of detail is that the methodology is "only a conceptual overview and the details are quite situation-specific". However, as will be discussed in this thesis, there are a wide range of features of a model assisted qualification that are common to all inspections and therefore many of these implementation details can be specified. This lack of detail is the most probable reason why this methodology has not been widely adopted.

An alternative qualification document is the one provided by the nuclear industry body, the European Network for Inspection and Qualification's (ENIQ) guidance on the topic [2]. This document focusses on two main principles: the use of technical justification and the performance of practical trials. The former consists of collating sufficient evidence to demonstrate the capability of the

technique for the desired application which can consist of technical arguments and evidence from previous inspections using the same technique. The latter is again the use of experimental trials to provide evidence of the capability of the technique to the specific application. The document does provide a detailed list of the various stages in a qualification campaign in an appendix to the methodology (specifically Appendix 2) as well as a series of Recommended Practice documents providing further advice on each stage. Within this framework, numerical modelling fits into the technical justification and may be used to provide evidence of the technical suitability of the technique. The ENIQ Recommended Practice 6: The Use of Modelling in Inspection Qualification [13] details the challenges associated with using any numerical model, from in-house code developed specifically for NDT applications to general modelling codes developed commercially, such as finite element codes, which may be applied to NDT scenarios. It suggests six ways of using numerical models as part of the technical justification: "to predict signal amplitudes from postulated defects", "to quantify the influence of parameters related to the inspected component", "to determine the most difficult defects to detect from amongst those in the defect specification", "to interpolate between cases covered by experimental data", "to predict inspection capability for components of similar but slightly different geometry" or "to provide physical insight that can be used further in technical arguments". Crucially, it again does not provide any insight into how these may be performed and this lack of implementation details hinders its use as a model assisted qualification methodology.

A significant challenge with any model assisted qualification methodolgy is to provide validation of the model which essentially requires confirmation of the underlying assumptions and that the model is being operated in its regime of validity. The ENIQ framework addresses these concerns and provides suggestions of how this can achieved, primarily through careful design of experiments against which to compare the model. The focus of these trials on samples with simple features which can be easily characterised however highlights one problem that hinders all qualification campaigns: that is finding samples that are realistic examples of defects that may be present in service. A canonical example of an artificial defect is electrical discharge machining (EDM) notches. These are used to develop and validate inspections, but, they are not necessarily a good representation of defects that will develop in service. Whilst methods have been suggested to overcome this deficiency when discrepancies arise, such as the work of Harding et al. [14] which uses a transfer function to account

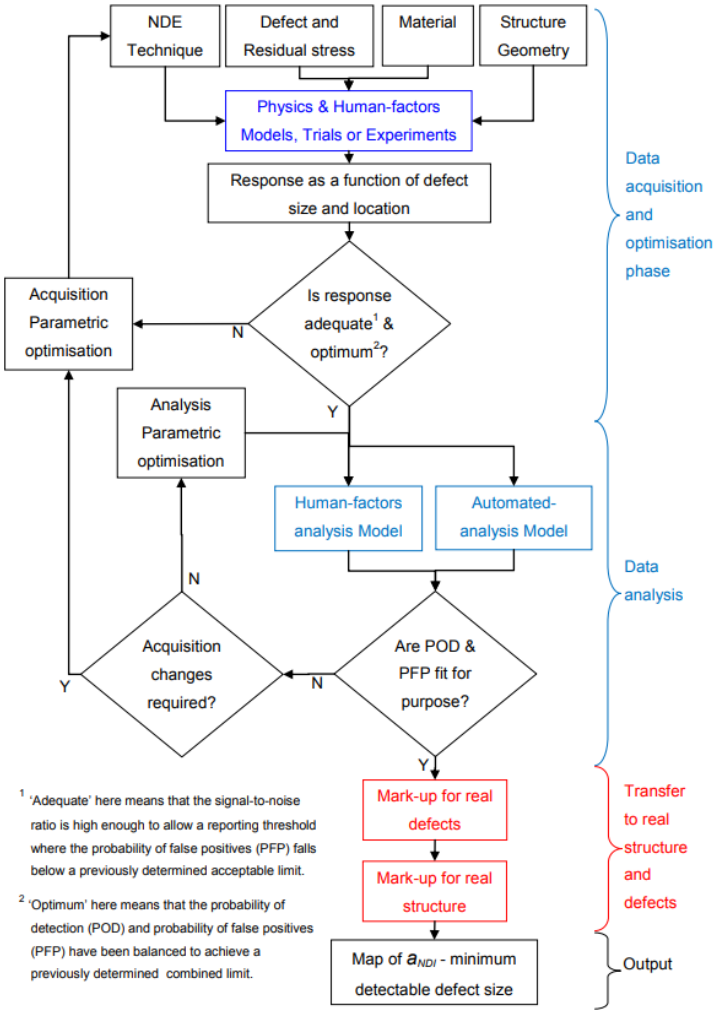

FIGURE 2.2: The flow diagram for the qualification methodology proposed in MASAAG Paper 122 [5].

for the differential between real and artificial defects, it is a problem that has to be considered in any qualification campaign. Modelling approaches have the potential to overcome this burden as it is often easier to model a complex defect than it is to physically reproduce one. Within the context of this project, this is especially important as the MOD has a fleet of ageing aircraft that is no longer in production. Therefore whilst it may be possible to obtain samples from these aircraft, this would require downtime for that aircraft in order to inspect the defect therefore reducing operational capability. Furthermore, knowledge of the true nature of the defect, for example by performing an x-ray CT scan, would be required to accurately characterise the defect before the NDT trial of the new technique can be performed. Conversations early in the project with relevant stakeholders has highlighted that this is a major issue and therefore a method to qualify techniques which minimises the use of defect specimens is desired.

A more general approach to qualification has been suggested by Smith et al. in MASAAG Paper 122 [5]. This outlines a qualification methodology which forms the basis of the protocol being drafted and demonstrated as part of the wider Dstl program by TWI. It presents a clear workflow in the form of a flow chart that someone performing qualification can follow, shown in Fig. 2.2, highlighting the information required to perform each stage and the decision process of continuing onto the next. This clarifies many of the stages and whilst it provides some specifics as to what is required, it still does not provide a complete description of the qualification process.

2.2.2 Personnel Certification

Within the aerospace industry however, as in most industries, there is a greater emphasis on personnel certification rather than on specific technique qualification. The oft-quoted source for such certification is European Standard BS EN 4179:2009 Aerospace Series - Qualification and Approval of Personnel For Non-Destructive Testing (EN 4179) [3]. It outlines the requirements for the various Levels associated with operator certification as well as the duties each Level is certified to perform. Due to this, it is essential to be able to account for the limits of operators and how the certification process affects how inspections are both designed and performed to allow accurate qualification of a technique. The Personnel Certification Network [15] also provides guidance on the subject, giving detailed requirements of the knowledge and practical experience required to be considered certified. The great emphasis on work instructions, documents detailing the specific process of how an inspection is performed which, in theory, covers every variable, should minimise operator variability. However, this can never be truly eliminated and so there will always be some variation which must be accounted for. This does highlight a thorny issue that is yet to be addressed by the NDT community: what is required to consider someone certified to perform inspection qualifications? Presently, this is the subject of debate within the qualification community and a general consensus is yet to be reached. The ENIQ guidance sheds light on this issue, having a specific Recommended Practice on the topic [16]. This provides guidance on how to establish a body capable of qualifying inspections, focussing on the range of skills required and the need for independent adjudicators to provide impartial judgement on the quality of an inspection. If the use of numerical models in qualification, as

hoped, increases, the required knowledge of qualifiers will change and it is possible that many who are currently certified to perform qualifications may no longer be capable of doing so. This therefore becomes a political as well a technical challenge and presents a significant hurdle to introducing regular model assisted qualifications.

2.2.3 Summary

1823A, the ENIQ Framework and MASAAG 122 all demonstrate a lack of detail of how to perform a qualification protocol in anger, making their application to a given inspection difficult. Whilst it is not possible to create a single protocol which covers all eventualities, these guidance documents do have many features in common and it should therefore be possible to write a more general protocol that provides more details on how to implement a qualification campaign primarily using modelling.

The following sections present some examples of practical studies that have been performed to highlight the practical challenges of performing qualification. This is not intended to be an exhaustive search as many trials are performed behind closed doors within companies and therefore the results are not always publicly accessible.

2.3 Examples of Inspection Qualifications

2.3.1 Traditional Qualification Studies

There have been a large number of traditional, experiment based, probability of detection (PoD) trials performed to qualify a technique for use in service. The most common purpose of qualification is to demonstrate a technique's suitability to an inspection scenario, for examples see [17, 18, 19]. The goal of these studies is to generate a PoD curve and from this determine the minimum detectable defect size that can be reliably detected. Another application of PoD trials has been to compare the capabilities of two potential techniques to determine which is most appropriate, as in [20]. In that study, ultrasonic inspections were compared to radiographic inspections to determine which provided the best performance. Explicit comparisons of different modalities are quite rare,

especially in industrial scenarios, as the high cost of the resources necessary to perform two very different inspections is rarely justifiable. A more common use of technique comparison is to demonstrate the advantages of a new inspection over an existing technique, such as in [21]. All of these studies used a significant number of physical specimens to obtain a useful result. However they did demonstrate the capabilities of the various inspections for their respective applications. This highlights both the suitability of using experimental trials and the large associated cost in terms of both time to perform the trials and the expense of obtaining the samples. They also only focus on the investigation of a single parameter of interest whereas it is possible multiple parameters may be of interest, such as the length and height of a crack, which would require multiple sets of samples to be obtained and assessed to obtain a measure of the reliability of the inspection.

There exists, therefore, a need for a qualification methodology that will significantly reduce the effort and expense of performing long trials on physical specimens. The use of numerical models is the prime candidate to achieve this.

2.3.2 Model Assisted Probability of Detection Studies

The idea of using models to replace a significant proportion of experimental trials has existed for several decades and is commonly referred to as Model Assisted Probability of Detection (MAPOD). The idea is to simulate the variations that would be present in an inspection using numerical models. These results can then be assessed using the same methods as experimental trials to assess the reliability of the inspection. This will reduce the significant burden of procuring samples and performing inspections, thus reducing the cost and hopefully the time required to qualify a technique. This idea has gained increased attention over the past decade in particular with computing power now allowing models that are representative of real inspections to be evaluated quickly.

Within the literature, there are two distinct modelling approaches which are used, the transfer function approach (XFN) and the fully modelled approach (FMA). The XFN approach utilises the notion of performing a smaller set of empirical trials on a set of samples and fitting a parametric model to the data. This will allow the data to be interpolated to any intermediary samples and, in theory, extrapolated to specimens with properties that differ in a quantifiable way. There are many examples of this approach being used, such as [22, 23,

24]. The XFN method minimises the computation time however it is still susceptible to many problems associated with traditional PoD trials, in particular the difference between manufactured and realistic samples. A key step in this process is the establishment of the transfer functions which relate in service defects with the manufactured defects and the differences between performing an inspection in the laboratory and in the field. An example of this is that performing an inspection on a runway in the Arctic at the end of a long shift in freezing conditions is very different to performing the same inspection in the comforts of an ideal laboratory. These transfer functions are by no means trivial to establish and there is no consensus either on the form of these functions or on how these should be established. This is demonstrated by Harding, Hugo and Bowles [25]. who observed a very marked discrepancy between the responses obtained from EDM notches and fatigue cracks. This discrepancy was attributed at least partially to crack closure, and it is suggested a further transfer function is required to account for this. It would require significant effort to accurately map the transfer function for a given inspection due to the large variations between different defects in different inspections. Given this, and the number of samples required to accurately establish the transfer function, it is a significant burden to obtain a realistic result. Subsequent work further demonstrated the differences between artificial defects and real defects, including fatigue cracks [14]. The point is raised that it is essential to capture as many realistic features as possible, such as surface finish and complex substructure, as these affect both the defect response and how operators inspect the component, leading to changes in the end result. Work by Demeyer and Jenson [22] provides further evidence of this. Therefore whilst the XFN approach may reduce the number of specimens required compared to a purely experimental trial, it still requires a significant outlay to obtain sufficient results to build an accurate model.

The FMA approach utilises a full numerical model of the inspection, for example using an analytic model or finite element model, and running simulations of the inspections for different samples. This approach has become more common in recent years and examples include [26, 27, 28, 29, 30]. The FMA approach uses a physics-based model to simulate the inspection rather than essentially performing the curve fitting that the XFN approach uses. This also increases the opportunity for optimising the inspection as part of the qualification process. Furthermore, as often the greatest cost of the FMA approach is the development and validation of the model, should the inspection need to be qualified for a similar scenario then a significant proportion of the modelling

work can potentially be re-used, often requiring only a change in inputs rather than the implementation of new modelling techniques. Carboni and Cantini used a numerical model of ultrasonic inspections to validate a rotating ultrasonic probe used to inspect train axles [31]. The modelling was performed using the numerical simulation tool CIVA [Extende, Massy, France] and with a single source of variability, the probe position. The final result was an extremely good match between the experimental PoD and numerical PoD, yielding encouraging results. This was however performed on a small sample of the axle material with artificial defects. Further work on real samples with real defects [32] yielded a significant difference between the simulated PoD and the experimental PoD. This highlights the need for a broader modelling approach which encompasses more of the variables present in performing an inspection in the field. Studies have also been performed to model eddy current inspections [33, 34]. These demonstrate the suitability of the FMA approach to techniques other than ultrasonics and in these studies very good agreement was found between the experimental and calculated PoD curves. However, these again were simple case studies and further work [35, 36], where variations that are likely to be present in real inspections were accounted for, yielded a more significant discrepancy between the experimental and numerical PoD curves. The PICASSO project [37, 38], a large study using CIVA to model ultrasonic weld inspections, accounted for more parameters and had greater success in accurately modelling real inspections, demonstrating the viability of this method. However, this was a study performed over a significant amount of time and thus, for FMA to be used regularly within the MOD, the process needs to be greatly accelerated.

A combination of the FMA and XFN methods has been proposed for a range of inspection techniques, see [39, 40] for examples. These use a combination of experiments to generate expected responses and modelling to aid the development of transfer functions which modify the experimental response. This has the benefit of being able to produce realistic PoD curves based on experimental data however it does not alleviate the need for realistic samples and the time and expense of performing inspection trials. It does however allow some parameters, such as random noise, to be modelled and the results combined with numerical models. This process is often referred to as the modular approach as, where possible, models are isolated if they are independent of one another so that the results may be reused in future qualifications. This is a method of reducing duplicated work over the course of a series of qualifications and is therefore very desirable for practical applications.

The majority of these studies, using both the XFN and FMA methods, follow a statistical analysis methodology akin to that set out in 1823A [4], which was developed to be used for experimental trials. Some of the studies, notably the PICASSO project [37, 38] used Monte Carlo integration to calculate the metrics. The choice of analysis method has an impact on how the model is used, such as what combinations of parameter values have to be evaluated, and the use of modelling presents opportunities for different ways of quantifying the reliability. These analysis methods and their merits are discussed in detail in the next chapter.

The modular approach is the most flexible of the methods discussed and is therefore the approach that will be taken in this project. However, there are some parameters, most notably human factors, that cannot be easily modelled numerically but do have a significant impact on the outcome of the inspection therefore these must be accounted for.

2.4 Human Factors in Inspections

One of the largest sources of variation in performing inspections is the effect of having human operators perform them. This introduces a range of parameters, such as the position of a probe on the surface, the amount of couplant used and device settings to name but a few, that may vary between inspectors and even between inspections performed by the same operator. This leads to an uncertainty in how the inspection will be performed which in turn affects the outcome of the inspection. An accurate quantification and incorporation of these factors into the qualification methodology will be necessary for an accurate result.

The variations caused by human factors can be separated into two categories: those which have an effect which can be quantified as a variable in a model and those which cannot. The former of these include parameters such as the location of a probe in a manual inspection or the effect of incorrectly setting up the equipment. The latter of these are parameters are often psychological, such as the effect of the time of day on the result of sentencing. These can be quantified through experimental trials and included as transfer function models in the calculation of reliability metrics. This is an example of how a modular approach will be very useful when multiple qualifications are performed: these

effects are often independent of other processes in the inspection so can be treated as independent and be reused in other studies.

There is a significant body of literature on the subject of human factors within inspections, see [41, 42] for examples, and the majority of it focusses on the best methods of mitigating their impact. There are a range of methods that can be used to assess the effects of humans on inspections, such as observing inspectors performing inspections [42], performing trials where inspectors assess the same data [43] and performing round robin studies where the same specimens are inspected by different organisations [44, 45, 46]. All of these yield a transfer function that essentially shifts the PoD curve, reducing the reliability of an inspection from its ideal case. This is calculated for a given value β of the parameter of interest x_c as

$$\mathrm{PoD}(x_c = \beta) = \mathrm{PoD}_0(x_c = \beta) - \Delta(\beta), \tag{2.3}$$

where $\mathrm{PoD}_0(x_c = \beta)$ is the untransformed PoD and Δ is the transfer function that quantifies the effect of the human factor. Given that the optimal method of including human factors in reliability assessment would itself be sufficient material for another thesis, in this work they will be accounted for through the use of transfer functions where their effect cannot be quantified in a variable in a numerical model. It should be noted that these parameters will have a reduced impact in automated inspections however humans are always involved in setting up automated inspections therefore they can never be completely ignored.

2.5 The Definition of Reliability

Many qualification methodologies focus on the need to demonstrate the reliability of a technique. However, there is a lack of consensus of what constitutes reliability. Given that a general approach to qualification will require this definition or at least some quantifiable metric that can be used as a measure of reliability, its definition merits further discussion. This section presents a review of the definition of reliability given across a range of guidance documents and qualification studies to highlight the differing opinions (or lack thereof) currently in use.

Perhaps the simplest definition of reliability which is applicable to an inspection is given by the Oxford English Dictionary [47] as “the degree to which

repeated measurements of the same subject under identical conditions yield consistent results". An ideal inspection would occur under ideal, identical conditions every time it is performed however the large number of possible variations that may occur in an inspection mean a broader definition is required. A survey of qualification documents shows that terms such as "sufficient reliability" [6] or "The procedures must be reliable and repeatable" [7] are often written as desired aims of an NDT trial without defining what these statements mean. The following have also been used as quantitative definitions of reliability: Probability of Detection (PoD, also known as the Sensitivity); Probability of False Alarm (PFA, also known as the Specificity); Probability of Acceptance (the probability of true negative); Probability of Rejection (the sum of the PoD and PFA). Of these, the PoD is the most commonly used metric. Carvalho [20] uses PoD as a measure of reliability and assumes that a high PoD constitutes a high reliability. A direct link between the two is made, claiming that "a high PoD and consequently increase in inspection reliability" however there is no discussion or awareness of the implications of this not necessarily being a good measure of reliability. Kurz et al. [19] consider the reliability of NDT methods in general before implicitly making the connection between reliability and PoD. They define the PoD as "the proportion of cracks that will be detected in the total number of existing cracks in a component by an NDE technique when applied by qualified operators to a population of components in a defined working environment". They do highlight the need for a quantification of NDT reliability, especially with the advent of damage tolerant design where the capabilities of the NDT method are integral to the design process.

The Health and Safety Executive (HSE) has taken a sizeable interest in this area and provide guidance on NDT for plant integrity management [8] as well as how to perform NDT measurements on metal structures [6]. The former document defines reliability as "the probability of detection and sizing accuracy". This is again an ambiguous definition as it describes two independent properties - the probability of detecting a defect and the accuracy of sizing that defect - whilst making no reference as to how to combine these two measures. It highlights that PoD or combining PoD with false calls in Receiver Operator Characteristic (ROC) curves can be used to measure a technique's ability to detect flaws. It discusses the importance of human factors and makes the explicit link between variability of human operators (and thus decision makers) with the level of reliability of a technique. Again, however, it makes no suggestion as to how to measure this in either a quantitative or qualitative

manner. It also implicitly makes the link between defect characteristics, such as size, orientation and position, and reliability, discussing how they can impact the "effectiveness of inspection". The latter document also fails to present a definition of reliability, including in the terminology appendix. It does discuss methods of describing reliability, through traditional PoD curves or ROC curves although it makes no judgement as to which of these two is a better measure of reliability or even if the two methods are a good measure of reliability.

The differences between PoD and reliability are highlighted by Rummel [48]. It is suggested that reliability can be described by three independent factors, namely reproducibility, repeatability and capability. These are, respectively, the ability to produce the same result using the same technique on the same specimen many times, the ability to use the same technique on different samples to produce accurate results, and the ability to produce the required discrimination level and/or probability of detection. This further highlights the lack of a single definition of reliability in the context of NDT.

One of the most notable absences of a definition is in MIL-HDBK-1823A [4]. It does not provide a definition of reliability and only notes in its definition of repeatability and reproducibility that "these definitions are not universally agreed on and the usage of "reliability", "repeatability", "reproducibility", "variability" and "capability" are often contradictory". In the presented methodology, the 90/95 metric is used as a quantitative metric of the reliability of a technique and as this is an oft quoted authority on the subject, this measure has become widespread as a definition, especially in the aerospace industry.

The nuclear industry is another area where the qualification of NDT methods is vital. A set of guidelines have been drawn up to qualify inspections [2] however this document makes no specific reference to reliability of a method and places an emphasis on technical justification for qualification, i.e "all evidence on the effectiveness of the inspection, including previous experience of its application, laboratory studies, mathematical modelling, physical reasoning and so on". Even in the expanded glossary there is no definition of reliability. This is itself a rather vague and general description of the qualification process. It is another good example of where reliability is merely glossed over despite it being a key part of NDT inspections and qualifications.

It has been noted that the ENIQ qualification guidelines [2] are somewhat vague and there have been attempts to quantify these guidelines utilising a Bayesian approach [49, 50]. A framework is presented that allows a quantitative

assessment of the reliability however, crucially, it does not attempt to present numerical values for what can be deemed "reliable". This was followed up with another study [51] in which the framework was applied to ferritic welds. In this study the qualification framework was shown to work well however a number of issues were raised with the Bayesian approach. Firstly, the subjective weighting of the various parameters involved in the Bayesian models was noted and the difference in the weightings assigned by different qualification bodies was highlighted. Secondly, the idea of having a single measure of reliability for a technique is dismissed as being essentially impossible as there are too many variables which can affect a measurement. In this regard it is suggested that a solution to this would be to quote the reliability for different types of defects (essentially assigning a reliability to sets of parameters). It is also noted that there is no consideration of sizing errors in the methodology. Nonetheless, there is no mention of what can be regarded as being "reliable" let alone any quantification of a measure of "reliable".

In general, there appears to be an acceptance to jump to the conclusion that obtaining a PoD is a sufficient measure of reliability and specifically that there is a correlation between a higher PoD and greater reliability. Some authors define the reliability as the PoD for the method applied to any scenario however others note that there can be extreme variations between samples and thus PoD should be quoted for different sets of parameters. This evidently leads to a natural variation in what parameters are considered and how these are discretised which makes it difficult to specify a general definition of PoD and thus reliability using PoD. It is evident from the definitions above that there is no single definition of reliability and that it is difficult to define one. Nonetheless, it is necessary to define one based on the need for a comprehensive definition for a qualification technique. The primary challenge in defining reliability is the difficulty in defining the aims of an inspection. Often the need to have a high probability of detecting defects is the main objective, such as in the case of large, expensive equipment such as aircraft and satellites. In this scenario, where missing a defect would be a critical mistake, the number of false positives is significantly less important than missing a real defect therefore the key metric becomes minimising the number of false negatives. Essentially, a high false positive rate is tolerable so long as the false negative rate is minimised. In this case, reliability will be characterised by minimising the false negative rate which is implicitly achieved by maximising the PoD. However, in some scenarios, such as high volume manufacturing of cheap components, a high false call rate may

incur a much greater cost than missing some defects through a lower probability of detection. In aerospace, a false positive can be an expensive mistake if it causes an expensive component to be needlessly scrapped. Within the military air domain, the cost is often downtime of an aircraft, or in some cases the entire fleet of a certain aircraft, which results in reduced capability of the military to perform its role. In this scenario, where it is important to minimise false positives, it may be advantageous to increase the decision threshold to reduce the likelihood of false positives however this comes at the cost of increasing the number of smaller defects that will be missed, assigned as false negatives. In this case, the reliability of the NDT technique will be its ability to minimise false positives whilst detecting not necessarily the smallest cracks. This dovetails with the notion of damage tolerance, the idea of designing components to be able to withstand a certain amount of damage and still be operational. These examples highlight the differences in requirements across inspections and the need for different measures of reliability. Any qualification methodology must therefore be able to generate assessments of reliability based upon the specific requirements of the scenario and not just calculate the probability of detection. Clearly a single reliability metric will be a function of the PoD and PFA however the literature does not provide any guidance as to what this function should be. This is discussed in more detail in Chapter 4.

A common theme in all of these definitions is that the inspection is treated as a single process. In reality, the process of applying a NDT method to a sample to determine if there is a defect with a parameter of interest of a magnitude is a measurement consisting of at least two distinct processes. The first is the ability of the NDT apparatus to detect if there is a defect within a sample, i.e. its ability to obtain a response from the defect. This will include the setting up of equipment and the act of performing a measurement. The second process is the interpretation procedure whereby the response is quantified to determine the size (and/or other characteristics) of the defect(s). This covers both intrinsic interpretation, such as an automated system, or an extrinsic interpretation, such as a human operator where the size of the crack has to be determined before it can be decided whether it is above or below a decision threshold. The accuracy of this sentencing method will have a significant effect on the reliability of any inspection in which sentencing is based on sizing. This links with the modular approach to qualification as, for example, changing the sentencing method, such as employing a new automated decision system, will lead to a change in the reliability of the system however this will have no effect

on the collection of the data. Therefore it would be significantly more efficient to quantify the reliability of only the independent processes that have changed rather than the entire system.

2.6 Summary

This review has highlighted that there is a range of guidance across a range of industries on how to qualify or demonstrate the reliability of a technique. These are all rather vague on how these can be implemented in practice and miss out many of the salient details required to do this. The focus of all of them, in one way or another, is to quantify the effect of variations in how an inspection is performed on the outcome of the inspection and thus infer how reliable the inspection is through some quantified metric. This project aims to develop a methodology that fills in many of these details so that it can readily be used in anger. This thesis will outline this protocol and the writing of a document appropriate for use in industry is being undertaken by TWI on behalf of Dstl, into which this thesis will feed. This review has also demonstrated that there is no single consistently used definition of reliability thus there is a need in any proposed methodology to be able to accommodate any definition that is required. The following chapter discusses the calculation of common reliability metrics in detail and presents the outline of a general methodology to quantify reliability.

Chapter 3

A General Theory of Qualification

The previous chapter discussed the features and limitations of contemporary qualification protocols. This chapter presents a general approach to performing a qualification campaign. It also reviews methods of calculating inspection metrics and presents a general formulation of calculating probability of detection and probability of false alarm, discussing both the theory and practical computational considerations.

3.1 A General Methodology for Assessing the Reliability of an Inspection Primarily Using Numerical Models

The qualification methodologies discussed in the previous chapter all have a large number of similarities. They all, with varying degrees of clarity and brevity, instruct the user to attempt to quantify the variations that are inherent in an inspection and to demonstrate that the technique can achieve a desired reliability. These can all be distilled into a general process which describes the stages of demonstrating the reliability of a technique. The following is the outline of such a general methodology that can be applied to any inspection modality.

The first requirement of a qualification is to determine the objectives of the process, essentially determining what is to be demonstrated. The most common requirement is to demonstrate that a minimum probability of detection will be

achieved for a defect characteristic of a certain value, for example finding a crack of length 5 mm in 90% of inspections. Another common requirement is to demonstrate a false call rate below a desired level. The qualification process however can also be used as an optimisation tool therefore the objective may be to determine the optimal inspection configuration as well as demonstrating a desired level of reliability.

Once the objective has been established, the parameters that vary in the inspection, such as the defect parameters and the parameters that a human has an influence over amongst others, need to be determined. These have an effect on the choice of the numerical model as ideally it should be able to incorporate the effect of changing each of these parameters. Of these parameters, those that are better suited to having a transfer function model derived from either experimental measurement, such as electrical noise, or a separate numerical model should be partitioned. The parameters better suited to the transfer function method are those whose effect is known to be independent (or a set of parameters whose effect as a collective is independent) of the other parameters. The relative cost, both time and resources, of incorporating these parameters into an existing numerical model or performing the necessary experiments to build a transfer function can be weighed to determine whether an experimental or numerical model is the most efficient. Incorporating these models into the calculation of the metrics is discussed in more detail later. Once these, and any other, requirements on the model are determined, an appropriate numerical model can be chosen. This should ideally already be experimentally validated so that it is demonstrably an accurate representation of the inspection, however further experimental validation may be required to satisfy the requirements of the qualifying organisation or a qualification body. A common approach to validation is to use the corner cases, that is the extreme values of the parameters, and if these are shown to be valid then it is generally a reasonable assumption that the model is valid across the entire parameter space.

How this model is used will depend upon the method that will be used to calculate the inspection metrics. These are discussed in more detail in the next section however they all involve generating some numerical data that represent outcomes of the inspection. The model should be evaluated at the required sample points to generate the requisite data. Once this has been completed, these results, alongside any transfer function models created, are used to assess the capability of the technique. Depending on the method of calculating the inspections metrics used, the information generated may be sufficient to provide

significant insight into how the parameters affect the inspection and thus can be used for optimisation of the inspection. If this is used, the reliability of the inspection may have to be reassessed and this process repeated until a suitable inspection is defined.

This process is the outline of the methodology that is being written as part of the wider Dstl project. Given that a major shortcoming of many contemporary qualification methodologies is a lack of specific implementation details, this thesis goes into significant detail on many aspects of this process to provide the finer details necessary to implement this process in anger. The hope is that this will be sufficient, alongside the written protocol, to use this methodology regularly to qualify techniques.

3.2 Contemporary Inspection Metric Calculations

The choice of method for calculating inspection metrics determines what models must be evaluated, therefore these must be understood in order to determine the optimal way of using models to demonstrate the reliability of the technique. The calculation of these metrics requires the combination of information about the possible outcomes of an inspection and the probabilities of these responses occurring. There are two main types of data which can be used to calculate metrics of an inspection: hit/miss data and response data. The former uses information of simply whether a detection has been made, that is whether the inspection has resulted in a positive or negative sentencing without requiring any quantitative characterisation of the response. The latter uses a quantified response of the inspection, such as the amplitude of an ultrasonic signal, and knowledge of a threshold on this response. Both of these require some knowledge, or assumptions, of probability of the variations in the outcome or response of the inspection. The latter is more general and is to be focussed on in this project. The reader is directed to [52] for a good description of the former method.

There are three main methods that are used presently to calculate metrics of inspection capability using response data: the $\hat{a}$ vs a method, the extension to this known as the Multi-parameter PoD model and Monte Carlo methods. Of these, the first method is the most commonly used with the second method

seeing only limited application. The third method is more suited to model assisted approaches and is rarely, if ever, used for experimental trials but has seen some use for numerical studies. This section discusses these three methods and their limitations.

3.2.1 $\hat{a}$ vs a

The $\hat{a}$ vs a method is presently the most common method used to calculate metrics such as PoD and PFA. It is referenced as the method to use in a range of qualification methodologies including the United States Department of Defence Military Handbook 1823A Nondestructive Evaluation System Reliability Assessment [4] and the European Network for Inspection and Qualification's (ENIQ) guidance on the topic [2]. The following derivation is based on the description of the method in [52]. It involves fitting a linear function to the response of the inspection, $\hat{a}$, as a function of the parameter of interest a, that is

$$\hat{a} = \alpha + \beta a + \delta, \tag{3.1}$$

where α and β are parameters that are estimated by performing a fit to the data and δ is a noise term. The noise term δ is assumed to be a normal distribution with zero mean and constant standard deviation σ for all values of a. The value of σ can be empirically determined from experimental data. The determination of α and β is typically performed using Maximum Likelihood Estimation (MLE). The idea of MLE is to find the parameters of the function that maximise the likelihood of the measurements being observed. For a linear function, using the linear least squares method (strictly using vertical rather than perpendicular offsets) with the assumption of constant normal variance is equivalent to MLE. Linear least squares can be calculated by expressing the problem in a linear matrix form as

$$\hat{a} = AB, \tag{3.2}$$

where A is the a $n \times 2$ matrix of input variable values for the n measurements performed with each ith row being $[1, a_i]$, and $B = [\alpha, \beta]^T$. B can be estimated by finding the minimum value of the Euclidean 2-norm given by

$$\min||\hat{a} - AB||^2. \tag{3.3}$$

In practice, this process can be highly optimised and the reader should see the LAPACK documentation for the DGELSD function [53] for more details.

The definition of PoD for a given value of the parameter of interest a is the probability that the response is greater than the response decision threshold, $\hat{a_{th}}$, that is

$$P(\hat{a} \geq \hat{a_{th}}) = 1 - \Phi\left(\frac{\hat{a_{th}} - \hat{a}}{\sigma}\right), \tag{3.4}$$

where Φ is the normal cumulative distribution function. Using the symmetry of the distribution and substituting in Eqn. 3.1, the PoD for a given value of a can be calculated as

$$PoD(a) = \Phi\left[\frac{\alpha + \beta a - \hat{a_{th}}}{\sigma}\right]. \tag{3.5}$$

This process however makes the following assumptions:

1. There is constant variance throughout the parameter range.
2. The variance is normally distributed about the mean value.
3. The response $\hat{a}$ is linearly proportional to the defect characteristic a.
4. There are no saturated responses (high or low).

The first assumption is specific to the MLE method used to estimate the parameters α and β and can sometimes be mitigated through the use of suitable transforms, such as taking logs [54]. However this may result in a change in the variance of the response throughout the range of a and thus invalidate assumptions 2 and 3. If 2 and 3 are not valid in the first instance then a transform may yield a data set that satisfies these assumptions however the resulting function may no longer be linear. Assumption 4 can only be overcome by truncating the data which will reduce the range of a over which the PoD is calculated, potentially reducing the usefulness of the resutling PoD curve. Another limitation of this process is that it can only calculate metrics for one parameter at a time and cannot investigate the interactions between parameters. Given that these assumptions are inherent to this method of calculating the PoD, if any of these assumptions are not true in a given scenario then the analysis will result in a misleading answer. These limitations are well documented and the relevant appendix of the Military Handbook 1823A [4] carries the following warning:

"If any of these assumptions is false, or, if the model is a line and the data describe a curve, then the subsequent analysis will be wrong. You may be able to coerce the software into producing POD plots, but they will be wrong. This is true of any analysis software (finite element codes for example) - If the input

is flawed the output will be wrong. Input includes the assumptions on which the analysis is based, not just the input data. Thus it is prudent practice - in statistics and in engineering - to state all analysis assumptions explicitly so that the customer can evaluate their relevance and veracity."

The error in this calculation is due to the error in the linear fit and the error in the calculation of σ. The error in the former arises from the error in the estimation of the parameters a and b. For a derivation of these errors the reader should see [55] and the salient result is that the errors in these quantities scale as $N^{-\frac{1}{2}}$ for N independent measurements. The error in the latter ϵ_σ is given by [56]

$$\frac{\epsilon_\sigma}{\sigma^2} = \sqrt{\frac{2}{N-1}}, \tag{3.6}$$

again the key point being that it scales as $N^{-\frac{1}{2}}$. Therefore it becomes increasingly costly to reduce the error in these predictions. Whilst it may be possible to obtain a more precise result with an increased number of specimens, it will only be accurate if the assumptions underpinning the method are valid. Therefore results created using this method should not be taken at face value and evidence of the satisfaction of the assumptions should be presented alongside to provide confidence in the result.

3.2.2 Multi-Parameter PoD Model

A proposed extension of the $\hat{a}$ vs a approach to account for multiple influencing parameters is the Multi-Parameter PoD model [57]. This alleviates this last limitation by casting the response as a function of multiple influencing parameters, that is $a_{MP} = f(a_1, a_2, ..., a_n)$. In this case the response is now written as a linear function of this parameter, that is

$$\hat{a} = \alpha + \beta a_{MP} + \delta, \tag{3.7}$$

and the same analysis method is used as in the $\hat{a}$ vs a method. However this is somewhat of an obvious step as the outcome of an inspection most often depends on multiple parameters. In the $\hat{a}$ vs a method described above, the effect of influencing parameters other than the parameter of interest is incorporated into the noise term and careful design of experiments is necessary when selecting specimens to account for these variations. If the inspection had only a single influential parameter then the PoD curve would be a step function as the PoD

would be zero when a is such that $\hat{a}(a) < \hat{a_{th}}$ and one when $\hat{a}(a) \geq \hat{a_{th}}$. This perhaps explains why this method has not achieved widespread usage as its only advantage over traditional $\hat{a}$ vs a is to reinforce the notion that the PoD does depend on multiple influencing parameters. It is also noted that due to the number of measurements required to obtain a good number of data points to perform this analysis, it is better suited to numerical model based trials. Similarly to the $\hat{a}$ vs a approach, it is still limited by the assumptions of MLE and is therefore not always appropriate.

3.2.3 Monte Carlo Calculation

A different approach, which makes no assumptions as to the nature of the probability distributions nor involves any form of curve fitting, is to use Monte Carlo integration to directly calculate the metrics. Monte Carlo methods are used in a wide range of fields beyond NDT, for examples of its application to NDT reliability studies see [37, 38, 26]. The idea of Monte Carlo methods in metric calculations is to sample the input values for the parameters of a model from known probability density functions and then classify the response of the model as either a pass or fail. In the case of PoD calculation, the classification is it either being above or below the response decision threshold. This process is repeated many times and the ratio of the number of responses obtained that are above the threshold to the number of responses evaluated equals the PoD, that is

$$\text{PoD} = \frac{\text{Number of responses} \geq T | \text{defect present}}{\text{Number of models evaluated} | \text{defect present}}. \tag{3.8}$$

The ratio of the responses below the threshold to the number of models evaluated yields the probability of false negative (PFN), equivalent to 1-PoD. As the number of model evaluations increases, the values of PoD and PFN converge to their true values. The benefit of the Monte Carlo approach is that it makes no assumptions as to the variance of the response, its linearity and can calculate metrics for several parameters simultaneously. This method does have two main disadvantages, that it can take a large number of model evaluations to determine the metrics and that changing any of the probability distribution functions requires the whole process to be repeated. The error in the calculation scales as $N^{-\frac{1}{2}}$ [58] for simple Monte Carlo integration schemes, that is it becomes increasingly costly to reduce the error in the calculation as the number of model evaluations increases. It is possible to improve this rate of convergence through the use of quasi-random sequences such as Sobol sequences [59, 60] or Halton

sequences [61]. The latter factor is the most significant hindrance as the models can potentially have a long evaluation time therefore incurring a significant time cost to recalculate the metrics should the probability distributions change. Ideally, the number of models that have to be evaluated should be minimised and the probability distributions should be independent of the choice of model evaluations so that they can be arbitrarily modified without having to perform further model evaluations.

3.3 Beyond $\hat{a}$ vs a

The limitations of current qualification methods highlight the need for a more general formulation of the calculation of inspection metrics and this section describes such an approach. Consider an inspection with n parameters that may vary, $x_1, x_2, ..., x_n$. Define the parameter space of possible variations which may occur in the inspection as the n dimensional unit hypercube

$$\Omega = [0, 1]^n \tag{3.9}$$

where the values of all parameters have been scaled to the interval $[0, 1]$. This step is important as different parameters can have scales that vary by several orders of magnitude and when performing multi-dimensional interpolation and sensitivity analysis this normalisation step ensures that each parameter has equal importance. The response of an inspection at a coordinate $\mathbf{x}$ in Ω is given by the function $R(\mathbf{x})$. The probability of a coordinate $\mathbf{x}$ occurring when an inspection is performed is given by the probability function $P(\mathbf{x})$. The determination of this function is discussed in more detail later.

Combining the information in $R(\mathbf{x})$ and $P(\mathbf{x})$ allows quantitative metrics of the capability of an inspection to be derived. The definition of PoD and PFA require two values to be specified to classify a response: a decision threshold on the response, T, and a decision threshold α on a parameter of interest x_c. The former indicates when a detection has been made and the latter determines whether the response is from a defect with a significant value of the parameter of interest. These values will also be determined from the definition of the inspection, for example T may be an amplitude threshold which determines whether a component is passed or failed and α may be a physical defect size above which a component is removed from service.

The PoD for a defect with a given characteristic parameter x_c equal to a specific value β is the probability that the response of the inspection is greater than the response decision threshold, T, given that a defect of that magnitude is present, that is

$$PoD(x_c = \beta) = p(R(\Omega) \geq T | x_c = \beta). \tag{3.10}$$

This can be calculated from analytic definitions of the response function $R(\mathbf{x})$ and the probability function $P(\mathbf{x})$ as

$$PoD(x_c = \beta) = \frac{\int_\zeta P(\mathbf{x}) d\mathbf{x}}{\int_\omega P(\mathbf{x}) d\mathbf{x}}, \tag{3.11}$$

where

$$\zeta = \Omega | (x_c = \beta, R(\Omega) \geq T), \tag{3.12}$$

and

$$\omega = \Omega | (x_c = \beta). \tag{3.13}$$

The PoD can be computed by using a discrete formulation of Eqn. 3.11, effectively performing numerical integration.

The PFA is the probability that the response of the inspection is greater than the decision threshold given that the critical parameter is less than its threshold α, $p(R(\Omega) \geq T | x_c < \alpha)$. The PFA may be calculated as

$$PFA(x_c < \alpha) = \frac{\int_\kappa P(\mathbf{x}) d\mathbf{x}}{\int_\xi P(\mathbf{x}) d\mathbf{x}}, \tag{3.14}$$

In the most general sense, a false call is an indication in an inspection which would cause the inspector to take an action which would otherwise not be taken if the true nature of the defect is known. In this case, the definition of a false call is

$$\kappa = \Omega | (x_c < \alpha, R(\Omega) \geq T), \tag{3.15}$$

and

$$\xi = \Omega | (x_c < \alpha). \tag{3.16}$$

Some organisations categorise false calls into two categories. The definition given by Eqn. 3.15 and 3.16 is designated as an overcall if a defect is present, that is a detection of a defect is made however it is determined to have a greater

magnitude than it has in reality. The other category is a false alarm in which no defect of the type of interest is present, that is $x_c = 0$, and in this case the integral region is

$$\kappa = \Omega|(x_c = 0, R(\Omega) \geq T), \tag{3.17}$$

and

$$\xi = \Omega|(x_c = 0). \tag{3.18}$$

In this work, the definition given by Eqn. 3.15 and 3.16 is used. These formulations may also be extended to cover multiple conditional probabilities, such as the probability that a crack has a certain length and height. This is achieved by redefining the integral regions to cover this condition. This is a significant advantage of this method over the $\hat{a}$ vs a method in that rather than just a threshold on the response, any sentencing method which can be expressed as a condition of the parameters and the response can be used for the integral regions. Similarly, the response can be converted into binary hit/miss data or some other quantified metric if a complex sentencing method is used and the integrals can still be performed on the resulting data to evaluate the reliability of the technique.

As these calculations are independent of the model evaluations, a single set of response data may be used to generate multiple reliability assessments using different metrics, a significant advantage over the $\hat{a}$ vs a method. Furthermore, it is possible to change the probability distributions without having to perform any further model evaluations whereas the Monte Carlo algorithm would have to be performed again thus incurring a significant number of needless model evaluations. The general method presented here is therefore a more efficient method of calculating metrics if the response function can be accurately mapped in a reasonable time, of at least the order of the time for Monte Carlo methods. These calculations can also be computed directly using Monte Carlo methods which would not require re-evaluation of models if the probability function is changed, although further model evaluations may be required to obtain the desired level of accuracy. The optimal way of mapping the response function is addressed in the next chapter. The next section presents a comparison of these methods for a simple example inspection.

3.4 Definition of Example Response Functions

This section defines two analytic functions that are used throughout this thesis as example response functions for the comparison of methods. Given the desire for generality of the methods presented in this thesis, these functions are not designed to be representative of the physical response function of a particular inspection. However, they do possess properties which may arise.

The first is a purely additive function of the four parameters x_0, x_1, x_2, x_3 and is defined over the four dimension unit hypercube $\Omega = [0, 1]^4$. It is defined as

$$R(\Omega) = ax_0 + bx_1 + cx_2 + dx_3, \tag{3.19}$$

where the parameters a, b, c and d are constant coefficients. This function is linear in all of the four variables, it is monotonically increasing and the parameters are all independent, that is the response function can be expressed as a sum of functions of single variables.

The second equation is a more complicated function of three parameters x_0, x_1, x_2 and is defined over the three dimension unit hypercube $[0, 1]^3$. The function is defined as

$$\begin{aligned} R(\mathbf{x}) &= 1 + (\frac{3x_0}{2})^3 + x_0x_1 + 2\sqrt{x_1 + x_2}, \\ R(\mathbf{x}) &< 2.0 = 2.0, \\ R(\mathbf{x}) &> 5.5 = 5.5. \end{aligned} \tag{3.20}$$

This function is not monotonically increasing, it has interaction between the parameters and saturation of the response is present. This is chosen as some or all of these features may be present in a real response function.

3.5 Comparison of Calculation Methods

The effect of using these different calculation methods can be demonstrated using a simple numerical function. Consider the response function given by Eqn. 3.20. A decision threshold on the response of 3.75 is used to distinguish a detection. The parameter x_0 is given a uniform distribution, x_1 is given a normal distribution with a mean of 0.3 and a standard deviation of 0.2, and x_2 is given a normal distribution with a mean of 0.5 and a standard deviation of

$0.1+\frac{x_1}{2}$. The PoD is calculated as a function of the parameter x_0, that is $a = x_0$. The empirical data required for applying the $\hat{a}$ vs a (in this case effectively $R(\mathbf{x})$ vs x_0) method is generated by randomly sampling 31 points for 31 values of x_0, a total of 961 samples which is significantly greater than would be used in an experimental PoD trial.

Figure 3.1(a) shows that this a non-linear function with non-constant variance across the range of x_0. This function was chosen specifically to not satisfy the assumptions of the $\hat{a}$ vs a method, as it is non-linear, has non constant variance that is not normally distributed and has saturation. Therefore this does not satisfy the assumptions of the $\hat{a}$ vs a method. The cubic dependence of the response on x_0 in Eqn. 3.20 would suggest that a natural logarithm would be an appropriate transform therefore this has been applied to attempt to linearise the function. The result of this is shown in Fig. 3.1(b) and is still not a linear function. A more robust method would be to derive the linearising transform explicitly from Eqn. 3.20 however in practice this function is very unlikely to be known making this process nigh on impossible. The result of this transformation still does not satisfy the assumptions of the $\hat{a}$ vs a method although for illustrative purposes the PoD is calculated for both data sets. The raw data appears to result in a better linear fit, validated by having an R^2 value of 0.79 compared to 0.53 for the transformed data. The transformation also has a significant effect on the variance of the function as the transformed data has a much larger range of variances across x_0 compared to the original data. The saturation of data is also an important factor and this has an effect on both the linear fit and on the calculation of the PoD as it assumes a continuous error distribution which saturation invalidates. An average standard deviation can be calculated across the range of x_0 and is used as the σ required to calculate the PoD using Eqn. 3.5.

The curve for the general method is calculated using full knowledge of the response function. This allows this curve to be calculated using direct numerical integration, specifically using the Fortran QUADPACK library [62] which is wrapped in SciPy. The reader is directed to the documentation in [62] for more details of integration methods. These methods are susceptible to rounding error in the calculation however as a relative fraction of the value of the integral they are of at worst 10^{-10}. Given this accuracy, no error bars are plotted on these curves. In practice, this method requires accurate mapping of the response function and the error in the calculation is dependent on the accuracy of the mapping. Methods of achieving this are discussed in detail in

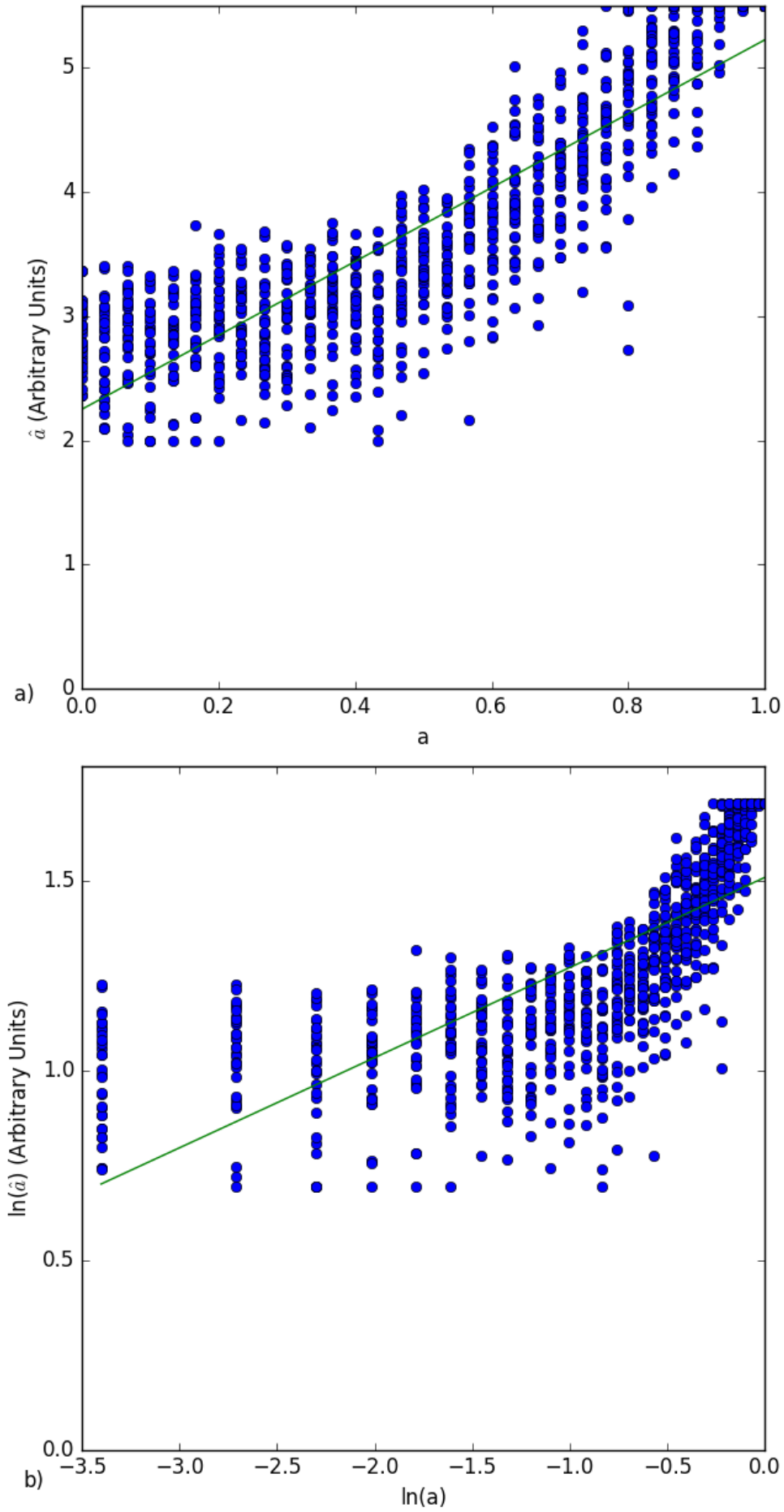

FIGURE 3.1: The response of the function plotted (a) linearly and (b) with a log-log transform applied.

the next Chapter.

The Monte Carlo method should give the same result as the general method and this is tested by taking the same number of randomly sampled points, 31, for each value of x_0. This process was repeated twenty times and an average error in the calculation was determined, shown by error bars on the PoD curve. The PoD curves for each of the three methods are shown in Fig. 3.2. This shows that the raw and transformed data result in very different PoD curves which differ significantly from the result of the Monte Carlo and general calculation method. As expected, the Monte Carlo and the general method results match to the error of the Monte Carlo method. The significant overestimation of the PoD in the middle of the x_0 range highlights the shortcomings of the $\hat{a}$ vs a method as it may yield an inaccurate representation of the capability of a technique, in this case suggesting the technique is more capable than it actually is. The effect of calculating the PoD using a smaller number of samples is demonstrated in Fig. 3.3 in which only 3 values at each value of a_0 were used and the error in the prediction calculated by repeating the calculation of the PoD curves 10 times. This shows that the $\hat{a}$ vs a requires fewer samples to obtain a more precise result than the Monte Carlo method however it has no effect on the accuracy if the underlying assumptions are not satisfied. The greater error of the Monte Carlo method for this example is not surprising given that Monte Carlo methods are known to be inefficient for low dimensionality integrals however it does still yield a more accurate result than the $\hat{a}$ vs a method. It is also interesting to observe that the error is largest for both methods in the mid range of x_0 and the Monte Carlo method has a much lower error at small and large values of x_0 than the $\hat{a}$ vs a calculation. In the Monte Carlo method, as the sampled responses at these values of x_0 will always be below or above the response threshold respectively, little if any error is introduced and a response will not be incorrectly classed as moving over the threshold. In contrast, the error in the linear fit of the $\hat{a}$ vs a calculation causes a shift in the mean value of the response at a given value of x_0 and hence introduces a change in the PoD at that value. In the mid range of x_0, the converse is true and the $\hat{a}$ vs a method provides a more precise, although still inaccurate, measure of the PoD. The small number of samples used in the Monte Carlo method does not provide an accurate representation of the PoD. Effectively, the use of 3 samples allows only four possible values of the PoD at any given value of x_0: [0.0, 0.33, 0.67, 1.0]. This results in the large error bars which are the effect of averaging over this large range of values.

The large discrepancy in shape demonstrates that it is evidently important

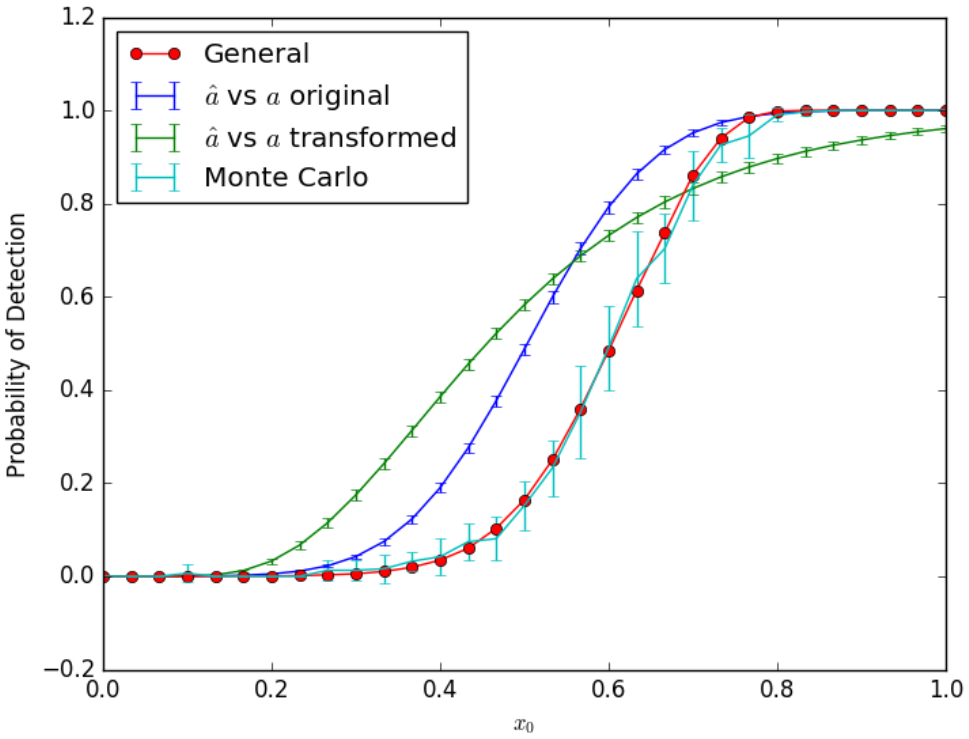

FIGURE 3.2: The PoD curves calculated using different analysis methods of the same data, using 31 points at each value of x_0.

that the assumptions of the $\hat{a}$ vs a method are satisfied otherwise a very different result may be obtained. In this case, whilst it is possible to analytically obtain a linear relationship between the response and defect size, clearly the variance of the response is neither constant nor normally distributed. The presence of saturation in the response is also having an effect on the calculation. In real scenarios, it is unlikely that all of these assumptions will be met therefore the results obtained are likely to be misleading. This is a strong motivation for the need for a more general method that moves away from these assumptions.

3.6 Incorporation of Transfer Function Models

The primary issue with the general approach is that as the number of parameters increases, the RAM required increases to a power law and the available memory will be quickly exhausted. A typical desktop computer has 16 GB of RAM and a 9 parameter response space with 10 values each would require 8 GB of RAM to store as 64 bit double precision floats thus it is not practical to store large parameter spaces. It is therefore desirable to split the parameter space into smaller independent spaces which can be mapped separately. This

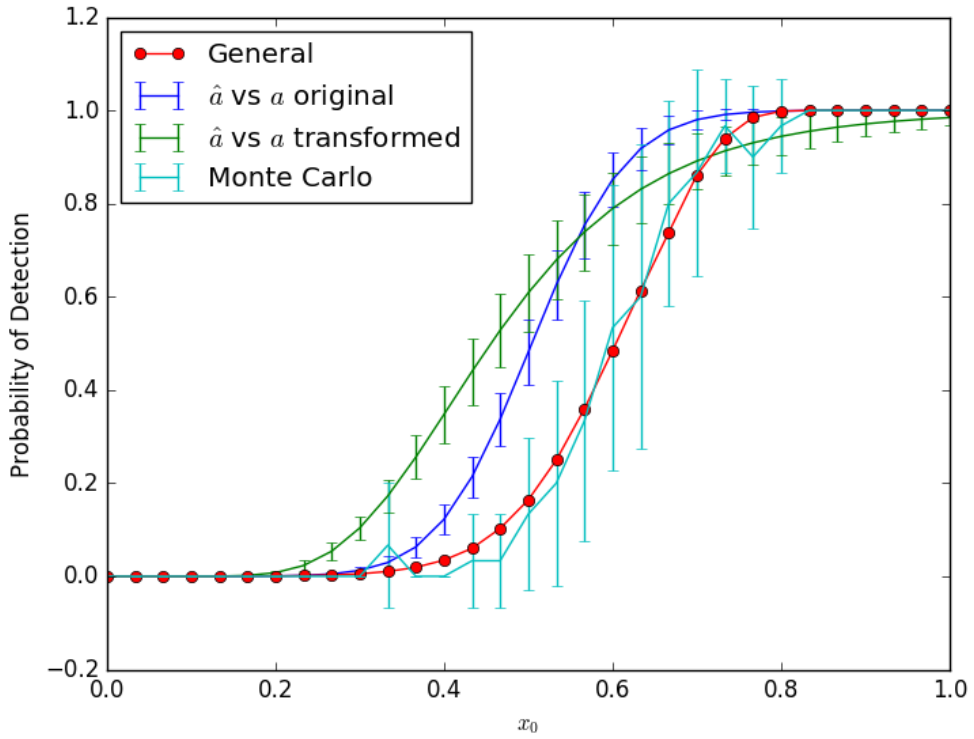

FIGURE 3.3: The PoD curves calculated using different analysis methods of the same data, using 3 points at each value of x_0.

would reduce the effective size of the parameter space with no loss of accuracy if the parameters are independent. There are also some independent parameters, such as ultrasonic couplant thickness or electrical noise, that are better suited to generating a model from experimental data rather than using a numerical model. In the literature this is referred to as the transfer function approach. Therefore a method is needed which allows these adjoint spaces and the transfer function models to be used in conjunction with numerically modelled data. An alternative method is to directly incorporate these parameters into calculation of metrics such as PoD and PFA as independent parameters or sets thereof. This is suitable for parameters which are known to be independent of all other parameters, such as electrical noise and coupling thickness. Consider two parameter spaces Ω and Ψ which have response functions, $R_\Omega(\Omega)$ and $R_\Psi(\Psi)$, and probability functions $P_\Omega(\Omega)$ and $P_\Psi(\Psi)$, respectively. They are independent if $R(\Omega,\Psi) = f(R_\Omega(\Omega), R_\Psi(\Psi))$ and $P(\Omega,\Psi) = P_\Omega(\Omega)P_\Psi(\Psi)$. These parameter spaces can then be combined by calculating a modified version of Eqn. 3.11 as

$$PoD(x_c = \alpha) = \frac{\int_\zeta (P_\Omega(\mathbf{x}) \int_\Delta P_\Psi(\mathbf{y}) d\mathbf{y}) d\mathbf{x}}{\int_\omega P_\Omega(\mathbf{x}) d\mathbf{x}}, \tag{3.21}$$

where $\mathbf{x}$ and $\mathbf{y}$ are dummy variables for the parameter spaces Ω and Ψ respectively. The choice of the integral region is dependent upon whether the response of the two spaces is additive or multiplicative. In the former case, the region is defined as

$$\Delta = \Psi|(R_\Omega(\Omega) + R_\Psi(\Psi) \geq T), \tag{3.22}$$

whereas in the latter the region is defined as

$$\Delta = \Psi|(R_\Omega(\Omega)R_\Psi(\Psi) \geq T). \tag{3.23}$$

Multiple independent parameters can be included in this method by combining individual functions into a single function and applying the formulas above. This method provides a significantly more computationally efficient method of including these parameters.

For N numerically modelled parameters and M independent parameters, this has the effect of reducing the memory required for including these parameters from $2NM$ to $2(N + M)$. The computational efficiency can be further improved by pre-computing a cumulative distribution function of the independent parameters which uses a minimal amount of memory. This approach can also be applied to the PFP and other metrics using a similar formulation. Discrete formulations of these metrics can be formed which allow these metrics to be calculated efficiently computationally.

3.7 Definition of Probability Function

The $\hat{a}$ vs a makes implicit assumptions about the probability of variations occurring, specifically that they are normally distributed with constant variance. The Monte Carlo method and the general method both, however, require the probability distributions to be explicitly defined. This can either be measured experimentally, although would require very costly observation of operators performing trials, or estimated from engineering judgement. For example, structural mechanics may predict a distribution of crack rotations whilst observing operators performing the inspection will give the probability distributions of the human controlled factors. In general, it is assumed that resources expended on training operators are effective and that they are capable of performing a better inspection than just arbitrarily guessing, therefore the probability function will be non-uniform and peaked around the optimal set of

inspection parameters. The worst case would be that there is no knowledge of likelihood of variations occurring in which case uniform distributions can be used for all parameters. Limits also need to be established for the parameters which will be derived from considering how an inspection is performed. For example, if the position of a probe is a parameter of an inspection, then it will be limited to, most conservatively, the probe being on the specimen. In practice, this may be further limited to a smaller area in a region of interest on a specimen where defects are known to potentially occur and where the inspector has been trained and instructed to look, such as around a prominent feature. If each of the n parameters are independent, $P(\mathbf{x})$ is by definition

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i), \tag{3.24}$$

where $P(x_i)$ is the probability of variations occurring in the i^{th} parameter. It is plausible that variations in parameters may not be independent and that probability functions based on the interactions will need to be established. An example of this is the optimisation of the position of a probe in two axes. In order to find the optimal position it may be required to optimise the positions of both simultaneously. In this case the probability function will not be the product of two probability functions, $P(\Omega) = P(x_1)P(x_2)$, but rather a complex function of the two variables $P(\Omega) = P(x_1, x_2)$. In order to calculate metrics of inspection capability such as PoD numerically, the probability function needs to be discretised into probability voxels P_v to perform numerical integration. Each P_v gives the probability that the inspection will occur in that voxel in the parameter space, i.e. it is the integral of the probability density function over the voxel. In the case that the functions are independent, the voxel value P_v can be calculated exactly over the range as

$$P_v = \prod_{i=1}^{n} C_i(x_{i,v} + \frac{\Delta_i}{2}) - C_i(x_{i,v} - \frac{\Delta_i}{2}), \tag{3.25}$$

where C_i is the cumulative distribution function for the parameter x_i, $x_{i,v}$ is the value at the centre of the voxel and Δ_i is the voxel width. This method results in exact calculation of the discretised probabilities throughout Ω. In the case where the probability function is a more complex function, a probability voxel can be calculated as

$$P_v = \int_{\mathbf{x}_c - \frac{\Delta \mathbf{x}}{2}}^{\mathbf{x}_c + \frac{\Delta \mathbf{x}}{2}} P(\mathbf{x}_v) d\mathbf{x} \tag{3.26}$$

This can be approximated numerically as

$$P_v \approx P(\mathbf{x}_v) \prod_{i=1}^{n} \Delta x_i. \tag{3.27}$$

Attempts have been made to define these probability distributions in other projects, for example the PICASSO program [37]. Two examples of estimating distributions are shown in Fig. 3.4 and Fig. 3.5. The former shows the centre frequency and bandwidth of a sample of ultrasonic probes provided to the current author by Olympus, advertised as 5 MHz probes. It is evident that there is a range of bandwidths and centre frequencies, none of which are at 5 MHz. This information can be used to generate a probability distribution of the frequency spectrum of ultrasonic probes. The latter figure is an example of measuring a probability distribution experimentally. The amplitude of the reflection from a fixed back wall of a block of aluminium was measured repeatedly by the current author, controlled such that the only variation between measurements is the application of coupling. This was attempted to be applied consistently however, as Fig. 3.5 demonstrates, a large variation in the amplitude was observed. This also demonstrates the importance of human factors in inspections and the need to account for the variations they cause in assessments of inspection capability. The worst case scenario is that all parameters are equally likely which corresponds to the operator randomly guessing the values of parameters over which they have control, such as the position of the probe. Experience from trials and engineering judgement should allow this to be estimated more accurately. As this distribution may change as more knowledge is gained or training has a greater effect, it is beneficial to be able to alter these probability distributions without having to re-evaluate models therefore the general approach discussed previously is particularly suited to this. It is possible that these probability distributions will change over the course of a qualification campaign based on improved knowledge or optimisation of the inspection. This highlights one advantage of the general method discussed previously, that the changing of the probability distributions does not require any further model evaluations.

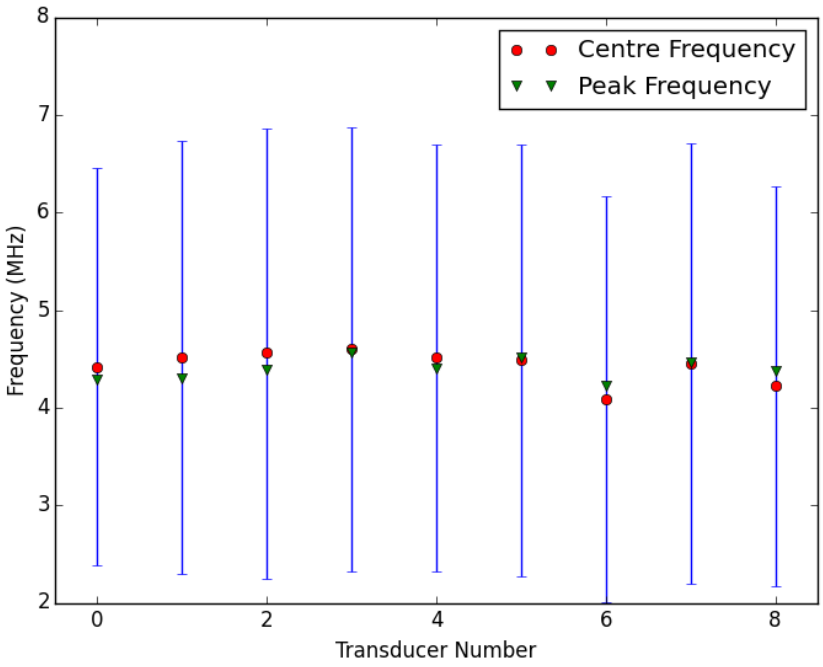

FIGURE 3.4: The properties of a set of ultrasonic transducers that are sold nominally as 5 MHz probes. The centre frequency of the probe is the frequency that is at the centre of the bandwidth range, shown by the blue lines, which is defined as the frequencies above and below the peak frequency that the amplitude of the output of the transducer falls to 6 dB below the peak frequency.

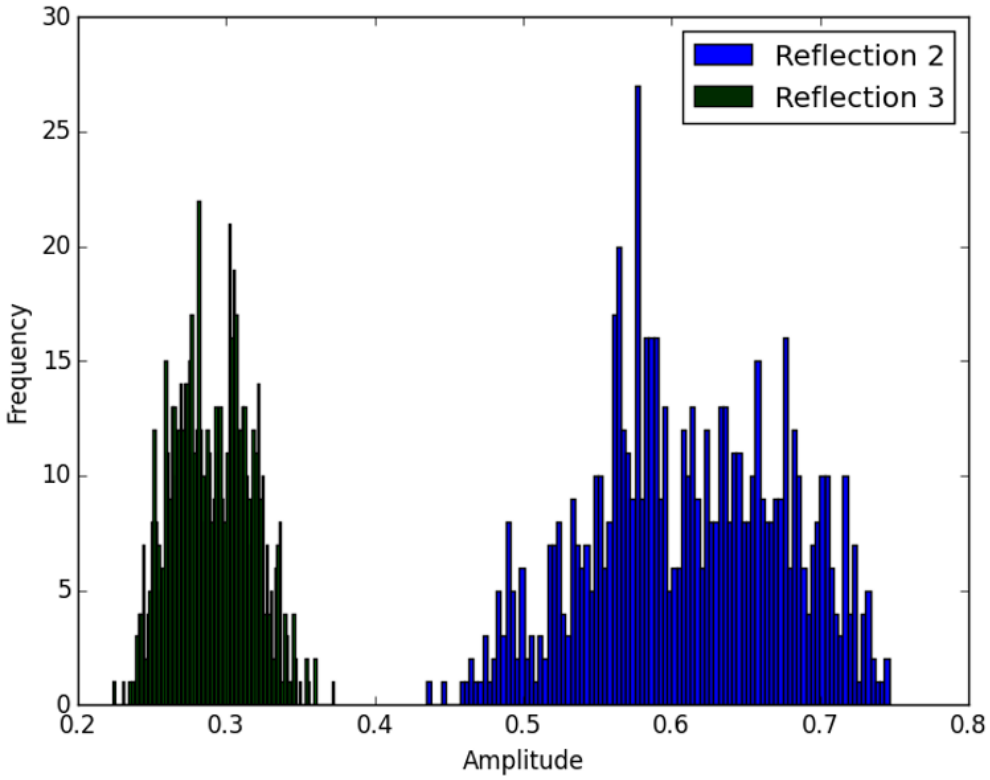

FIGURE 3.5: The amplitude of the ssecond and third reflections from the back wall of an aluminium block, normalised to the amplitude of the first reflection, generated using an inspection performed nominally identically. The change in the thickness of the applied coupling causes the significant variation in the measured amplitude.

3.8 Summary

This chapter has discussed methods of calculating quantitative metrics of an inspection through different methods and presented a generalised way of calculating metrics such as PoD and PFA amongst others, making no assumptions as to the nature of the response or probability distributions. This is a significant advantage over current methods and the decoupling of the response of the inspection from the probability of variations occurring allows the latter to be changed without having to re-evaluate any models. It is therefore trivial to generate significant information about the capability of a technique by accurately establishing the response function and probability function as well as gaining insight into how to optimise the technique. However, the response function must be established accurately in a reasonable time. Methods of achieving this using numerical models whilst minimising the number of model evaluations is discussed in the next chapter.

Chapter 4

Mapping the Response Function

A key part of assessing the capability of a technique is to establish all possible outcomes of the inspection throughout the parameter space Ω, mapping the response function $R(\Omega)$. As the number of parameters considered increases, the volume of Ω scales to a power law, therefore it quickly becomes impractical to evaluate $R(\mathbf{x})$ at every $\mathbf{x}$ in Ω. This is especially true for models of inspections which may take a significant amount of time to run. Thus a method is needed which allows $R(\Omega)$ to be mapped using the smallest number of model evaluations to minimise the total qualification time. As the response is typically a quantified metric, such as an amplitude, a phase or an area on an image, it is possible to sub-sample Ω and approximate $R(\Omega)$ to a high degree of accuracy using numerical interpolation. This process can be applied to any inspection for which there is a model. It is also possible to apply this methodology to an experimental qualification campaign if the samples and conditions of the inspection can be accurately controlled although this may become a very expensive exercise.

This chapter presents methods which can map $R(\Omega)$ through sampling and interpolation using a small number of model evaluations. This methodology assumes that the model has been appropriately validated for the inspection and validation is discussed in more detail in later chapters. In practice, it is possible that not all parameters have a significant effect on the outcome of the inspection. If it were possible to ignore some parameters, the dimensionality of the parameter space can be reduced and thus the number of model evaluations reduced. This is possible through the use of quantitative sensitivity analysis and is discussed in this chapter.

4.1 Sampling and Interpolation

The process of sampling and interpolating over a parameter space to build an approximation of a function is a common challenge across a broad spectrum of fields and is often referred to as surrogate modelling. The goal is to use a small number of model evaluations to build a numerical surrogate model, typically some form of analytic or numerical interpolating function. Many numerical models, with the exception of Monte Carlo based models that are used for example in models of radiographic inspections, are deterministic, that is running a model multiple times with the same inputs will result in the same output to within numerical error. Both finite element methods and ray tracing algorithms fall into this category. This has the advantage of not having to evaluate a model multiple times for a given coordinate in Ω, thus reducing the number of model evaluations required. The following work assumes that a deterministic model is used. If a Monte Carlo-based model or another model with some variance in its outcome is being used, the following methodology may be applied if the degree of variance in the outcome of repeated model evaluations is small, that is a single model evaluation is an accurate representation of the true result. If this is not true then an additional adjoint parameter space which quantifies the variance of the response throughout Ω will be required. This will result in a greatly increased qualification time and direct Monte Carlo integrals of the inspection capability metrics will likely be a more efficient method of quantifying reliability metrics.

The general outline of the sampling and interpolation process is as follows. An initial set of samples is chosen, $S_0 \subset \Omega$, and the response of the model is evaluated at these points. A predictor is built from this data set and its quality is tested by calculating a predictive error. Methods of determining the predictive error are discussed in the following section. A subsequent set of samples, $S_1 \subset \Omega$, is chosen such that there is no overlap between the sample sets, that is $S_0 \cap S_1 = 0$. The model is evaluated at S_1 and the predictor is rebuilt using the combined set $S_T = S_0 + S_1$. If the error of the predictor is not below the desired threshold, another set of samples $S_2 \subset \Omega$ is chosen, again such that $S_T \cap S_2 = 0$, the models evaluated at these points and the predictor built using the new total set $S_T = S_T + S_2$. This process is repeated n times until the use of sample set S_n results in an error metric that is below the desired threshold.

Three questions therefore arise: what is an appropriate method to determine S_0 through S_n, what choice of interpolation method can can yield a good prediction of $R(\Omega)$ and what method can be used as a test of the predictive error. For generality, it is assumed that there is no prior knowledge of any features of $R(\Omega)$ and that $R(\Omega)$ is non-linear. It is not strictly assumed that the response function is continuous over Ω, only that the response function is continuous on the length scale of the interpolation method such that the interpolation is valid.

4.1.1 Testing Interpolation Quality

A key stage in this process is how to determine if the predictor is accurate. There are a broad range of methods to do this however they all involve testing the predictor at a set of sampled points which are not used in the construction of the predictor. The primary reason for the independence of the sets is that often numerical predictors, such as linear interpolators, guarantee that the interpolation is exactly the sampled value at the sampled coordinates therefore the error would be zero always.

One possibility is the independent error set method. This generates a set of samples S_E that is distributed throughout Ω such that it is independent of all the sample points used for building the predictor, that is $S_T \cap S_E = 0$. The benefit of this method is that the interpolator will be tested throughout Ω and that the number of error points and their location is determined independently of the choice of samples to build the predictor. The disadvantage of this method is that effort is effectively wasted on model evaluations which do not contribute to the building of the predictor. It is possible, although unlikely, that the number of model evaluations required to build a good predictor is less than the number of points in the error set. In this case, more than 50% of the computational effort will have been expended gaining information which will not be used which is therefore inefficient. Methods of choosing S_E are discussed in the following section in more detail as the same issues are faced in choosing this set and choosing the sets S_0 through S_n.

An alternative method is to use Generalised Cross Validation (GCV) [63]. The idea of this method is to remove a subset of the sampled points, build a predictor based on the reduced set and test it at the removed subset. If a predictor is of a good quality then the removal of some points from the predictor should not degrade it significantly. The question then arises of what is the best

subset to remove which involves partitioning the sample set in some manner. This can be done in either a systematic or random manner with different sizes of sample sets however these will always be susceptible to the choice of error subset and which points end up in the building of the predictor. The impact of this choice can be mitigated to some extent by repeating the process with different subsets such that all points are in the predictor at least once, so called exhaustive cross validation. This will still be sensitive to the the choice of error sets therefore the logical extreme of this process is to apply this to every possible subset of size 1 to $N-1$ for N sampled points. The number of possible subsets, n_s, is given by

$$n_s = \sum_{k=1}^{N-1} \frac{N!}{k!(N-k)!}, \tag{4.1}$$

where k is the number of points in the subset. This will quickly become very large as N increases, for example 100 sampled points has 1.27×10^{30} possible sample sets, and is therefore not a feasible solution. The impact of degrading the predictor through removing sampled points whilst not generating an excessive number of subsets can be minimised by using the Leave-One-Out Cross Validation (LOOCV) [64] method. In this process, each sampled point is in turn removed from the building of the predictor and the subsequent predictor is then tested at the omitted point. This avoids the inefficiency of the independent error set method in which effort is wasted evaluating points that are not used in building the predictor. However this method is still susceptible to the choice of error set which is implicitly determined by the choice of the sample sets S_0 through S_n. A potential disadvantage of this method is that for N sample points, the predictor has to be constructed N times. Some complex prediction algorithms require an appreciable time to build therefore this has been found in practice to become a prohibitively expensive method as N increases. As the number of sample points required to accurately map $R(\Omega)$ is evidently not known prior to the mapping, the independent error set method is used in this work as this provides a fixed cost of evaluating the quality of the predictor rather than GCV which can potentially become prohibitively time consuming.

Whichever method of choosing the error set is used, a quantitative error metric is required to evaluate the quality of the predictor. The most common metrics used for testing the quality of model predictors are the root mean

squared error (RMSE) and the mean absolute error (MAE). The former is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R(\mathbf{x}_i) - R'(\mathbf{x}_i))^2}, \tag{4.2}$$

for N points at which the error is evaluated, $R'(\mathbf{x})$ is the predicted value at the coordinates $\mathbf{x}$ and $R(\mathbf{x})$ is the evaluated response at the coordinate. This gives a greater weight to errors with larger variance. The MAE is defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|R(\mathbf{x}_i) - R'(\mathbf{x}_i)|, \tag{4.3}$$

which gives equal weight to all magnitudes of error and is used in this work.

4.1.2 Latin Hypercube Sampling

Given that there is no prior knowledge of any features of $R(\Omega)$, it is reasonable to distribute the points in $S_E, S_0, ..., S_n$ throughout Ω. This can be achieved in a number of ways, such as using a regular grid or random sampling. An alternative method is to use Latin Hypercube Designs (LHDs, also known as Latin Hypercube Sampling) [65, 66]. These are defined as each value of each parameter being sampled once and only once. This has the benefit of testing many values of each parameter and, with appropriate design, providing a set of samples that are well distributed throughout Ω. The definition of a sample set being well distributed is subject to debate, for a discussion see [67], however in this work the optimality criterion proposed by Jin, the ϕ_p criterion, [67] is to be minimised, defined as

$$\phi_p = \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^{-p}\right]^{\frac{1}{p}} \tag{4.4}$$

where N is the number of points in the design, d_{ij} is the Euclidean distance between points i and j and p is a chosen number. The best value for p is itself an area of debate. In the limit $p \to 1$, the total distance between all points will be maximised whereas in the limit $p \to \infty$ the shortest distance between any two points will be maximised. Therefore a trade-off between these two cases is needed and following the work of Jin, $p = 50$ is used. Given this optimality criterion, there are a range of methods which can be used to generate the design. All designs consist of N points in each of n_v variables and are built

in the space $[0, N-1]^{n_v}$ before scaling to the values of the input parameters. This allows the same design to be reused if the values of one or more parameters are changed and avoids the issue of disparate scales of variables which will affect the calculation of distance within the design. It is also important to note that for the purposes of this project it is not necessary to find the globally optimal LHD, only one that is highly optimal. Therefore methods which are capable of finding a design that achieves this rather than definitively globally optimal are appropriate.

The simplest method of generating a LHD is to generate n_v vectors of random ordering of the vector $[0, 1, ..., N-2, N-1]$. A simple optimisation method is to repeat this process a number of times and choose the design that has the lowest ϕ_p value. An alternative optimisation method is to perform pair-wise column-wise switching in which a pair of values in two vectors is chosen and swapped. The ϕ_p criterion is then calculated for this new design and the process is repeated either a given number of times or until a desired optimality is reached. Whilst this may result in a highly optimal design, it is also likely to result in a design that is far from optimal therefore a method that more reliably results in a highly optimal LHD is required.

There are a wide range of optimisation algorithms that can be applied to this problem however given that the input parameters are discrete, the problem is better suited to stochastic methods rather than any that involve gradient descent, such as Newton-Raphson iteration [68]. Within stochastic methods, there are a wide range of techniques such as simulated annealing [69] and genetic algorithms [70]. A method that has been demonstrated to work well for LHDs is a modification of simulated annealing, the Enhanced Stochastic Evolution (ESE) algorithm [67]. The algorithm is outlined here however the reader is directed to [67] for a more detailed discussion.

Enhanced Stochastic Evolution is based on the idea of jumping around the design space to find a more optimal design and reducing the size of the jumps to move towards a local minimum which is ideally the global minimum. The size of the jumps is determined by a single scalar parameter, the temperature of the simulation as in simulated annealing, and as the optimisation cools the size of the jumps decreases. The key difference between ESE and simulated annealing is that ESE has a temperature curve which rapidly decreases the temperature, thus rapidly converging on a locally optimal design, before rapidly increasing the temperature again, thus looking globally for a more optimal design. This

process is repeated several times until some termination conditions are met. Whilst this does not necessarily produce the globally optimal design, it does rapidly produce a highly optimal design and can incorporate information from previous designs and results.

An alternative approach to generating optimal designs efficiently is the method of Viana et al. [71], the Translational Propagation algorithm. This approach uses an initial seed design which is propagated to generate the complete design. The use of different initial seeds results in different final designs whose optimality can be evaluated using the ϕ_p criterion and the best design chosen. The process of propagating a seed is demonstrated in Fig. 4.1 for a two parameter LHD. The primary benefit of this algorithm is that it is significantly faster than the ESE algorithm as it involves only vector addition and significantly fewer optimality evaluations which are computationally expensive.

These methods can be compared by generating LHDs for a range of sizes and dimensions to compare the optimality value. This is demonstrated for a LHD of three parameters with a number of points varying from 11 to 111. The random variation in the methods is accounted for by taking the average of 21 designs at each size and calculating the standard deviation of the optimality. The optimality criterion used is the ϕ_p criterion as in Eqn. 4.4. The results of this are shown in Fig. 4.2. This shows that the the ESE and TPLHD methods consistently generate more optimal designs than the random method.

4.1.3 Choice of Error Set Size

In the case where an independent error set is used to calculate the predictor error, the question remains of how many points should be used for the error set. The calculation of the average error throughout Ω is equivalent to the numerical integration of the error function. Each point in the independent error set is an approximation of the error in its local vicinity. Therefore an increase in the number of points in the error set and thus its density reduces the volume of the voxel around each error point and increases the accuracy of the approximation. For a smooth function, the maximum deviation in the error is likely to be at the furthest distance from any error point, as this provides the weakest approximation of the error value. Thus, the maximum distance of any point in Ω from any point in the independent error set can be used as a characterisation of the quality of the independent error set. The addition of

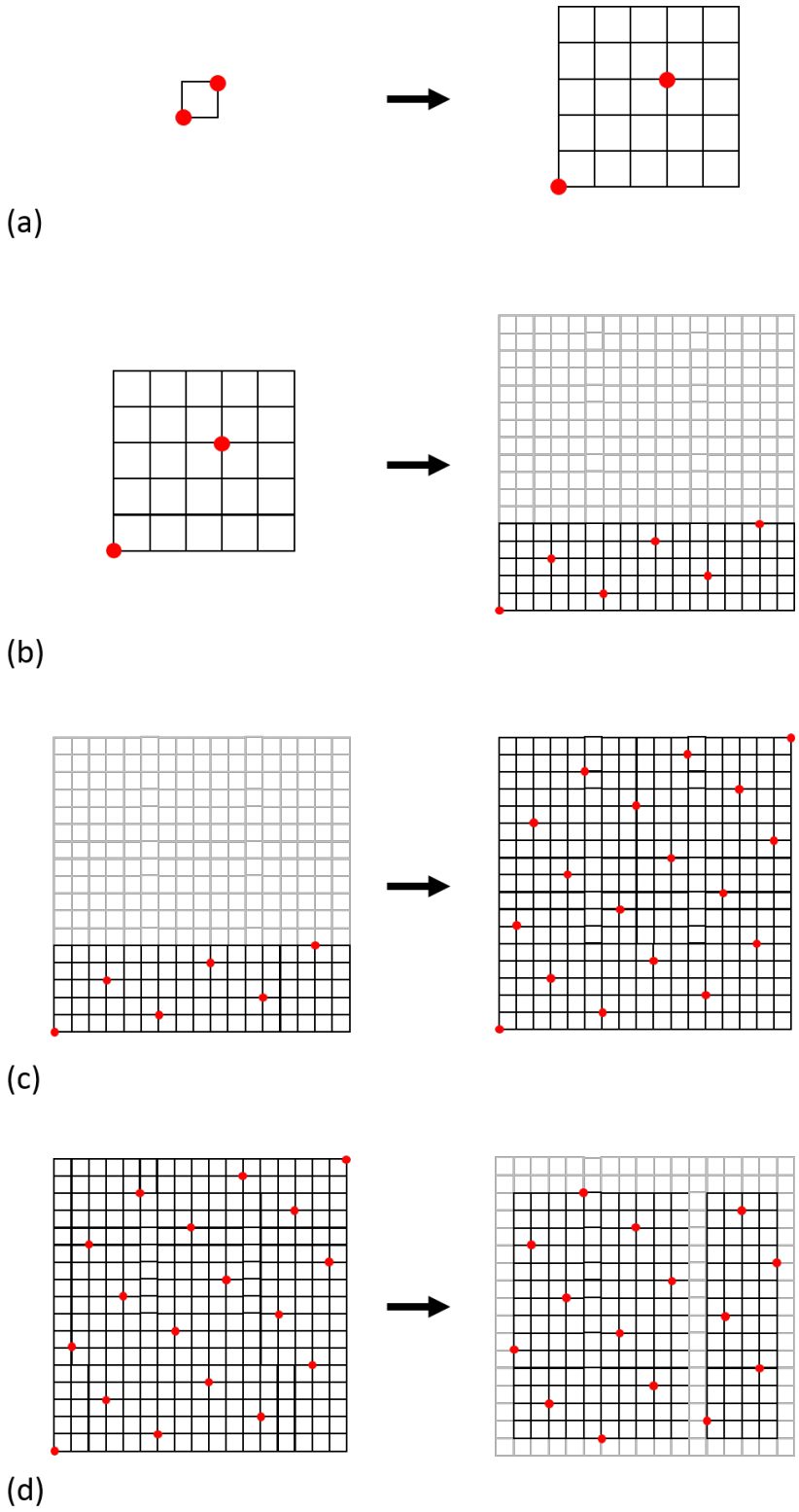

FIGURE 4.1: The development of a 18 point Latin Hypercube Design from an initial 2 point seed design. (a) The initial 2 point seed is scaled to an appropraite block size. (b) The block is propagated through the first dimension. (c) The result of (b) is used as the seed for the propagation in the second dimension. (d) In the case of generating a smaller design, a larger design may be reduced by removing both the points and the levels associated with those points.

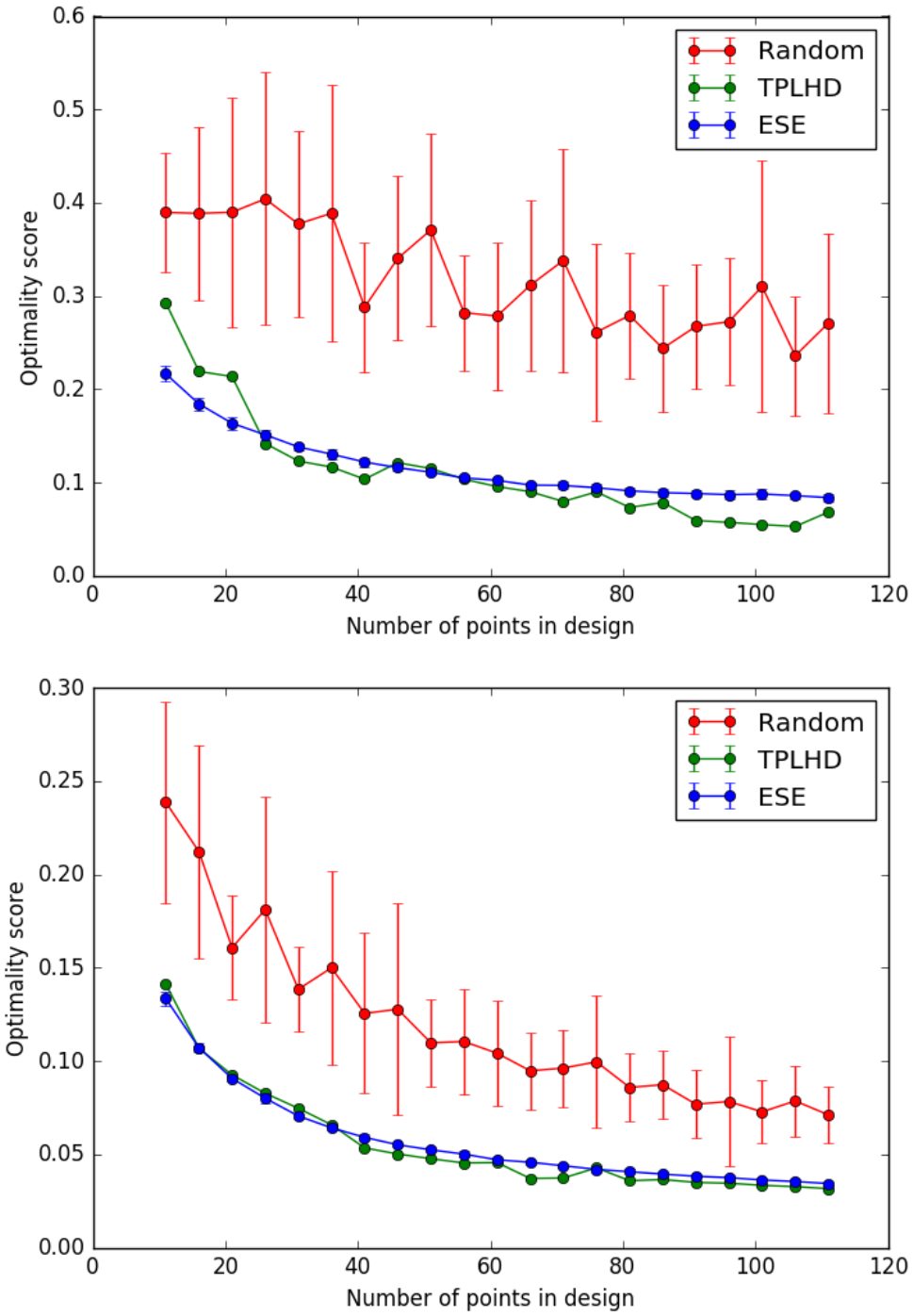

FIGURE 4.2: A comparison of methods for generating Latin Hypercube designs of 3 parameters (top) and 5 parameters (bottom) for a range of number of design points by calculating the optimality of the design. Three methods were used for generating designs: random generation, the translational propagation algorithm (TPLHD) and the enhanced stochastic evolution algorithm (ESE).

more points will reduce the maximum distance any point in Ω is from any point in the independent error set. This process however has diminishing returns and therefore there is a number of error points above which the computational cost of adding in more points outweighs the decrease in the maximum distance. This process can be investigated numerically using Voronoi diagrams [72] which determines the volume around a sample point in which all points in that volume are closer to that sample point than any other. An example of this for a two parameter space with 20 points is shown in Fig. 4.3. The maximum distance between any sampled point and its Voronoi nodes gives the maximum distance between any point in the parameter space and any sampled point. This process naturally extends to higher dimensions so can be used for any dimensionality and size of design. This method could also be used as an alternative to the ϕ_p method discussed previously as a metric of LHD quality however given that it requires several orders of magnitude of time greater to calculate, it is less suited to being used as a quality metric in an optimisation process.

The Voronoi diagram however must be constrained on the edges of the volume for this process to work. The diagram will be constrained if the corners of the volume are included as sample points but this is not possible with LHDs as the definition of each value of each parameter being sampled once and only once precludes this possibility. This is shown in Fig. 4.3 (top) in which it is clear that the boundary of the unit hypercube is poorly constrained. In this example, the sampled point closest to [0,0] is further away from this point that any of its Voronoi vertices therefore a calculation of the maximum distance only using these vertices would be incorrect. This can be mitigated however by reflecting the sample points in all faces of the parameter space. This has the effect of creating points just outside of the original parameter space which can be used to create the Voronoi diagram and results in a constrained plot, as shown in Fig. 4.3 (bottom) where the plotted area has been constrained for clarity. The calculation of maximum distance is performed in n-dimensional Euclidean space and is computed for each Voronoi node which is inside or on the unit hypercube for each sampled point. The maximum value for each node is then compared to obtain the maximum value for that design.

This process was applied to LHDs on a unit hypercube generated using the ESE algorithm with varying numbers of parameters and sample points. For each combination of parameters and sample points, 28 designs were generated to obtain a measure of the variance of the maximum distance. The results of this are shown in Fig. 4.4 alongside the maximum distance, d_{max}, for a regular

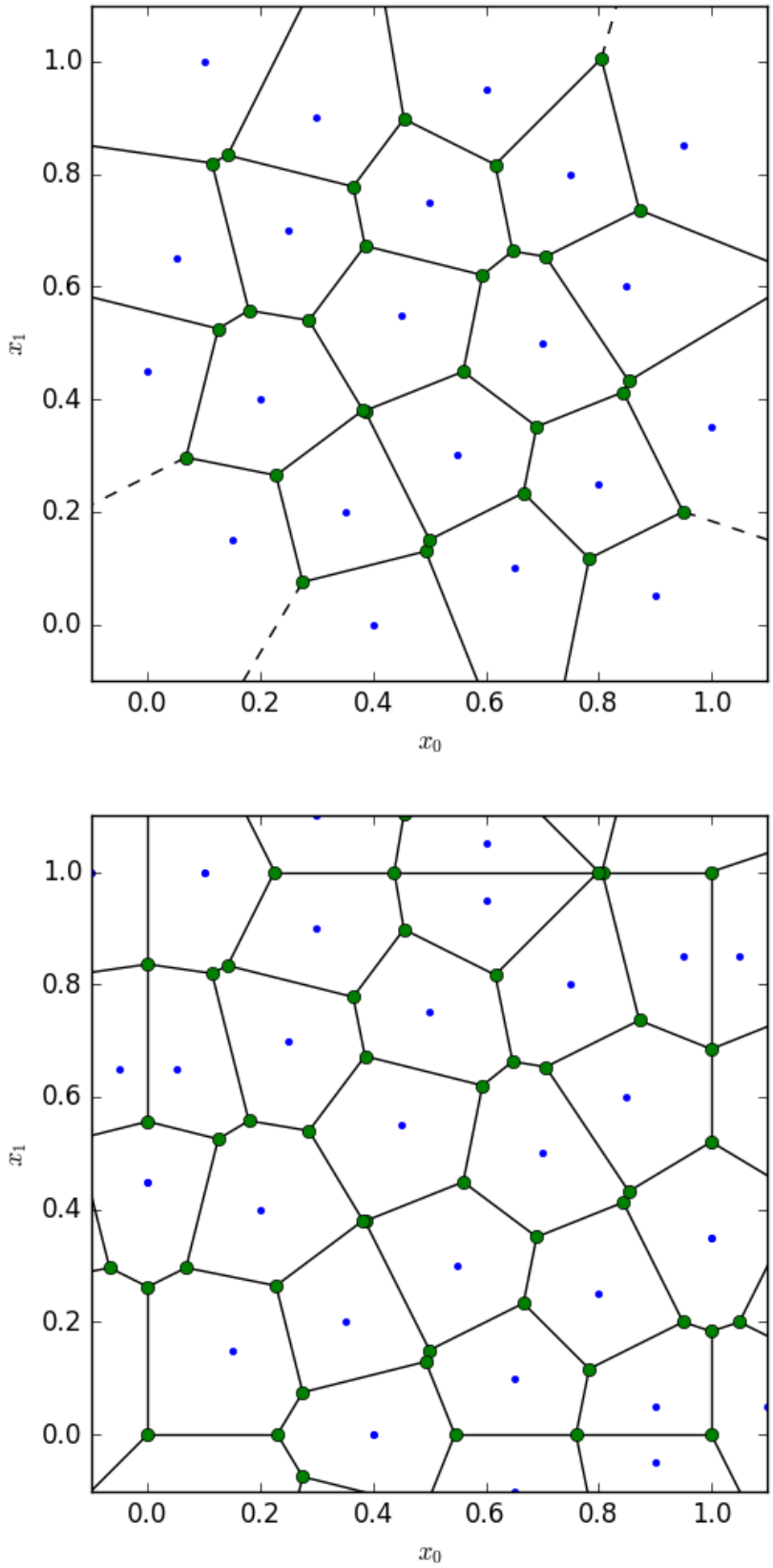

FIGURE 4.3: The Voronoi diagram for the original Latin Hypercube Design (top) and with it reflected in all faces of the unit hypercube (bottom), providing a better constraint of the boundary of the unit hypercube. The sampled points of the design are in blue and Voronoi vertices in green. The solid black lines indicate finite ridges and the dashed lines indicate the infinite ridges, that is the Voronoi vertex at one end of the ridge is at infinity.

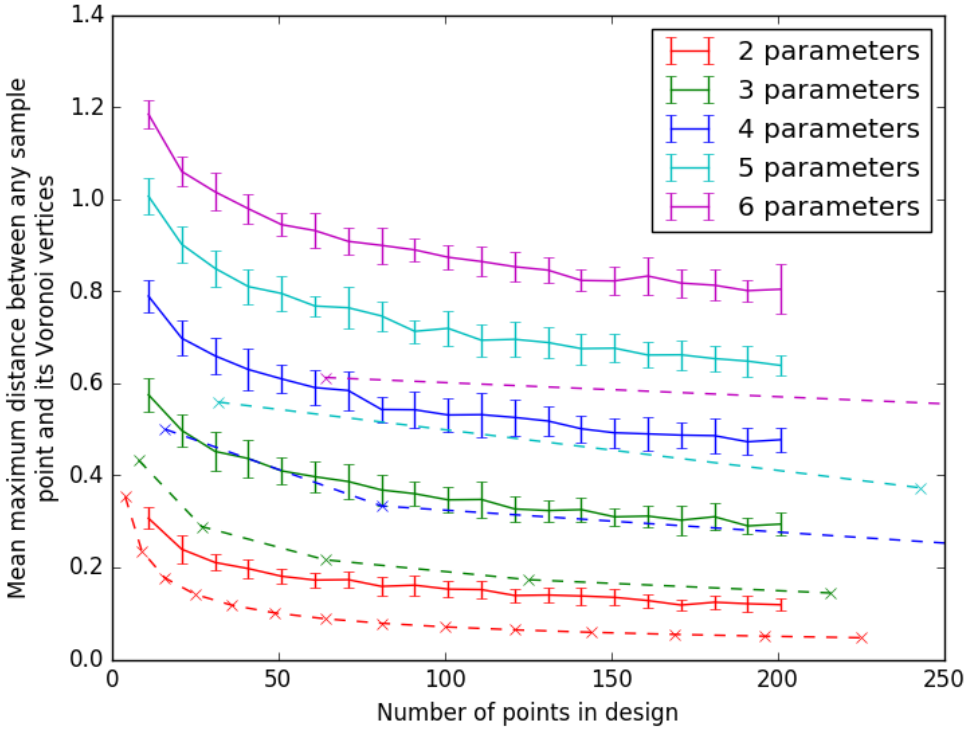

FIGURE 4.4: The maximum distance between any point in the unit hypercube and any sampled point for both Latin Hypercube Designs (solid line) and sampling on a regular grid (dashed line).

grid. These are calculated for a grid of N points per dimension for each of n_v dimensions as

$$d_{max} = \frac{\sqrt{n_v}}{2N}. \tag{4.5}$$

This demonstrates that a regular grid produces smaller maximum distances however for larger numbers of parameters, as the number of points in a regular grid scales as N^n, it becomes prohibitively expensive to add in more points to the independent error set. For example, for a 6 parameter space a regular grid of $N = 2$ has 64 points whilst a regular grid of $N = 3$ points has 729 points. This also has the disadvantage of testing only a few values for each parameter. Given that the LHD method has more flexibility to produce designs of arbitrary numbers of points and that it does produce designs that test multiple values per parameter, this method will be used. It should be noted that, as discussed later in this chapter, using LHDs for function approximation is shown to work well therefore it is a valid method to use.

Figure 4.4 demonstrates that for all numbers of parameters, as the number of points in the design increases, the maximum distance decreases at an ever slowing rate. Therefore this plot can be used to determine a good choice of

error set size to be used as a trade off between the time required to generate the error set and the maximum distance any point is away from the error set. For a given number of dimensions of the LHD n_v, the average volume around each of the N points in the design is proportional to $N^{-\frac{1}{n_v}}$. Figure 4.4 suggests that there is an offset in the intercept for each of the different values of n therefore it is appropriate to fit a function of the form

$$f(N) = a + \frac{b}{N^{\frac{1}{n_v}}}, \tag{4.6}$$

where a and b are parameters to be determined. The estimated values of a and b for each value of n_v are summarised in Table 4.1 and the functions are plotted in Fig. 4.5. This can be used to define a convergence rate. In general, this is defined as a proportional increase in the number of samples, CN resulting in a proportional change in the value of the function, D, that is

$$\frac{f(N) - f(CN)}{f(N)} = D. \tag{4.7}$$

Substituting Eqn. 4.6 into Eqn. 4.7 and rearranging allows the number of samples required to achieve a desired convergence rate to be defined as

$$N = \left(\frac{b(1 - C^{\frac{1}{n_v}} - D)}{Da} \right)^{n_v}. \tag{4.8}$$

For example, for a 4 dimensional LHD with a convergence criterion of a change in volume of 10% ($D = 0.1$) at a cost of doubling the number of points in the design ($C = 2$), a design of 186 points is required. This information can be used to determine how many points are required in an error set to give a desired level of coverage of Ω as well as how many additional points need to be added to the error set to improve the coverage of Ω by a desired amount. In practice, the choice of the size of the error set is a trade off between desired accuracy and the cost of a model evaluation. The higher the dimensionality of Ω, the more points required should be used in the error set. In this work, a minimum of $11n_v$ has been used for the size of the error set as a trade off between the cost of generating the error set and the providing good coverage across the space. In practice, a higher number may be appropriate to obtain a more accurate estimate of the error of the prediction however this will be determined on a case by case basis as a trade off between the cost of evaluating the additional models and the benefit of obtaining a more accurate estimate of the error.

Dimensions	a	b
2	0.069	0.800
3	0.124	1.032
4	0.180	1.124
5	0.187	1.322
6	0.202	1.445

TABLE 4.1: The values of the fit parameters in Eqn. 4.6 for different numbers of dimensions.

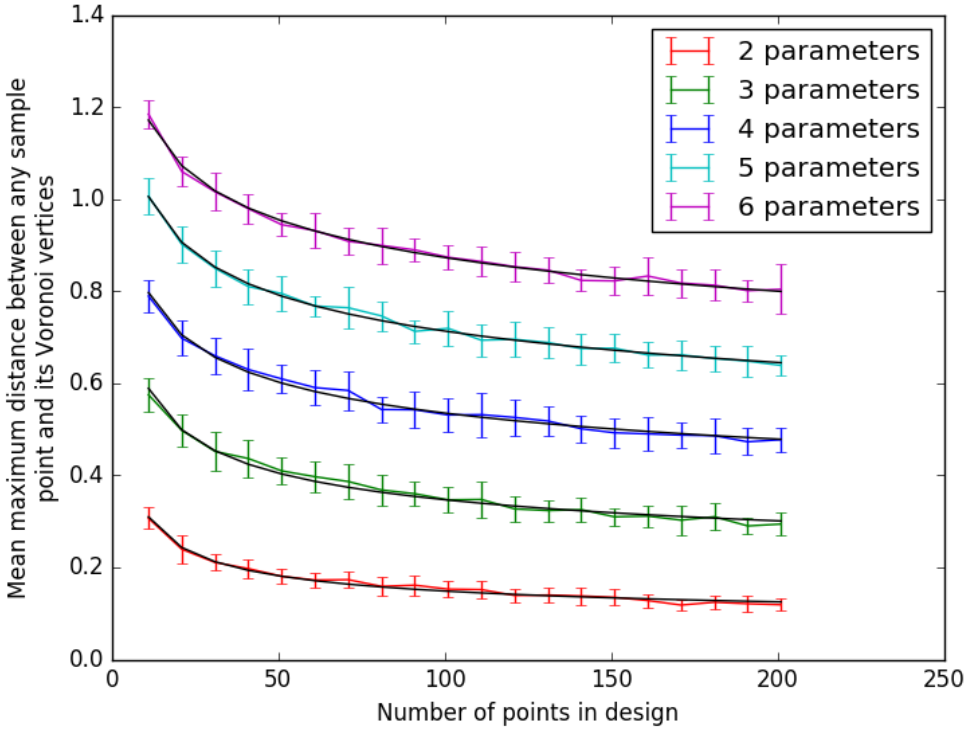

FIGURE 4.5: The maximum distance between any point in the unit hypercube and any sampled point for both Latin Hypercube Designs and the fit of Eqn. 4.6 (black lines).

4.1.4 Parameter Space Mapping using Latin Hypercube Designs

Latin Hypercube Designs can be used to perform general parameter space mapping when coupled with an interpolation method. Using the ESE method, after the first iteration, S_T can be passed into the generation function of the next set of sample points S_i and the optimality of the combination of $S_i + S_T$ can be maximised. The condition that there is no overlap between the set S_i and S_T is also maintained. This method therefore builds up a distribution of points throughout the space with each value of each parameter being sampled n times after n sets of sample points have been added to S_T. This process was also considered using the TPLHD algorithm using a shift of the initial design. It was found however that it was not always possible to find a shift with no overlaps between the sets therefore this method is not suitable.

4.1.5 Interpolation Methods

Given any sampling method, an interpolation method is required to estimate the response across Ω from the sampled points. There are a wide range of techniques that can be used for multivariate interpolation. It was decided that, where possible, pre-built interpolation packages would be used as these tend to consist of highly optimised code and the focus of this project was not to develop a highly efficient multivariate interpolation routine. The following presents and compares multivariate interpolation methods which can be used.

Linear Interpolation

The simplest interpolation method that naturally extends to higher dimensions is linear interpolation [73]. The most common method of linear interpolation for arbitrary numbers of dimensions is to use linear barycentric interpolation. This involves calculating the Delauney triangulation [74] of the sampled points to create a set of simplexes and performing linear barycentric interpolation within each simplex to approximate the response value. Each simplex in n_v dimensions has $N_{\text{vert}} = n_v + 1$ vertices. The barycentric coordinates $\mathbf{x}_B$ at any given point within a simplex is the weighted sum of the coordinates of the

vertices, that is

$$\mathbf{x}_B = \sum_{i=1}^{N_{vert}} \alpha_i \mathbf{x}_i, \tag{4.9}$$

where α_i is a weighting coefficient. These coefficients act as weights for interpolation and the value of the function at $\mathbf{x}_k$ can be approximated as

$$R(\mathbf{x}_k) \approx \sum_{i=1}^{N_{vert}} \alpha_i f(\mathbf{x}_i). \tag{4.10}$$

The coefficients can be calculated by treating the coordinates as a system of linear equations and solving for the coefficients and normalised such that they sum to 1. Any interpolation method which is not capable of also performing extrapolation, which linear barycentric interpolation is not, requires the convex hull of the input to cover the whole of Ω. This will not be possible with LHDs alone therefore the corners of Ω must be sampled first to ensure that this condition is met. This interpolation algorithm is reasonably fast and significant research has already been expended on developing optimised code, therefore the SciPy wrapper of the QHull implementation of the Quickhull algorithm [75] is used. The primary disadvantage of this method is that a piecewise linear representation may not be a good approximation of the response function if it is rapidly varying however it will be very accurate in regions where the function slowly varying and thus is linear to a good approximation. Another issue is that it does not guarantee that the interpolator is smooth, that is the first derivative is not continuous across the boundary of two simplexes.

Radial Basis Functions

A common non-linear interpolation method is radial basis function (RBF) interpolation [73]. A radial basis function $\psi(r)$ is one whose argument is the distance r of the interpolated point $\mathbf{x}$ away from the sampled point $\mathbf{x}_i$. A common choice is the Euclidean distance, that is $r = |\mathbf{x}_i - \mathbf{x}|$ although any metric of distance can be used. The function itself can be any which has a sole argument and common ones are linear, Gaussian and multiquadric. These functions are used as an interpolator by calculating the weights β_i for each of N data points such that the error between the function values and actual values

is minimised. The function approximation is therefore

$$R(\mathbf{x}) = \sum_{i=1}^{N} \beta_i \psi_i(|\mathbf{x} - \mathbf{x}_i|). \tag{4.11}$$

The coefficients can again be calculated by treating the calculation as a set of linear equations and solving for the coefficients. This method requires selection of the basis function and choice of any parameters in these functions, for example the variance of a Gaussian basis function must be chosen. As these interpolators require a short time to construct, several basis functions may be used and the one that gives the lowest predictive error used.

Multivariate Adaptive Regression Splines

An alternative approach to non-linear interpolation is the multivariate adaptive regression splines (MARS) algorithm. This was first proposed by Friedman [76] as an extension of the recursive splitting algorithm. For clarity, the algorithms are summarised here, for exact forms see Algorithms 2 and 3 of [76] for the forward and backward stepwise algorithms respectively. The basic idea of MARS is to develop a function which consists of a sum of basis functions through recursive splitting of domains.

Initially, the model has a single domain, the entire parameter space, and a constant basis function, β_0, is fitted. Thus after the first pass in the forward step, the predictor of the response function $R'(\Omega)$ is of the form

$$R'(\Omega) = \beta_0. \tag{4.12}$$

The value of β_0 is chosen to minimise the mean square error as defined in Eqn. 4.2. It should be noted that unlike linear and RBF interpolation, MARS does not force the predictor through the sampled values therefore ϵ is not always zero. The next pass finds the pair of reflected basis functions whose addition to the function minimises the RMSE of $R'(\Omega)$ with respect to the sampled points. The basis functions are commonly referred to as hinge functions although by definition they are truncated power basis functions and are defined for a single variable x_i as

$$h(x_i, c)^{\pm} = max(0, \pm(x - c)), \tag{4.13}$$

where c is the knot location. This function has a discontinuity in the first

derivative and thus is able to accommodate discontinues in the first derivative of $R(\Omega)$. These may be added as single functions or products of these functions, in one or more parameters, which accounts for interaction between variables and non-linear trends. All function coefficients are recalculated upon the addition of a pair of basis functions. Therefore β_0 is recalculated as part of this pass. The function, in a simple case, after the second pass is thus in the form

$$R'(\Omega) = \beta_0 + \beta_1 h(\mathbf{x}, c_1)^+ + \beta_2 h(\mathbf{x}, c_1)^-. \tag{4.14}$$

Note that the hinge function is expressed in terms of a vector of coordinates, $\mathbf{x}$, but acts only in a single variable. This results in two domains, one on each side of the knot. Importantly, upon the splitting of a given domain by the addition of a pair of reflected basis functions, the parent basis function is retained. This process is recursively repeated for splitting each domain by adding pairs of reflected basis functions until either ϵ is below a tolerance or a maximum number of basis functions M is added, the value of which is chosen by the user.

The end result of the forward step is an overly constrained model. It is desirable to have a model consisting of fewer basis functions as this fits the global trends rather than fitting local noise. There is therefore a trade off between the accuracy of the model and the number of terms and it is desirable to remove some terms from the over constrained model. Thus a pruning process is employed in a backwards step to remove basis functions. The metric used to assess a model consisting of w basis functions is a weighted version of the General Cross Variance (GCV) [76], given by

$$\mathrm{GCV} = \frac{\mathrm{RMSE}}{\left[1 - \frac{w(1+d)}{N}\right]^2}, \tag{4.15}$$

where d is a cost per basis function and all other terms retain the same meaning. The value of d can be decided upon, however, a value in the range of 2 to 4 is generally chosen on the advice of Friedmann [76]. The backward step goes through every w value from M to 2 and finds the basis function whose removal improves the fit by the most or least degrades the fit, by calculating the GCV for the removal of every eligible function. The constant basis function is never eligible for deletion which ensures that all points in the space being mapped always have a functional value. The number of basis functions which minimises the GCV is the selected model as this will have the best trade off between

error and number of basis functions. The resultant function can be cast in a number of ways. A useful form which allows the interaction of parameters to be determined is

$$R'(\Omega) = \beta_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + ..., \quad (4.16)$$

where the first sum is over all basis functions which involve a single parameter, for clarity expressed as the function f, the second is over functions which involve two parameters hence showing the interaction between them, the third sum is over all functions which involve 3 parameters and so on. This is known as the Analysis Of Variances (ANOVA) decomposition and allows the interactions between the parameters to be investigated. Summing the variances of these functions allows the relative importance of the parameters to be assessed however this requires accurate fitting of the model. Given this limitation and the desire to assess the relative importance of parameters using other interpolation methods, alternative sensitivity metrics are discussed in the following sections. Two implementations of the algorithm were used in this work. The ARESLab [77] implementation was initially used which is a MATLAB based library however the pyEARTH package in Python was found to be more flexible as well as being easier to integrate into the wider tool chain. This method is more flexible than the others previously discussed however it does have a significantly greater degree of tuning required to obtain an accurate result and computationally it is significantly more expensive.

4.1.6 Comparison of Interpolation Methods

These interpolation methods can be compared using the example functions defined in Chapter 3. Consider the first function, Eqn. 3.19 with $a = b = c = d = 1$. This function was sampled at the corners of the parameter space Ω, requiring 16 samples, and an increasing number of LHDs, to a maximum of 10, of 21 points each generated using the algorithm discussed previously. Linear interpolation and the MARS algorithm are compared with two radial basis functions, defined as

$$\psi_1(r) = r^3, \quad (4.17)$$

and

$$\psi_2(r) = e^{-\left(\frac{r}{\beta}\right)^2}, \quad (4.18)$$

where β is a parameter that must be tuned to minimise the error of the interpolation. To account for variation in the choice of samples, this process was repeated 28 times to obtain an estimate of the error in the quality of the interpolators. The mean absolute error of the prediction at all points on a regular grid of 21 points per parameter was used as a metric of the interpolator quality, as defined in Eqn. 4.3.

The results of applying these interpolators are shown in Fig 4.6. This shows that the MARS algorithm and linear interpolation are capable of very accurately predicting this function given a very small number of samples. This is due to these being capable of producing a very good approximation of the underlying function. In the case of MARS, it is able to use purely linear basis functions for each parameter which when summed accurately represents the underlying function. Of the two RBF methods, the Gaussian radial function gives a relatively poor result and the cubic radial function gives a better prediction, comparable to the MARS and linear methods however requires a larger number of samples to reach the same level of accuracy. These results are due to these radial basis functions being unrepresentative of the underlying response function therefore they form poor predictors.

The same methodology of comparison is also applied to the more complex example function Eqn. 3.20. The results of this process are shown in Fig. 4.7. This shows that again the MARS algorithm, linear interpolation and cubic RBF are the best interpolation methods. The Gaussian method performs poorly, again this is due to it being a poor representation of the underlying function. It is also instructive to note that the error rapidly converges to a small value after approximately 100 model evaluations even for this complex function. This is a feasible number of models to evaluate if its evaluation time is not trivial and suggests that this process could work for complex response functions that may be present in reality.

As the performance of an interpolator is dependent upon how well it represents the underlying function, each of these methods will be suited to different scenarios. Given that the time taken to build a single instance of these interpolators is often significantly less than the time to run a model, it is worth trying all of these methods to determine which is the best. The best method will also reduce the number of model evaluations required to achieve a predictive error within tolerance therefore testing all three interpolation methods is the most efficient overall approach.

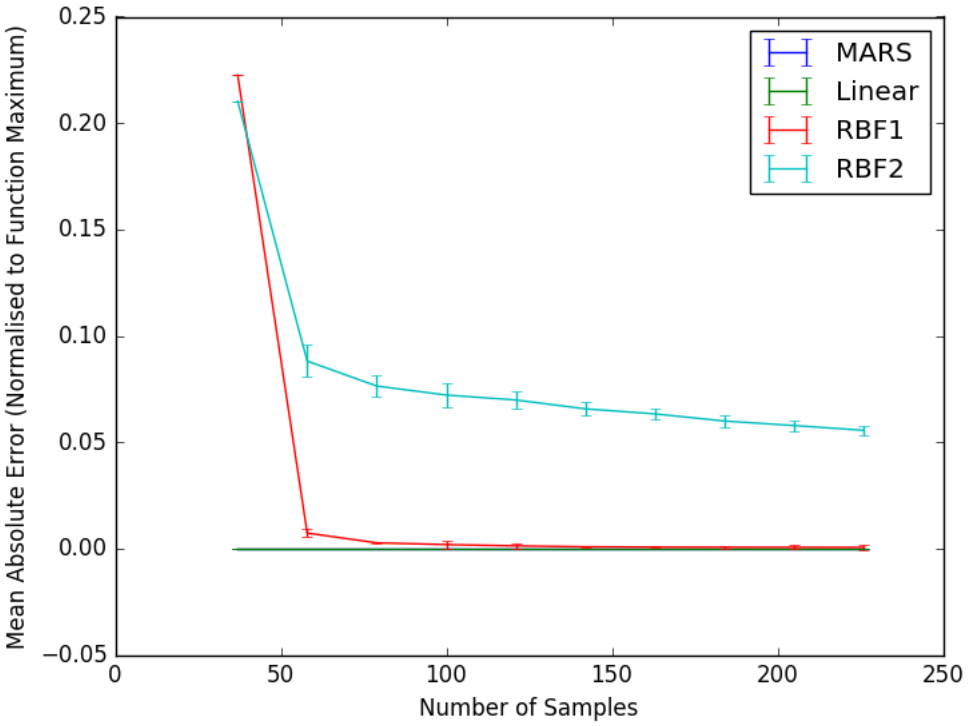

FIGURE 4.6: The result of applying the sampling and interpolation algorithm to the linear example function defined by Eqn. 3.19 with $a = b = c = d = 1$. Different interpolation methods are shown: Multivariate Adaptive Regression Splines (MARS), linear interpolation, cubic radial basis functions (RBF1) and Gaussian radial basis functions (RBF2). The results of the MARS algorithm are coincident with the results of the linear interpolation therefore both lines are not visible.

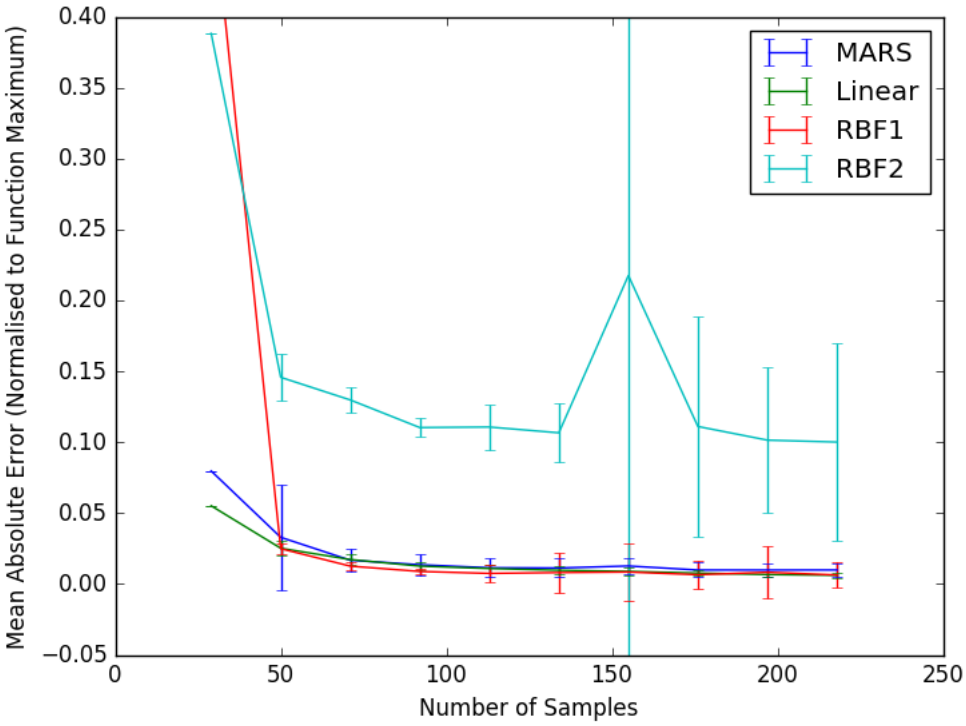

FIGURE 4.7: The result of applying the sampling and interpolation algorithm to the linear example function defined by Eqn. 3.20. Different interpolation methods are shown: Multivariate Adaptive Regression Splines (MARS), linear interpolation, cubic radial basis functions (RBF1) and Gaussian radial basis functions (RBF2).

Whilst these methods work, it could be more efficient to adaptively sample Ω in regions where the error is greatest, predominately caused by the response function varying rapidly, and not further increase the density of sampling in regions where the error is low, because the response function varies slowly. This is not easy to implement with LHDs therefore an alternative method is discussed in the next section.

4.1.7 Adaptive Sampling and Interpolation

Latin Hypercube sampling provides good coverage of the sample points throughout Ω however this may not be the best approach if the response function has regions where it is slowly varying and regions where it is changing quicker. In this case, a lower density of sample points will be sufficient in smoother regions of the response function however denser sampling may be necessary in regions with coarser features to capture the variations. It should be noted that the aim is to map the response function independently of the choice of probability function. Whilst sampling preferentially in regions of Ω where the probability density is highest may result in a more accurate assessment of reliability, if the probability function changes then the result may be less accurate. As discussed previously, one major advantage of the methodology is that the response function is decoupled from the probability function, therefore changing one has no effect on the other. This property should ideally be retained when mapping the response function so that changes in the probability function have no effect on the quality of the prediction of the response function and thus an impact on the calculation of inspection metrics such as PoD and PFA.

A method of generating sample points adaptively is the spatially adaptive sparse grid method. Sparse grids can be traced back to Smolyak [78]. However, were not regularly used until an efficient computational method was developed to apply them to the solution of partial differential equations [79]. They have subsequently been applied to high-dimension function approximation [80]. The properties of sparse grids are summarised here and the reader is directed to [81] for a detailed mathematical description of their construction.

The generation of sparse grids is somewhat similar to the recursive splitting used in the MARS algorithm. In this sense, each point in the sparse grid represents a subspace which may be further split by adding in a pair of new

points. The recursive splitting of these subspaces increases the density of sampling in a given region. Strictly speaking, they are complete hierarchical grids that are formed of a set of hierarchical subspaces whose tensor product forms a complete basis for the parameter space. Each subspace is constructed using a set of hierarchical basis functions, truncated using hat functions to limit the domain they act upon. Adaptive sparse grids are an extension of this in which, when moving from one level of the hierarchy to a lower one, not all child points are added to the grid. The effect of this is to create a hierarchical grid that is not fully populated and instead sample points are added based upon a criterion. In this case of interpolation, this criterion is based upon the form of the fitted function. Sparse grids form interpolators by fitting a sum of basis functions to the data, that is

$$R'(\Omega) = \sum_{i=1}^{N} \alpha_i \phi_i(\mathbf{x}_i), \tag{4.19}$$

where N is the number of sampled data points, α is a weighting factor and ϕ is a basis function. The choice of which points to add is based on hierarchical surpluses, that is the value of the weighting factors α_i. As the number of points added to a subspace increases, the weight of each basis function will decrease, more rapidly if the gradient is smoother in that subspace. Therefore, these weighting factors are a measure of the rate of change of the gradient of the function in the subspace. The addition of more points to a region will provide a better approximation of the function in that subspace, thereby decreasing the weights of the other basis functions assigned to that subspace. Therefore points are preferentially added to subspaces with larger weights for their basis functions, thus adaptively sampling the parameter space using information of the response function. This process is known as surplus refinement.

The application of adaptive sparse grids is demonstrated using an example twin peaked Gaussian function as shown in Fig. 4.8. The sparse grid created to map this function is also shown in Fig. 4.8. This shows that the adaptive method preferentially samples at regions where the function gradient is changing most rapidly whilst yielding a low density of sampling in regions where the function is slowly varying. This demonstrates the power of the technique as it is able to sample in regions where the predictive error is likely to be largest.

The primary issue with the use of sparse grids is the condition that all of a point's parents are known. This can lead to considerable wasted effort mapping the parent points which may exist in regions where there is already dense sampling. This is demonstrated in Fig. 4.8 in which it can be seen

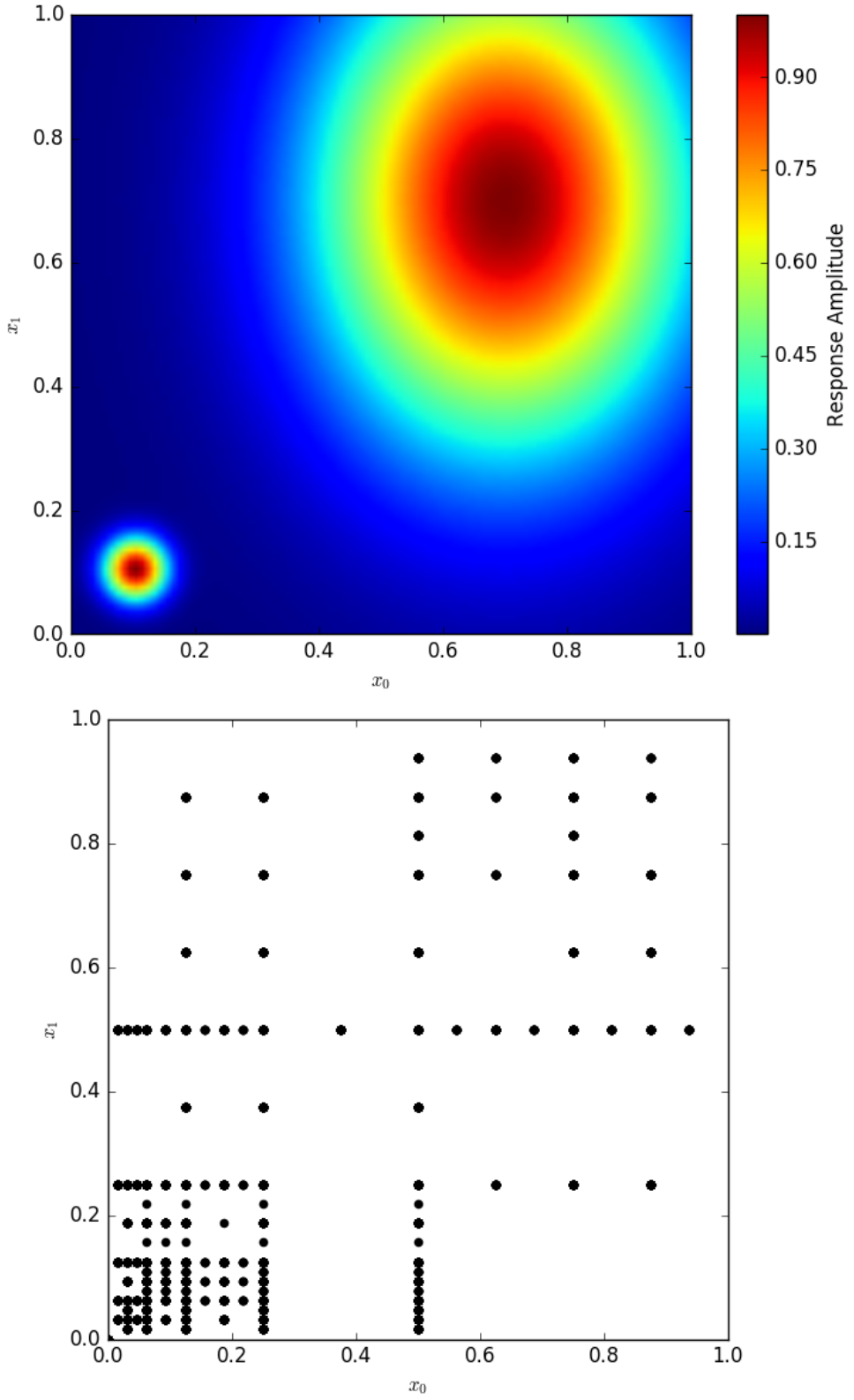

FIGURE 4.8: An example of the application of sparse grids to an interpolation. The map of the response function is shown (top) with the location of the sample points (bottom).

that there is excessively dense sampling around $(0.5, 0.1)$ and $(0.1, 0.5)$ even though the function is smooth in these regions. In the case where the numerical model has a significant run time, this can result in a large time requirement for the addition of a single point. Another problem is that it is sensitive to the choice of basis function and given that the choice of basis function determines the location of sample points on the grid, changing these will likely result in a change of sample locations and thus a re-evaluation of the model at these points. Therefore given that this method can potentially be very time consuming to use if the choice of basis function is not appropriate, it should be used with caution, as noted in [82].

Two sparse grid libraries were used in this work, SGPP [83] and TASMANIAN [84]. The way sparse grids are built and stored using in the two libraries are incompatible with each other, therefore it is not possible to interchange grids between them. Of these, SGPP has a limited, incomplete Python wrapper whereas TASMANIAN has a fuller library of functions. SGPP was initially used however its limitations were such that an alternative was sought and TASMANIAN was integrated into the tool chain.

4.2 Parameter Reduction Using Sensitivity Analysis

As previously discussed, being able to ignore parameters due to insignificance can reduce the dimensionality of Ω and thus reduce the time it takes to map $R(\Omega)$. Sensitivity analysis is again a broad topic that is widely researched and the desire is to utilise well documented methods to quantitatively determine the relative significance of parameters. This section discusses the application of the method of Sobol for calculating sensitivity indices to inspections and how these can be used to minimise the number of parameters that have to be considered.

4.2.1 Sobol Sensitivity Indices

There are a wide range of sensitivity metrics that can be used to assess the relative importance of parameters, for a good overview see [85]. Of these

methods, the method of Sobol is very intuitive and provides an efficient computational method of calculating indices which represent the relative importance of parameters.

The following is taken from [85] and [86] and is the definition of the sensitivity indices used in this work. For brevity, only the computational calculation of these indices using Monte Carlo integration is discussed and the analytic derivation of the following can be found in [87]. Throughout this work, a hatted value, for example $\hat{D}$, indicates the approximation of a value which has been calculated through Monte Carlo methods.

The primary idea of Sobol sensitivity indices is to determine the proportion of the total variance of the function $f(\mathbf{x})$ that can be attributed to each of the n input parameters and the interactions between them. This is a similar idea to the method of calculating the relative importance of parameters used in the MARS algorithm discussed earlier. This is based on the notion that it is possible to express a function as a linear sum of functions of increasing numbers of its n parameter arguments, that is

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n-1}\sum_{r>i}^{n} f_{ir}(x_i, x_r) + ... + f_{1,2,...,n}(x_1, ..., x_n). \quad (4.20)$$

This is known as the Analysis of Variance (ANOVA) representation of the function, the same principle used in Eqn. 4.16. The average value of f can be calculated as

$$\hat{f}_0 = \frac{1}{N}\sum_{q=1}^{N} f(\mathbf{x}_q), \quad (4.21)$$

where the function f has been sampled at N discrete points. The total variance of f can be estimated as

$$\hat{D} = \left(\frac{1}{N}\sum_{q=1}^{N} f^2(\mathbf{x}_q)\right) - \hat{f}_0^2. \quad (4.22)$$

The variance that can be attributed to the first order effect of a parameter, that is its main contribution in $f_q(x_q)$, can be estimated as

$$\hat{D}_i = \left(\frac{1}{N}\sum_{q=1}^{N} f(\mathbf{x}^{(1)}_{(\sim i)q}, x^{(1)}_{iq}) f(\mathbf{x}^{(2)}_{(\sim i)q}, x^{(1)}_{iq})\right) - \hat{f}_0^2, \quad (4.23)$$

where $\mathbf{x}_{(\sim i)}$ denotes the vector defining all coordinate values except that for

parameter x_i. The superscripts (1) and (2) denote two different choices of the values of the variable(s). Therefore each iteration of the summation uses pairs of function values at two points with the same value of the parameter x_i but different values of all other parameters $\mathbf{x}_{(\sim i)}$. It is convenient to compare the variance of a single variable to the total variance of the function therefore a sensitivity index for the parameter x_i can be defined as

$$\hat{S}_i = \frac{\hat{D}_i}{\hat{D}}. \tag{4.24}$$

This is the first order sensitivity index for parameter x_i. The sensitivity indices can similarly be calculated for higher order interaction terms by calculating $\hat{D}_{i,...,n}$ for any combination of parameters $i, ..., s$. Explicit formulas for these calculations are derived in [86] and are summarised here. Define

$$\overline{D_{i,...,n}} = \frac{1}{N}\sum_{q=1}^{N} f(\mathbf{x}^{(1)}_{(\sim i,...,n)q}, x^{(1)}_{(i,...,n)q}) f(\mathbf{x}^{(2)}_{(\sim i,...,n)q}, x^{(1)}_{(i,...,s)q}). \tag{4.25}$$

This again uses pairs of function values which have common values of the parameters $x_{i,...,n}$ and different values of the parameters $x_{\sim(i,...,n)}$. It can be shown that

$$\hat{D}_{i,...,n} = \overline{D_{1,...,n}} - \sum\nolimits^{>} \overline{D_{1,...,n-1}} + \sum\nolimits^{>} \overline{D_{1,...,n-2}} ... (-1)^{n-r} \sum\nolimits^{>} \overline{D_{1,...,r}} + (-1)^n \hat{f}_0^2, \tag{4.26}$$

where $\sum^{>} \overline{D_{1,...,r}}$ denotes the sum over all permutations of size r of the parameters in $i_1, ..., i_s$. This significantly reduces the total computation as it allows each $\overline{D_{1,...,n}}$ to be calculated then each $\hat{D}_{1,...,n}$ to be calculated using combinatorics by finding every combination of parameters. It is however more useful as a first screening method to consider the total effect of a parameter. This can be calculated by summing all $S_{1,...,i,...,n}$ terms which include parameter x_i. However, the explicit calculation of all the sensitivity indices involving a given parameter becomes increasingly computationally expensive as the number of parameters and thus the the number of higher order interaction terms increases. Noting that all sensitivity indices must sum to 1 by definition, the total sensitivity index for parameter x_i can be calculated as

$$\hat{T}_i = 1 - \frac{\hat{D}_{(\sim i)}}{\hat{D}}. \tag{4.27}$$

The calculation of $\hat{D}_{(\sim i)}$ can be performed in a single Monte Carlo calculation as

$$\hat{D}_{(\sim i)} = \left(\frac{1}{N} \sum_{q=1}^{N} f(\mathbf{x}_{(\sim i)q}^{(1)}, x_{iq}^{(1)}) f(\mathbf{x}_{(\sim i)q}^{(1)}, x_{iq}^{(2)}) \right) - \hat{f}_0^2. \tag{4.28}$$

This involves using function values at pairs of points with a common value of all parameters except x_i and varying x_i.

4.2.2 Calculation of Sensitivity Indices During Parameter Space Mapping

The above describes the general methods of calculating sensitivity indices however this still requires the definition of the sample locations at which to evaluate the function. The LHD algorithm is suited to calculating the first order sensitivity indices as the use of multiple samples at each value of each parameter but with different values of all other parameters, as required to avoid repeated sampling, satisfies the requirements of Eqn. 4.23. After n_I iterations of a design with N_P points there will be $N_P^{n_I-1}$ pairs of points which can be used to compute the first order sensitivity indices without using an interpolator to generate more response values from these sampled points. The accuracy of this process is tested by applying it to Eqn. 3.19 with $[a, b, c, d] = [1, 2, 3, 4]$ and Eqn. 3.20. The sampling process is the same as is used in the comparison of interpolation methods using a LHD of 21 points repeated 28 times. The results of this are shown in Fig 4.9. The large error bars show that this small number of samples is not sufficient to consistently obtain an accurate measure of the sensitivity indices. The decreasing size of the error bars as the number of sample points increases suggests that a much greater number of response values is necessary to obtain an accurate measure of the sensitivity indices. Performing greater numbers of response evaluations would alleviate this problem, however, this would come at a much greater time cost. An alternative is to use the interpolator to generate the necessary response function values to calculate the indices.

The use of an interpolator to generate the necessary points for the calculation can be tested by application to the complex example response function given by Eqn. 3.20. The simple example response function is not used here as the linear interpolation and MARS algorithm provide no error in the prediction against which to compare. For clarity, only the linear interpolation method was

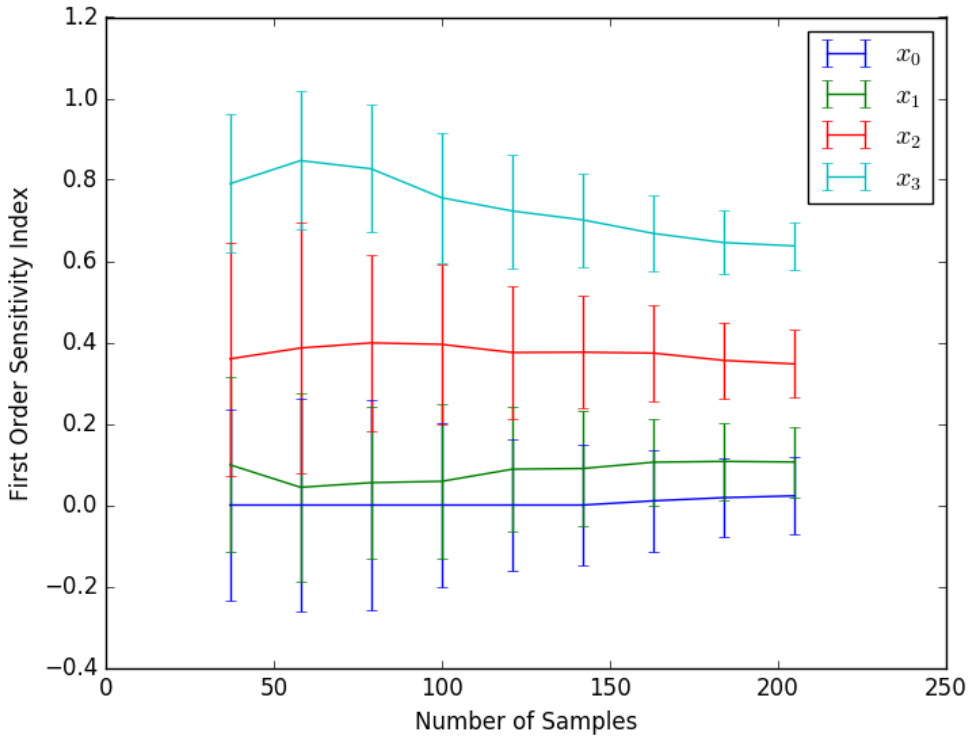

FIGURE 4.9: The calculation of the first order sensitivity indices using combinations of sampled points.

used and 50000 randomly generated pairs of samples were used to calculate the indices. This process was again repeated 28 times with different sets of LHDs to generate the interpolator to obtain a measure of the repeatability of this process. The results are shown in Fig. 4.10 which show that both the first and total sensitivity indices converge rapidly, they do not change significantly as the number of models evaluated increases. The small error bars in these graphs demonstrate the repeatability of the calculation which shows that it is not susceptible to the choice of sampled points. The indices as a function of the error in the prediction is shown in Fig. 4.11 which demonstrates little variation in the calculation of the indices as the error of the prediction decreases. This suggests that this process is capable of accurately estimating the sensitivity indices as the parameter space mapping process proceeds, making it a useful tool for potentially discounting parameters during the qualification. The accuracy of these calculations can be further improved through the use of low-discrepancy quasi-random sequences, such as Sobol sequences [88], to define the sample locations to be used in the interpolator. These types of sequences result in an increased rate of convergence of the calculations due to providing more uniform coverage of the space than random sampling. The result of this is that Monte

Carlo integrals performed using N_{MC} points converge as N_{MC}^{-1} using Sobol sequences compared to $N_{MC}^{-\frac{1}{2}}$ for random sampling. The reader is directed to [59] for further details.

It should be noted that the first order sensitivity indices are approximately equal to the total sensitivity indices. This suggests that, despite Eqn. 3.20 appearing to have significant interaction between the parameters, the response can be approximated to a high degree as a sum of independent functions. This potentially allows parameters to either be ignored due to having little significant or be treated as independent if they have negligible interaction terms. The former of these is especially important as, as well as potentially reducing the dimensions of the parameter space, for parameters whose inclusion may introduce significant modelling complexity, for example requiring a model to utilise another module or calculation algorithm, their removal may greatly reduce the qualification time.

4.2.3 Reduction By Insignificance

In the case that $T_k \approx 0$, the parameter x_k has very little relative effect on the outcome of the inspection. Therefore it is possible to ignore this parameter. In practice, it is not possible to simply remove a parameter from an inspection, it must be fixed to a given value δ such that

$$R(\mathbf{x}) = R(\mathbf{x}_{\sim k}) + R(\mathbf{x}_{\sim k}, x_k = \delta). \tag{4.29}$$

In this case a small error will be introduced into the calculation of metrics such as PoD. This is discussed in more detail later.

4.2.4 Reduction by Independence

It is possible that $T_k \approx S_k$. In this case, all of the second order and higher interaction terms are negligible, that is

$$\sum S_{k,\ldots,n} \approx 0. \tag{4.30}$$

If this is true then it is possible to re-express the response function as

$$R(\mathbf{x}) = R(\mathbf{x}_{\sim k}) + f_k(x_k), \tag{4.31}$$

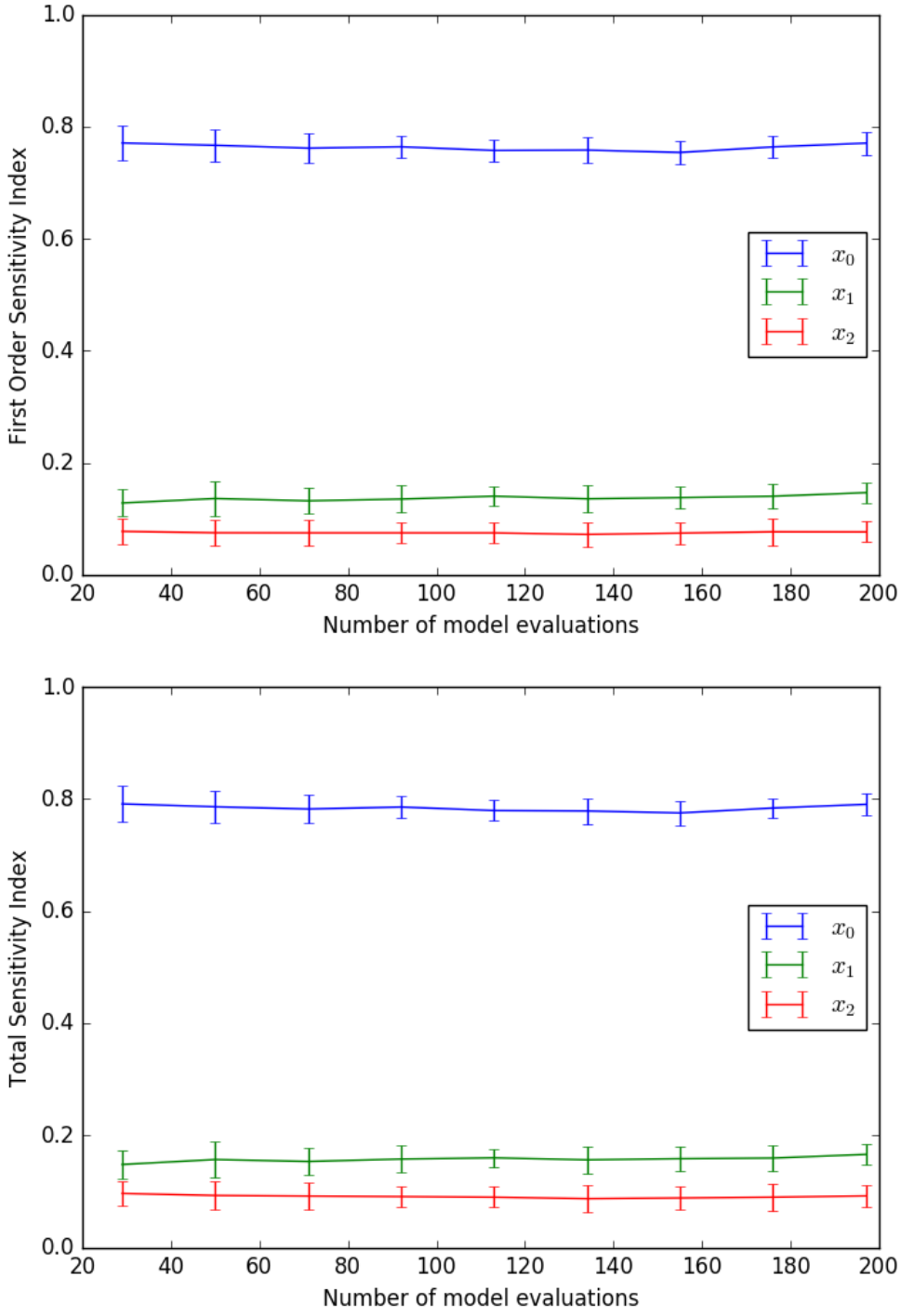

FIGURE 4.10: The (top) first order and (bottom) total sensitivity indices as a function of the number of sampled points.

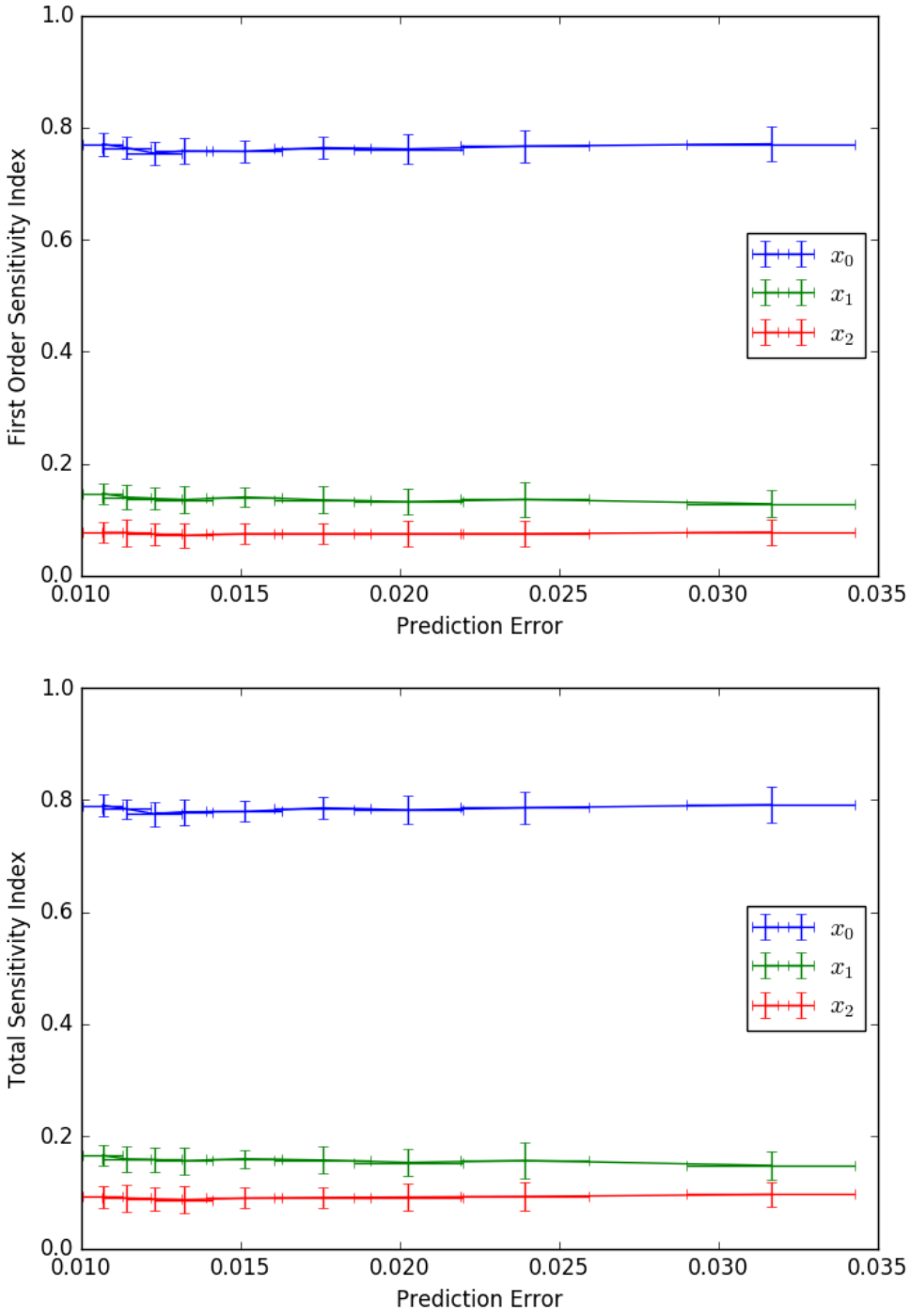

FIGURE 4.11: The (top) first order and (bottom) total sensitivity indices as a function of error in the prediction. The error in the prediction is normalised to the maximum function value.

where $f_k(x_k)$ is the first order function of variable x_k. This implies that the contribution to the total response of parameter x_k is independent of all other parameters to a degree of approximation. It is therefore possible to fix all of the other parameters and vary x_k to empirically build an explicit function for this parameter. This parameter may then be removed from the parameter space mapping process and included in the calculations of metrics of an inspection, due to its independence, via the methods discussed in Chapter 3.6. This still leaves open the question of how good the approximation $T_k \approx S_k$ has to be in order to apply this technique.

4.2.5 Optimal Parameter Fix Value

In the scenario where a parameter x_k is determined to have exactly no effect on the response or determined to be exactly independent, the ignoring of x_k will lead to no error in the calculation of reliability metrics. In practice, even if this is not exactly true, it may may be desirable to ignore some parameters as they account for little variance to reduce the dimensionality of Ω and thus reduce the mapping time. It is also possible that a parameter may have a small but significant effect however its inclusion in the model may cause a significant increase in its evaluation time, such as considering non-linear effects. Therefore it may be desirable to ignore a parameter to satisfy a time or resource constraint. In either case, the fixing of a parameter to a value will introduce some error into the calculation of inspection metrics such as PoD. Express the response of the inspection as

$$R(\mathbf{x}) = R(\mathbf{x}_{\sim k}) + R(\mathbf{x}_{\sim k}, x_k) = R_{\sim k} + R_k, \tag{4.32}$$

where $R_{\sim k}$ is the sum of all functions not involving x_k and R_k is the sum of functions involving x_k. In practice, ignoring a parameter means that it is fixed to a value $x_k = \gamma$ and it is taken that it does not vary. The function R_k may be a function of the other parameters if the higher order sensitivity indices, $S_{k,\ldots}$ are non-zero thus will still vary as a function of the other parameters. Therefore the error in the calculations of metrics such as PoD depend upon how much the function $R(\mathbf{x}_{\sim k}, x_k)$ varies compared to when the parameter x_k is fixed, that is $R_{k=\gamma} = R(\mathbf{x}_{\sim k}, x_k = \gamma)$. Define $R_{k,\min}$ as the minimum possible value of R_k and $R_{k,\max}$ as the maximum possible value of R_k. When $x_k = \gamma$, $R_{k,\min} \leq R_k \leq R_{k,\max}$. For metrics such as PoD and PFA, when $R_{\sim k} + R_{k,\min} \geq T$ and $R_{\sim k} + R_{k,\max} < T$, the error in the value of the response has no effect

on the metric as it does not cause the decision threshold to be incorrectly exceeded or undercut. Therefore the error only effects metrics in the region $T - R_{k,\max} \leq R_{\sim k} < T - R_{k,\min}$. There are two possible errors that can occur, firstly when the response is incorrectly overestimated and causes it to exceed the threshold, a false positive, and when the response is incorrectly underestimated and causes it to be below the threshold, a false negative. The former of these is the region $\Delta_A : R_{\sim k} + R_k < T \leq R_{\sim k} + R_{k=\gamma}$ and the latter $\Delta_B : R_{\sim k} + R_{k=\gamma} < T \leq R_{\sim k} + R_k$. In the case of PoD, the most common capability metric, the total error ϵ_p across the range of the parameters of interest x_c, whose range is normalised to the interval [0, 1], is therefore

$$\epsilon_p = \int_{x_c=0}^{x_c=1} \left[\frac{\int_{\Delta_A} P(\mathbf{x}) d\mathbf{x}}{\int_{\Omega} P(\mathbf{x}) d\mathbf{x}} + \frac{\int_{\Delta_B} P(\mathbf{x}) d\mathbf{x}}{\int_{\Omega} P(\mathbf{x}) d\mathbf{x}} \right] dx_c. \tag{4.33}$$

This cannot be further simplified without explicit forms of the probability function or the response function. The optimal choice of value can, however, be investigated numerically using example test functions as follows. Consider Eqn. 3.19 with a being variable, $b = c = 1$ and $d = 2$. In this case the response function has the form

$$R_1(\mathbf{x}) = ax_0 + x_1 + x_2 + 2x_3. \tag{4.34}$$

The effects of these parameters on this function are all independent, that is $S_i = T_i$ for all i. The relative importance of the parameters depends upon the value of a. If $a < 2$ then x_3 is the most significant parameter whereas if $a > 2$, x_0 is the most significant parameter. Each parameter has an independent normal probability distribution about the centres $\mu_i = [0.25, 0.5, 0.75, 0.5]$ respectively, giving the probability function for the parameters $P(\mathbf{x})$ as

$$P(\mathbf{x}) = \prod_{i=0}^{3} e^{-\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2}, \tag{4.35}$$

where σ_i is the standard deviation for parameter i and is $[0.5, 0.25, 0.25, 0.75]$ for each of the four parameters respectively. In the case that $a = 0.5$, x_0 is the least significant parameter and thus would be the most likely candidate to be ignored. The effect of doing so can be investigated by calculating the mean absolute error of the PoD curve, as in Eqn. 4.33, for another parameter for different fix values, in this case x_3 as it is the most significant of the other parameters. The result of doing this is shown in Fig. 4.12 as a function of the fix value, the corresponding cumulative distribution function (CDF) value and the

corresponding probability density function value. This shows that the optimal fix value is at approximately the centre of the CDF for the fixed parameter, its mean value, not the maximum value of the PDF, its most likely value. For the normal distribution used for x_0, these do not coincide as the distribution is defined over a finite domain with its mean not in the centre of the domain. It also demonstrates the importance of choosing an appropriate value to fix a parameter at if it is to be ignored otherwise this can result in a significant error in the PoD. The effect of changing the value of a is shown in Fig. 4.13, demonstrating that the minimum error in PoD is consistently at the mean value of x_0 and that the minimum error increases as the value of a increases. The minimum error as a function of the total sensitivity index for parameter x_0 is plotted in Fig. 4.14 which shows a linear relationship between the total sensitivity index and the error in the calculation of the PoD. It demonstrates that it is possible to obtain a small error in the PoD whilst ignoring a relatively large proportion of the total variance of function, in this case an error of 5% whilst ignoring 20% of the total sensitivity.

This simple, purely additive function is not necessarily representative of the response function that may be present in an inspection therefore the same methodology is applied to the more complex function. It was previously demonstrated that the parameters in Eqn. 3.20 are approximately independent therefore would yield the same result as the above example. Given the need to demonstrate this effects for which the higher order terms of the function R_k will not be trivial, a more complex function is used. This is defined as

$$R_2(\mathbf{x}) = \frac{ax_0^2}{10} - (x_0 - x_1) + 2x_1x_2 + 2e^{(2x_3 - ax_0)^2}, \tag{4.36}$$

where a is again a variable weighting parameter. The sensitivity plot for this function with $a = 1$ is shown in Fig. 4.15. This function has significant interaction between the parameters and is not monotonic. The error in the PoD is calculated for the most significant parameter, x_3, using Eqn. 4.33. The error as a function of fix value, CDF and PDF is shown in Fig. 4.12 which suggests that the optimal fix value is not at the centre of the CDF value or at the most likely value, the maximum of the PDF. Therefore it is not trivial to set the parameter x_3 at its optimal value in this case. This is further demonstrated by plotting the minimum error as a function of changing the weighting factor, as shown in Fig. 4.16. This shows that the optimal fix value changes as the weighting factor changes, further evidence of the optimal choice being very dependent upon the specific response function. It also demonstrates that it is again possible

to achieve a very low error in the PoD whilst ignoring a reasonably important parameter, in this case inducing an error of only a few percent whilst ignoring more than 50% of the total variance. This is potentially significant if this could allow several parameters with low total sensitivity indices to be ignored, significantly reducing the response mapping time by reducing the dimensionality of the parameter space.

This suggests that the optimal fix value is not easy to determine analytically and is very dependent on the response function being mapped however it is possible to ignore a reasonably large proportion of the variance whilst still obtaining a reasonably accurate result. Given the difficult of choosing the optimal fix value, in practice, it is most conservative to set x_k such that $R_{k=\gamma}$ is minimised, thus minimising the region Δ_A and not overestimating the PoD. This gives the worst case of inspection capability as the PoD will never be overestimated. This may be estimated numerically, optimising the fixed value of the ignored parameter by minimising the error in the calculated PoD when the parameter is fixed compared to its unfixed value. This does assume that when this process is performed the estimate of the response function is accurate and the methods described previously have been demonstrated to be able to accurately estimate response functions. This is therefore a feasible process in practice.

4.2.6 Sensitivity Indices as a Metric of Inspection Quality

It is common for the PoD, or more specifically the 90/95 value known as the minimum detectable defect size, to be used as a metric of the quality of an inspection. This does not, however, take into account the false call rate of an inspection which may be just as important. Sensitivity indices also provide a way of comparing the quality of different inspection methods. In an ideal world, a quantitative inspection would be sensitive to only the parameter of interest x_j and varying any combination of all other parameters would have no effect on the response of the inspection, that is $S_j = 1$, $T_{i\neq j} = 0$ and $R(\Omega) = R(x_j)$. In practice, it may be desirable for the location of the probe to have an effect on the inspection to allow localisation of a defect although it is still desirable for the parameter of interest to have the most significant impact. As $S_j \to 1$, the PoD for any value of x_j whose mean value is less than the threshold tends to

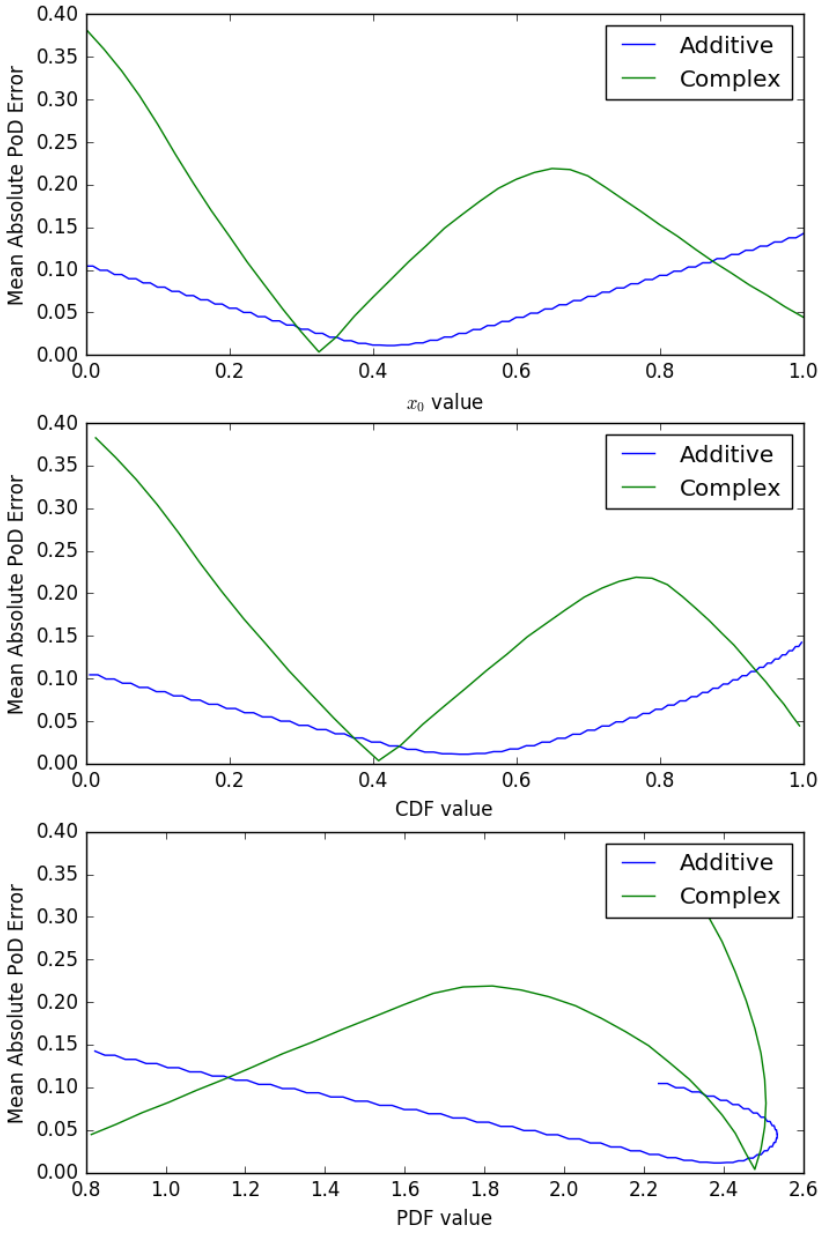

FIGURE 4.12: The error in the PoD value as a function of the value of the fixed parameter (top), the cumulative distribution function (CDF) of the fixed value (middle) and the probability density function (PDF) of the fixed value (bottom).

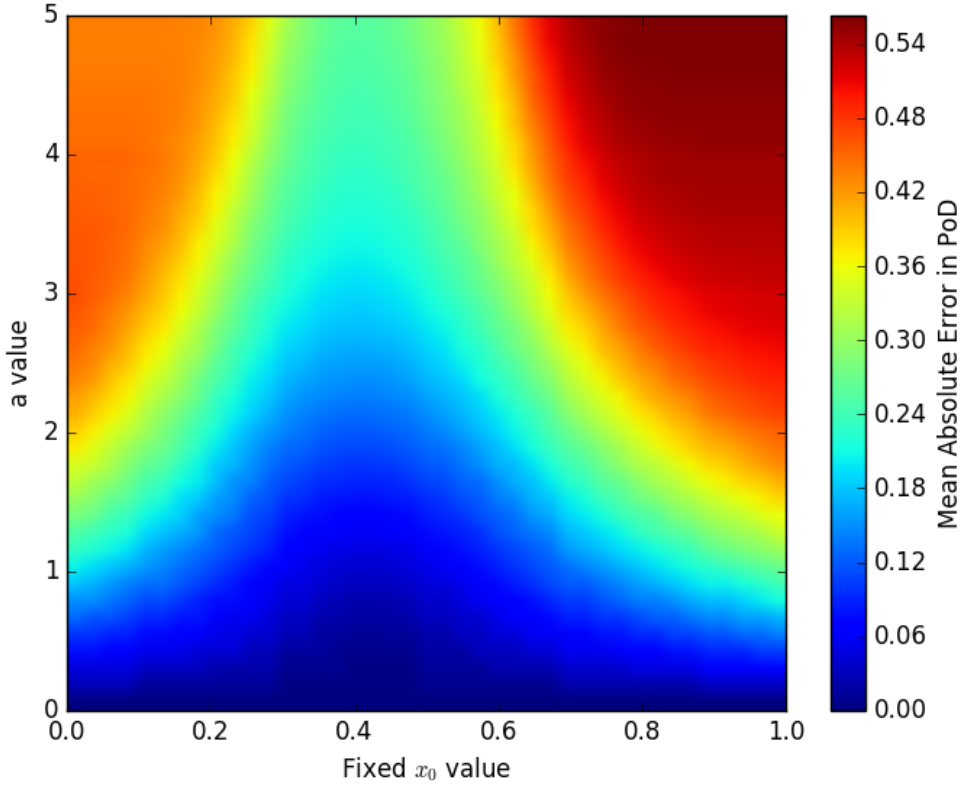

FIGURE 4.13: The error in the PoD as a function of the value at which the ignored parameter x_0 is fixed with the value of the weighting factor a changing, as defined in Eqn. 4.34.

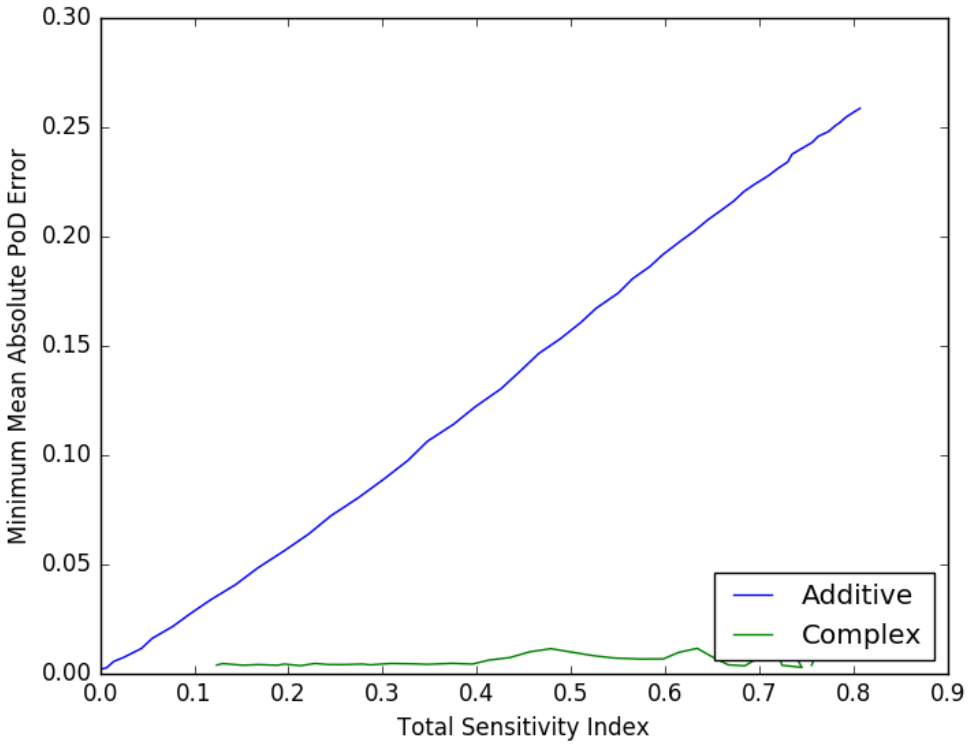

FIGURE 4.14: The error in the PoD value as a function of the total sensitivity index for ignored parameter.

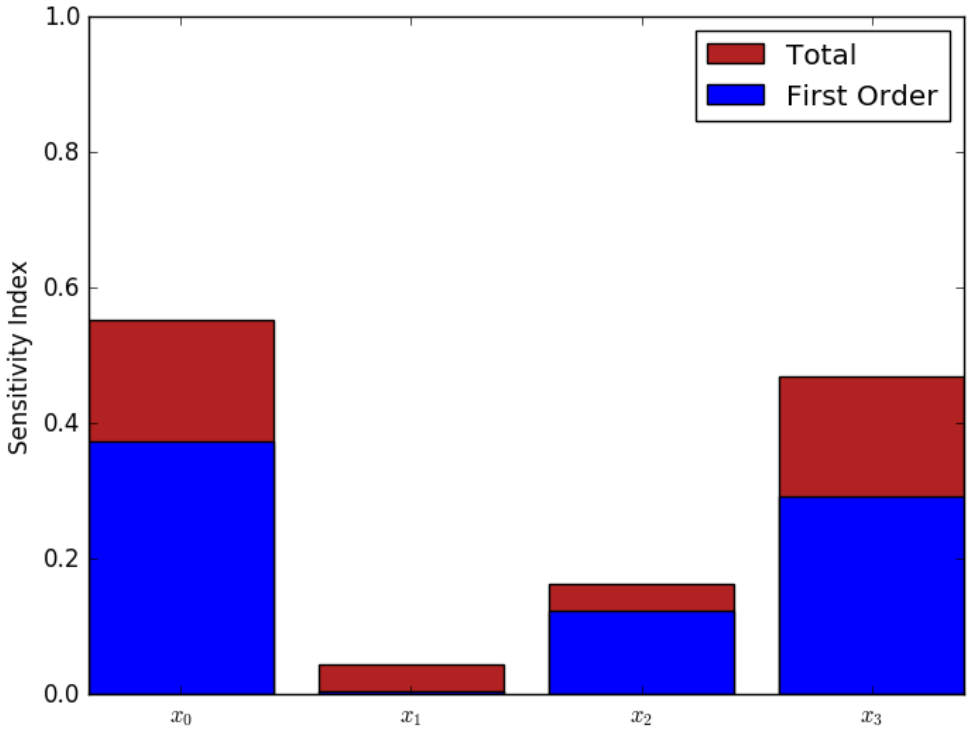

FIGURE 4.15: The sensitivity plot for the function given by Eqn. 4.36 with a=1.0.

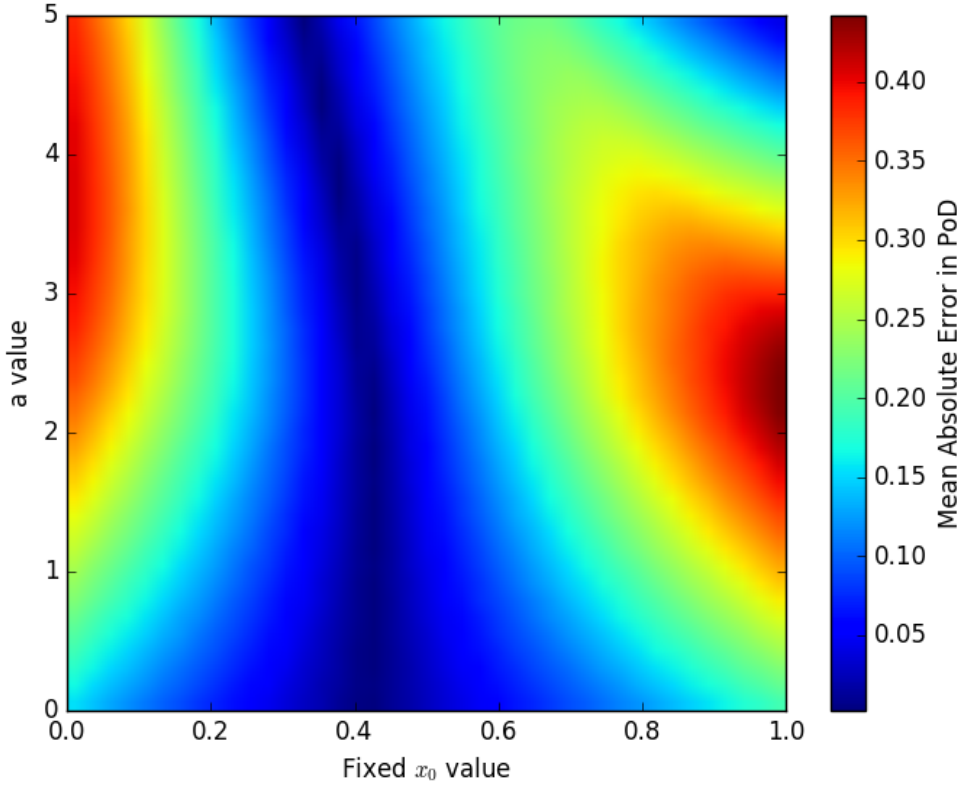

FIGURE 4.16: The error in the PoD as a function of the value at which the ignored parameter x_0 is fixed with the value of the weighting factor a changing, as defined in Eqn. 4.36.

zero and those whose mean is above tends to one. This results in a sharper PoD curve, becoming a step function for monotonically increasing $R(x_j)$, and thus a better inspection. This also reduces the false calls as, if the decision threshold on the response is set correctly, the probability of false alarm will decrease to zero. In the opposite case, as $T_j \to 0$, the PoD curve will tend to a constant value as changing the value of the critical parameter will have no effect on the response of the inspection, as $R(\Omega) = R(\mathbf{x}_{\sim j})$, and the outcome will be purely determined by the values of the other parameters.

An alternative way of thinking about an inspection is to consider the response of the inspection to be attributable to either the response from the defect or the response of noise in the system. In this case, the range of these responses can be used to build probability distributions of the response from the defect, G, and the response caused by other factors in the inspection which is noise, N. These will have a mean value and variance which can be derived for the distribution. For simplicity, express an arbitrary distribution d as a function of its mean and variance, that is $d(mean, variance)$. Considering a given value α of the defect parameter of interest x_j, the distributions of responses are $G = d(S_c, T_{j,x_j=\alpha} - S_j)$ and $N = d(N_c, 1 - T_j)$. Note that these are arbitrary probability distributions and that $T_{j,x_j=\alpha} < T_j$ if $T_j \neq S_j$. As $T_i \to 1$, the variance of these distributions tends to zero, that is the range of responses from either source decreases. As this happens, the responses will be easier to distinguish between signal and noise as there will be less overlap in the distributions if $N_c \neq S_c$. There will also be better characterisation of defects of different magnitudes as S_c will depend on the value of α. In practice, a good inspection will have $N_c < S_c$, otherwise the PFA is likely to be higher than the PoD, which is not a useful inspection.

As sensitivity indices can encapsulate information about both the PoD and PFA, it is a more useful metric than either on its own. It is also inherently both normalised and unit-less and it therefore ideal as a metric for comparing disparate inspections.

4.3 Summary

The accurate mapping of the response function is essential for accurately calculating metrics of inspection reliability. This chapter has discussed methods of doing this, such as using Latin Hypercubes and sparse grids coupled to an interpolation method, that will in theory allow the response function to be accurately mapped using a reasonable number of model evaluations. The use of sensitivity analysis can provide insight into the relative importance of parameters, highlighting those which can be ignored or treated as being independent. However, unless either of these conditions is exactly met, an error in the reliability metric will be introduced. The effect of this has been investigated and the results suggest that it may be possible to ignore a significant proportion of the total variance whilst inducing a reasonably small error into the calculation of the inspection metrics. This does require careful choice of the fix value of ignored parameters which may be performed numerically. The sensitivity index may also be used as a metric of inspection capability, allowing disparate inspections of the same defects to be compared. As well as information on the PoD, sensitivity indices also encode information of PFA, providing more information than either alone. The following section discusses the application of these methods to an example inspection qualification.

Chapter 5

Example Inspection Qualification

This chapter presents the application of the methods discussed in this thesis to a real inspection. A canonical example in aerospace is the inspection of cracks that emanate from fastener holes in wing skins. This provides a good example of an inspection on which to perform model assisted qualification as there are many parameters that can vary, both human factors and defect parameters. This inspection is also sufficiently complex to justify the use of a numerical model which has a significant computational burden. Whilst a single iteration of the model with a fixed set of parameters is reasonably straightforward to create, efficient automated model generation for parameter space mapping is not trivial and requires considerable work.

Initial work was done by Smith and Edgar [21] whilst the technique was under development at QinetiQ, where some experimental trials were performed to demonstrate the capability of the technique. This work was used as the basis of understanding of how the inspection is performed.

5.1 Definition of Inspection

A realistic representation of the inspection scenario is shown with dimensions in Fig. 5.1. It consists of a bolt hole in an aluminium wing skin from which a triangular shaped crack emanates radially. The dimensions and cord-wise rotation of the crack are variable parameters within the inspection, as shown in Fig. 5.1. This defect was inspected using a single element, unfocussed, 45° shear wave transducer with a centre frequency of 4 MHz. It is sentenced using a gated

Parameter	Parameter Type	Lower Limit	Upper Limit
Defect length c_L	Defect	1 mm	11 mm
Defect height c_H	Defect	0.5 mm	5.5 mm
Defect angle c_θ	Defect	-45^o	45^o
Probe lateral position p_x	Human	1 mm	11 mm
Probe perpendicular position p_y	Human	5 mm	15 mm
Probe rotation p_θ	Human	-10^o	10^o
Coupling thickness ϵ	Equipment	-20%	20%
Electrical noise η	Equipment	-5%	5%

TABLE 5.1: The six numerically modelled parameters and their ranges. The values of the probe's lateral and perpendicular position are relative to the centre of the hole.

threshold, that is measuring the maximum amplitude of the Hilbert transform of the signal in a fixed time gate whose position is determined by the operator. In this inspection the time gate is set such that its centre is at the time of arrival of a corner reflection from the base of a crack and its width is set to $\pm 10\%$ of this arrival time. A 3D finite element model was built and solved using Pogo FE [89]. Each model required approximately 20 minutes to evaluate, including pre-processing for the model including the generation of the mesh, and analysing the resulting time trace. Each model also required approximately 8 GB of GPU RAM therefore it is possible to evaluate only a single model at a time on the available hardware. In all, 8 parameters of variation are considered. These are summarised, along with their limits, in Table 5.1. This chapter demonstrates the application of methods for assessing reliability presented in this thesis to increasing numbers of these parameters thereby increasing the dimensionality of Ω.

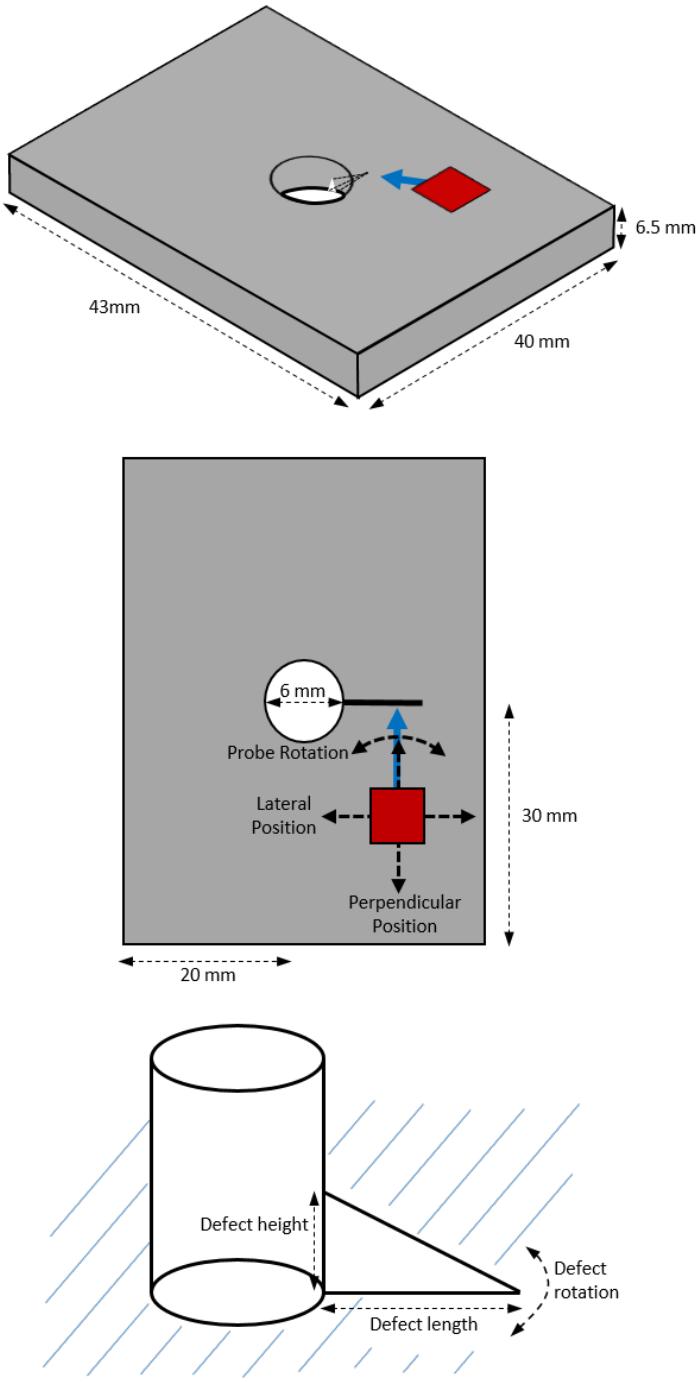

FIGURE 5.1: Schematic diagrams of the inspection investigated from an isometric view (top), plan view (middle) and a diagram of the crack emanating from the hole (bottom). The probe is shown by a red square and the ultrasonic beam by a blue arrow. In this inspection, the operator scans the probe around the fastener hole to check all possible root positions of the crack. Given the rotational symmetry of the inspection, only one crack root position is simulated.

5.2 Finite Element Modelling Using Pogo

The fundamental mechanics of elastic media when subject to an elastodynamic excitation are well established, for an overview see [90]. The primary challenge with the application of these formulae is solving them for complex scenarios, such as intricate geometries. This typically cannot be achieved analytically and in practice numerical methods are used. The most common approach is to use the finite element method to discretise a volume and numerically solve the equations on this mesh. The reader is directed to [91] for a good overview of the finite element method used in ultrasonic modelling.

The advent of general purpose graphical processing units (GPGPUs, henceforth referred to as GPUs) in the last decade, which can have thousands of cores, allows for massive parallelisation of this calculation using a desktop PC. Pogo FEA [89] uses nVidia's CUDA framework to achieve this parallelisation, accelerating calculations by a factor of 10-1000 times over CPU based explicit solvers such as Abaqus (Dassault Systems, Vlizy-Villacoublay France). For brevity, the details of how this is achieved are not covered and the reader is directed to [89] for details of the implementation. In its current form, Pogo consists of the solver along with other programs required to set up the problem on the GPU, using a non-proprietary binary input file format to define the models. This is therefore well suited to evaluating many models as these input files, and associated output files, can be generated and analysed using any scripting language so desired.

The main limitation of Pogo is that presently it does not support acoustic mesh elements therefore from an ultrasonic NDT perspective it is not able to simulate immersion inspections. It has been used for simulations of ultrasonic inspections using bulk waves, for example [92], thus it is suited to the demonstration inspection that for this project. This section discusses the use of Pogo as a tool for performing MAQ of ultrasonic inspections, particularly methods of optimising its use for efficient parameter space mapping.

5.2.1 Practical Considerations

Pogo, like most parallelised calculation software, is primarily limited by the amount of memory bandwidth on the bus between CPU and GPU. Therefore a GPU with high memory bandwidth is necessary to achieve good performance.

The nVidia GeForce cards, primarily aimed at gaming rather than numerical computing, have high memory bandwidth and relatively low cost, making them a good choice for Pogo. A desktop PC was purchased consisting of an Intel Core i7-6700 (4 cores, 8 threads), 32 GB of DDR4 RAM and a nVidia GeForce Titan X GPU, which has 3072 cores, 12 GB of RAM and a memory bandwidth of 480 GB/sec. A PCI-E solid state drive was also installed to allow rapid reading and writing of the files created in the generation of models. Typically this could be of the order of several gigabytes for a large model and can be a significant proportion of the total simulation time, therefore minimising the data transfer time is essential. This system cost approximately £2200 and was used for all Pogo simulations, making it a realistic option for a SME to use.

Given the potentially large number of simulations that will be run, automating the generation and running of models is essential. This is best achieved using a high level scripting language, such as Python, which allows the various programs necessary to run a Pogo model, such as the meshing tool, blocker and solver, to be called in sequence. Python therefore acts as the glue between these components and can also be used for the data analysis, sampling and interpolation. Python as a language is well suited to this work as it is designed to handle arbitrary objects, reading and writing of data and through the Scientific Python stack (SciPy, NumPy and MatPlotLib) it is ideal for analysing results and performing the necessary calculations required to generate a model. A significant advantage of Python over other popular scripting languages, such as MATLAB (MathWorks, Massachusetts USA), is that it is open source software, allowing any tools developed in this project to be easily utilised by other organisations without having to purchase expensive software licenses.

5.2.2 Automated Geometry Generation and Meshing

Given the large number of simulations that have to be completed, the automated definition of models is essential to mapping the response function in a reasonable time. The primary challenge in generating a model is the definition of the mesh and therefore the volume that is to be meshed. There is a large body of literature on the optimal generation of meshes and the advantages and disadvantages of different element types, for a good overview see [93]. Given that the use of cuboid (in 3D) elements can lead to incorrect reflection behaviour of angled surfaces [93] if they do not accurately map to a surface, tetrahedral elements were used regularly in this work as they are geometrically more flexible.

Again, given the significant research invested in optimal meshing algorithms, the software Tetgen [94] was used to generate meshes consisting of tetrahedral elements. This process is based upon performing Delauney triangulation of the convex hull of the volume to be meshed and performing iterative refinement based upon some termination conditions to generate the mesh. In ultrasonic FE calculations, typically the termination condition is when a desired number of nodes per wavelength has been achieved. Evidence suggests that approximately 10 elements per wavelength at the highest frequency is appropriate [93]. As the mesh coarsens and the average element volume increases, numerical inaccuracy arises which can cause incorrect predictions of the amplitude and velocity of waves in the model. Doubling the number of elements per wavelength halves the edge length of an element, increasing the number of elements in the model by a factor of 4 in 2D and a factor of 8 in 3D. As the edge length halves, the time step used in the calculation must also halve, resulting in an additional factor of 2 increase in the evaluation time. This leads to a significant increase in the run time of a single model and thus a large increase in the total time required to accurately map the response function. There is therefore a trade off between accuracy and evaluation time of a model.

This is further complicated by the fact that non-uniform meshes have a range of element sizes and the minimum distance between any two connected nodes determines the time step of the simulation, where a larger time step results in fewer time increments and thus a shorter simulation time. If the time step is too large then the model is numerically unstable and infinities arise in the output. Tetgen has the ability to refine a mesh until the volume of the largest element is below a desired value however, without modification to the source code, it is not possible to define a minimum value and this would require mesh coarsening within the Delauney triangulation framework. It was decided that this limitation would be acceptable given the effort required to either write a new meshing program or to modify these programs to accommodate a minimum element size. It should be noted that these programs have been highly optimised and are able to generate meshes quickly, significantly faster than the built in meshing routines of Abaqus.

A simple investigation was performed into the effect of element size on accuracy. In the context of this inspection, the use of a gated threshold requires an accurate estimation of the amplitude and arrival time of the reflected signals. The latter therefore requires accurate estimation of the wave velocity in the material. The simple model used is shown in Fig 5.2. This consists of a 10

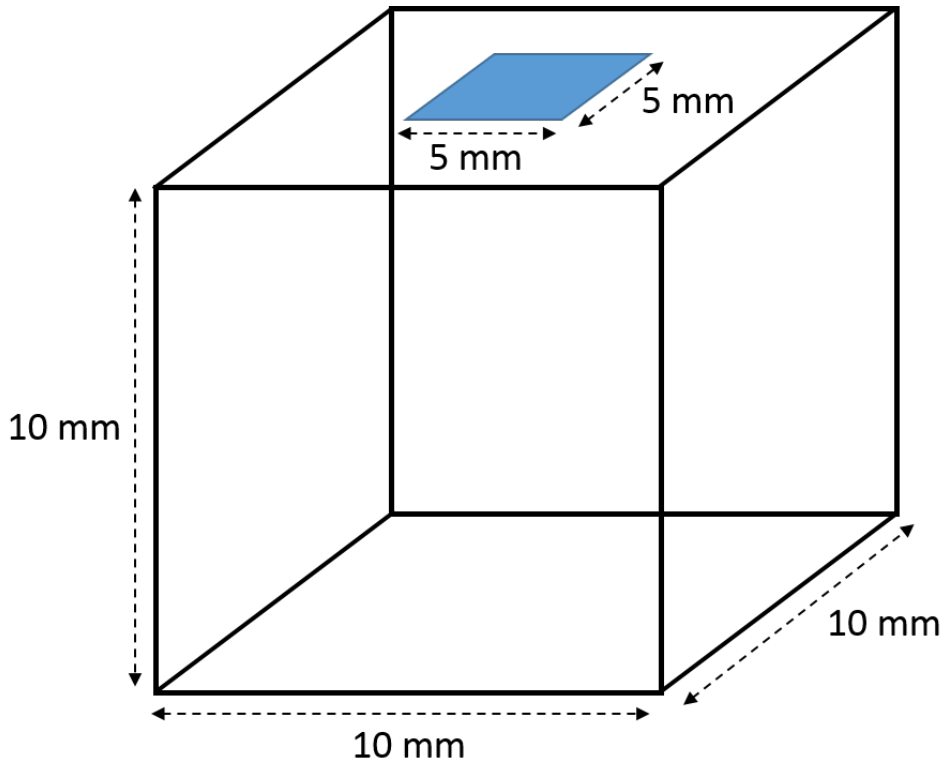

FIGURE 5.2: The geometry of the simple model used for the investigation into the effect of varying the volume of elements within the mesh. It consists of a 10 mm cube of aluminium (black) with a square element transducer of side 5 mm on top (blue).

mm^3 cube of aluminium with a 5 MHz, 5 cycle Gaussian tone burst input into the cube. A square transducer element was used of dimension 5 mm and the model time was approximately 6 μs to allow for a reflection off the back wall with some extra time to allow for variations in the ultrasonic wave velocity. Ultrasonic waves are induced into the model through the displacements of the nodes in the transducer footprint. These displacements are perpendicular to the surface and their amplitude varied to induce a sinusoidal wave. The other dependant properties of the model, such as the time step, are calculated in the generation of the model as these are a function of properties of the mesh.

The amplitude of the first reflection as a fraction of the amplitude of the input pulse is shown in Fig. 5.3 which demonstrates that the coarseness of the mesh has a significant effect on the response. It also has an effect on the velocity of the wave in the model, and therefore the arrival time of the reflection, as shown in Fig. 5.4. A coarser mesh reduces the measured velocity of the longitudinal wave in the model, with the velocity converging as the mesh is refined. Clearly a minimum value of the maximum element volume is required

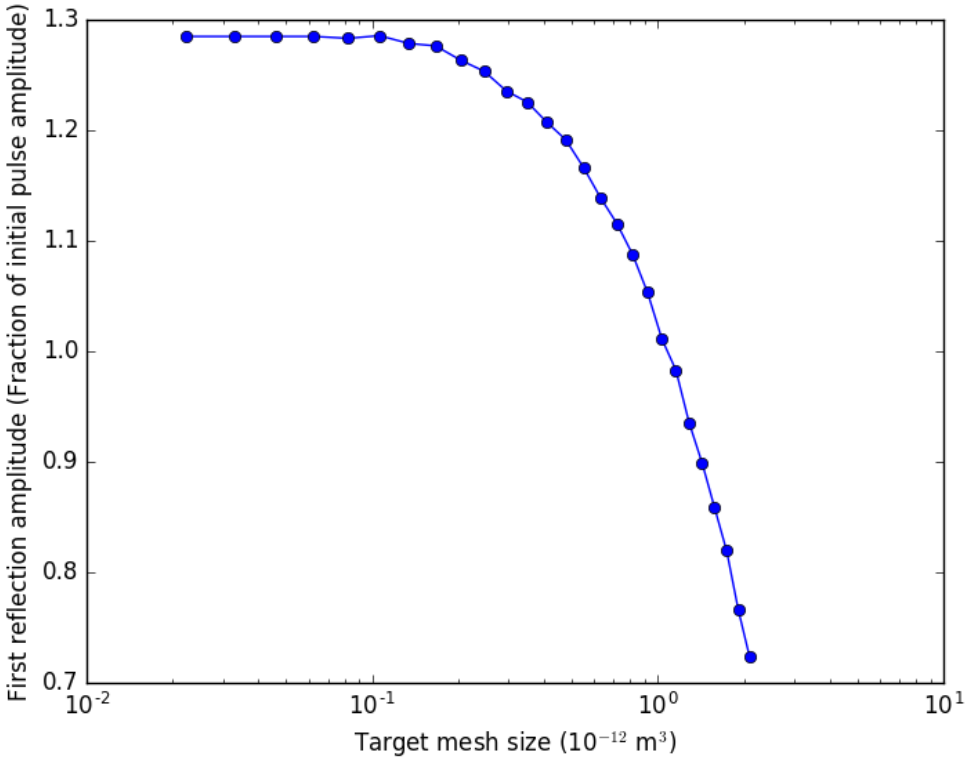

FIGURE 5.3: The amplitude of the first reflection from the back wall of the sample as a fraction of the initial amplitude, as a function of the target mesh size.

to achieve this convergence in both amplitude and velocity however refining the mesh beyond this point provides no further increase in accuracy whilst increasing the evaluation time of the model, as shown in Fig. 5.5. In the context of inspection qualification, this sweet spot should be aimed for to both minimise the time required for qualification whilst maintaining the accuracy of the result. In this case, the result of over-refining the mesh could result in an increase in the run time of a factor of four which, if many hundreds of models are evaluated, could potentially result in days of wasted effort. Both of these measurements converge at approximately 10^{-13} m^3. Therefore an aluminium sample with a 5 MHz input frequency requires approximately 10^4 elements per mm^3 for accurate modelling, or approximately 21 nodes per mm, thus this 1000 mm^3 sample requires approximately 10 million elements. In practice, real specimens will be significantly larger than this therefore will require on the order of tens of millions of elements.

The major bottleneck in this stage of generating the model is in the generation and transfer of the mesh information, the node locations and element assignments, into a Pogo input file. In practice a useful mesh may be of the

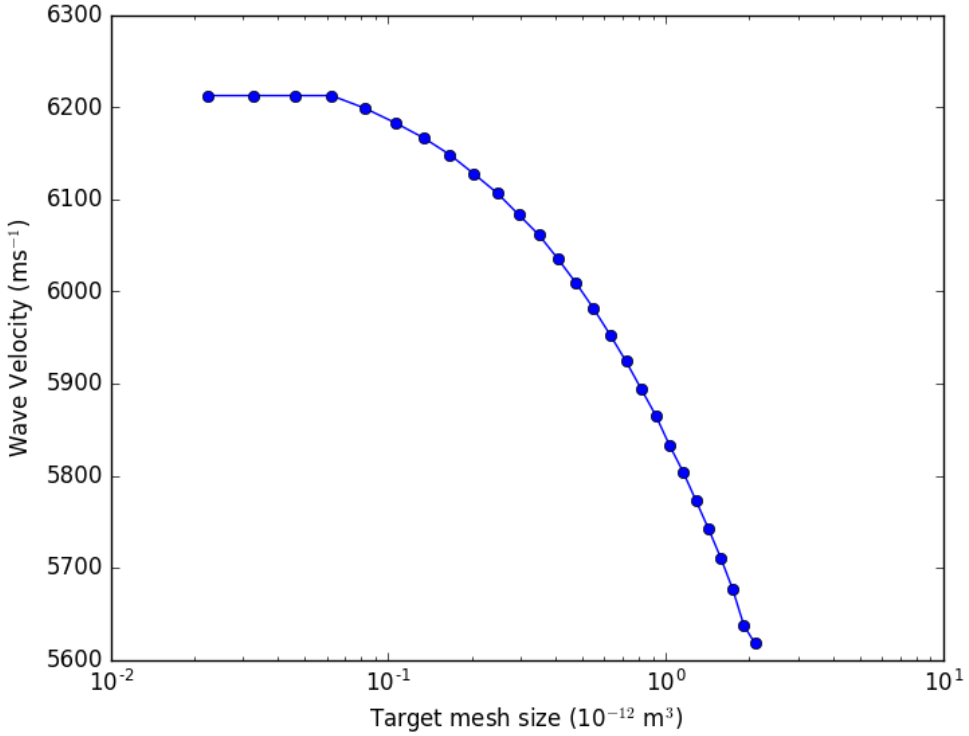

FIGURE 5.4: The wave velocity of the longitudinal wave in the model as a function of the target mesh size.

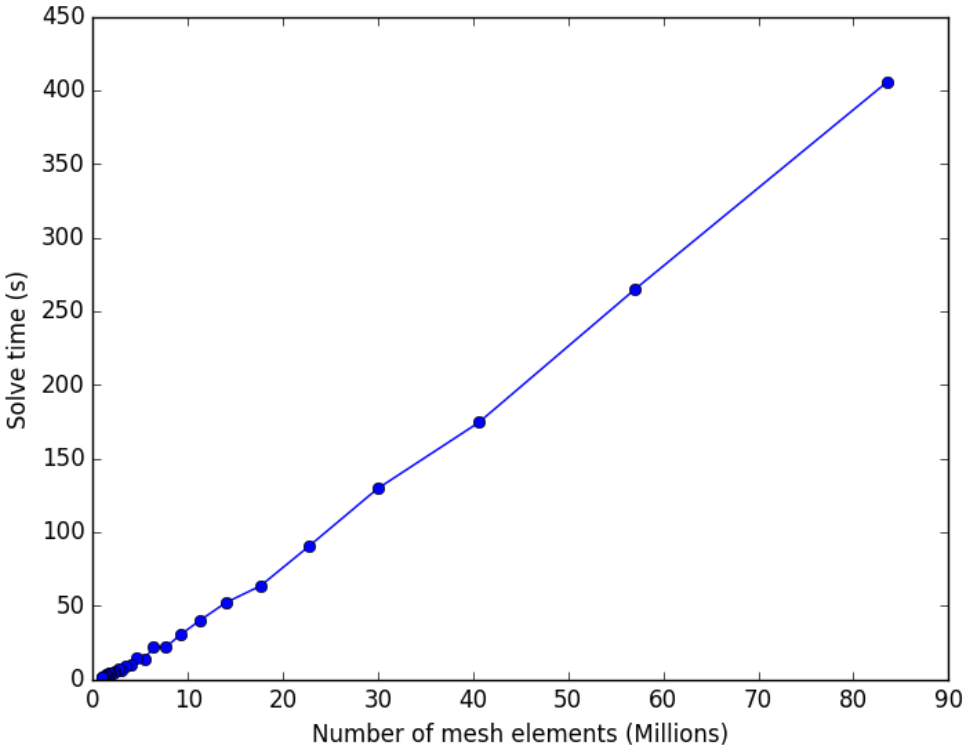

FIGURE 5.5: The time required to calculate the model as a function of the number of elements in the model.

order of tens of millions of elements and millions of nodes. Tetgen, like all meshing programs, requires significant amounts of system resources to generate the mesh, typically more than 16 GB of RAM for meshes of tens of millions of elements and takes an appreciable time. It is not always possible to parallelise this process given the required resources therefore this process is typically serialised. This is further complicated by the need for error checking of the output of the Pogo model and if necessary, due to either insufficient GPU RAM or the model being unstable, the mesh may need to be regenerated using a coarser element volume constraint. Therefore maximising the occupancy of the CPU and GPU to maximise the efficiency of model generation and execution is not trivial. An improvement of this process would be to create a parallel job submission system in which meshes to be generated and input files could be queued to the CPU and these input files could then be queued to the GPU for execution. Time did not allow for this in the project however this would be a very worthwhile avenue of investigation if this was to be performed in anger again.

Within this process, Tetgen uses a simple text file format to store the information about the elements and nodes in the mesh. For meshes with tens of millions of elements these files can be of the order of gigabytes in size. This is a significant amount of text to parse into a floating point binary format and for some models this part of the work flow may take up to 40% of the total time. Ideally, this could be accelerated either by integrating the code into the input file writer, for example in Python using a Cython wrapper, or, less effectively, writing the files out into a binary format which would save parsing the information on both ends of the transfer. Presently neither of these were implemented however ideally, should this work continue, they would be and the former would be significantly more efficient than the latter.

Prior to running a meshing program, the volume that is to be meshed needs to be defined. This problem is not unique to meshing but is common in a range of fields, primarily computer aided design (CAD). The generation of a single geometry that describes a volume is trivial to perform by a human in a CAD program however automating this process for variations in parameters of the geometry, such as the length and height of the crack, is not trivial and requires some coding. This therefore requires a defined format that describes the geometry which can be modified using code.

A common and useful format for defining a volume is a Boundary Representation (BRep). A BRep consists of defining the volume in terms of the

facets that enclose it and within this the polygons that define each facet. The polygons are described by the points that are the corners of polygons and the lines which join the points to make the polygon. This is a simple format which may be easily written into a text file and is known as the poly file format. A complete description of the format along with some example input files can be found here [94]. The advantage of this format is that the location of individual nodes can be redefined without having to redefine the entire facet, for example the tip of a crack can be moved without having to redefine any other lines or facets. This format is also easy to generate analytically, allowing for the easy automation of this process. The main difficulty is in combining objects, such as placing a crack on the side of a cylinder. This generally requires the modification of the facets of at least one of the objects so that they become attached however this can be done in an automated fashion. This process must also be tested to ensure that the geometries passed to the meshing tool are valid. It is not possible to generate and test every possible geometry therefore the use of the corner cases of the parameter space can be generated and validated as correct inputs of the meshing tool. This provides confidence that the automated geometry process is robust. It is also important to include error checking and handling in this process so that models that fail can be recorded and modified as necessary to obtain valid geometries. A significant amount of effort was expended developing the functions and work flow necessary to perform these operations in an automated fashion. The result of this process is a library of robust parametric definitions of geometries which were needed in this project which is now available in a public GitHub repository [95].

5.2.3 Precision

A question that arises in any numerical model is that of its precision and accuracy. The question of accuracy has been discussed previously and will be further considered where models are compared with experiments. This section deals with the issue of numerical precision of Pogo.

There are two main sources of error in numerical simulations: accumulated rounding error and error caused by data corruption in memory. The former is caused by having insufficient bits to precisely compute operations or the inability of the binary basis to represent some numbers. The latter is caused by a bit in memory undergoing some excitation which causes it to change state, thus changing the number stored in memory. The former can be reduced through

using more bits (using double rather than single precision) and using a large range of the binary representation by multiplying all numbers by a constant factor. The latter can be guarded against through the use of Error Correcting Code (ECC) memory which is present in nVidia's more expensive Quadro and Tesla cards however this does come at a cost of slightly reduced performance. There is evidence to suggest that this phenomenon is less prevalent than it was in the past due to improved manufacturing [96] although it may still be an issue [97]. Most errors of this kind are caused by defective hardware rather than external sources therefore it is worth performing rigorous acceptance tests of hardware before using it in anger. These errors were investigated by repeating a simulation to determine if any there is any difference between the results. The test scenario was a 2D square with an irregular triangular mesh. The mesh was created with a defined maximum element area therefore there are a range of element sizes within the mesh. The mesh was created to achieve a maximum of 5 elements per wavelength at twice the centre frequency of the input pulse. A 5 MHz, 5 cycle Gaussian tone burst was applied to a single node in the centre of an edge of the square, designed to be representative of signals likely to be present in ultrasonic NDT applications. The simulation consisted of 200,000 time steps for a total duration of 100 μs. This node is monitored along with 5 other randomly chosen nodes. This simulation was repeated for both the 32 bit and 64 bit versions of Pogo and these models were then repeated 100 times. No discrepancy was found between the 32 bit and 64 bit versions, both results were repeatedly identical. This suggests that there is minimal (if any) rounding error in the 32 bit calculations. Similarly, the results of repeating both the 32 bit and 64 bit are identical, suggesting that in both versions there is minimal (if any) random numerical error caused by bit flipping in either version. The 32 bit version had an average run time of 18.52 $\pm$ 0.16 s and the 64 bit version 22.04 $\pm$ 0.29 s. This is a difference of approximately 20%, similar time differences have also been consistently observed throughout this work. Given the identical nature of the results and the difference in run time, the 32 bit version was used for all simulations as minimising the run time is an important part of performing a qualification efficiently.

5.2.4 Modelling Transducers on Wedges

A common feature of ultrasonic inspections is the use of a wedge to direct the beam within a material. In practice coupling gel is used to provide improved

coupling between the wedge and the specimen, filling in the air gap that would otherwise be present in dry coupling. However, given that Pogo is unable to model gels as they are not elastodynamic, faithfully recreating this scenario in Pogo is not possible. The modelling of ultrasonic transducers is also complicated as they are complex devices, typically consisting of a piezoelectric element surrounded by some backing materials. It is therefore far easier to experimentally measure the output of a transducer and apply this as a load under ideal coupling conditions. The liquid nature of the couplant is accounted for in the model by only applying the load forces normal to the surface of the specimen. The primary effect of varying the coupling is to provide a change in gain of the signal, therefore a first order model of this effect is to apply a multiplicative factor to the amplitude of the measured displacements. This is particularly useful as the effect of coupling thickness can be included in calculations of metrics of an inspection as an independent parameter through the methods discussed in Chapter 3.

Whilst the wedge is not modelled, angled beams still need to be induced into the model to accurately model the inspection. Angled beams can be induced into a material by applying loads directly to the surface of the specimen and using appropriate delay laws to propagate the ultrasound at a desired angle. This concept is illustrated in Fig. 5.6. The delay Δt for a given node which is a distance Δx from the edge of a wedge at an angle θ is given by

$$\Delta t = \frac{\Delta x \sin(\theta)}{v}, \tag{5.1}$$

where v is the velocity of the wave of interest. This is calculated for each node in the footprint of the transducer and a bespoke load is defined for each node. If an analytically defined signal is used then the implementation of the delay is trivial however for experimentally measured signals, it is easiest to perform this delay in the frequency domain, multiplying the input spectrum by a time delay as

$$f(t + \Delta t) = F^{-1}(F(\omega)e^{-i\omega\Delta t}), \tag{5.2}$$

where $F(\omega)$ is the Fourier transform of the input signal $f(t)$, F^{-1} is the inverse Fourier transform and ω is the angular frequency. This process is also used on reception as the received signals have to be further delayed to simulate the return of them to the transducer at the end of the wedge. This is again achieved using phase delays applied in the frequency domain and adds extra run time to the model. However, this time is normally far less than simulating the wedge

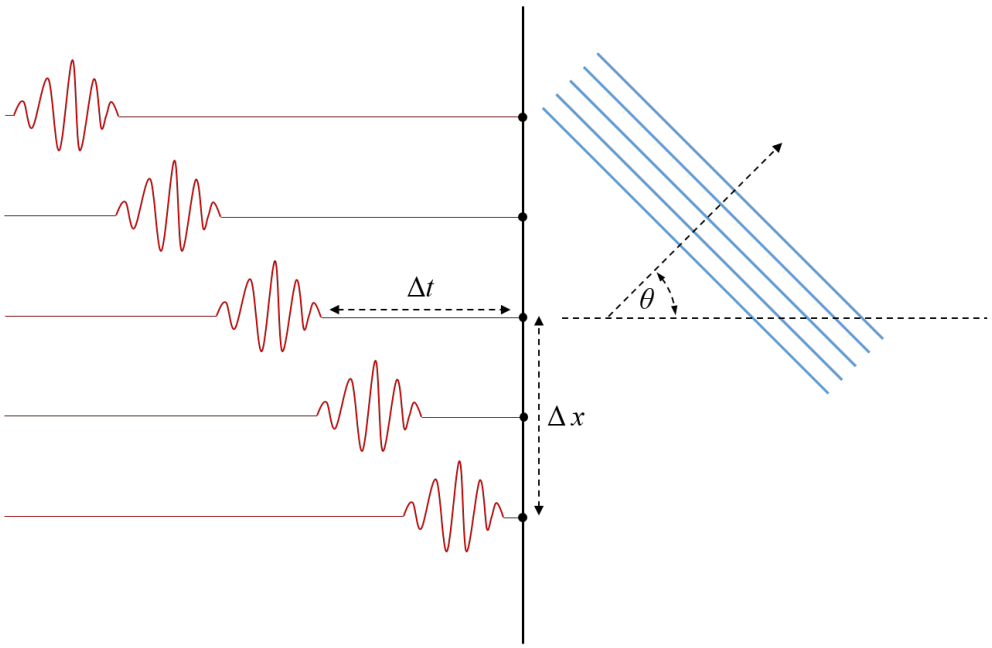

FIGURE 5.6: A diagram of the delay method for generating an ultrasound beam (blue) at an arbitrary angle θ in a material. The input signals (red) for each node (black dots) on the surface of the specimen (black solid line) are delayed by a time dt dependent upon the node's distance dx from the edge of the transducer and θ.

and is an easier process to automate. The limitation to this process however is the required RAM on the GPU to store all of these input signals. In the case of large models, this can potentially become prohibitively expensive as a significant proportion of the available memory will have to be dedicated to storing these signals. As Pogo requires a load value for each time step in the model, this also becomes an issue when there is a large number of time steps. In this case, when a short initial pulse is used in the model, a large number of zeros still have to be stored for the remainder of the time steps, wasting resources.

5.2.5 Pogo Simulations in Parallel

Modern graphics cards support Multiple Input Multiple Output (MIMO) instructions, practically this means that they are capable of running multiple models simultaneously if enough RAM is available. This is especially useful for Full Matrix Capture (FMC) simulations of phased arrays in which each transducer needs to be fired in turn and listened to on all others, requiring a separate model evaluation for each transmission. Parallel models can be easily implemented using parallel mapping functions where the limitations are the

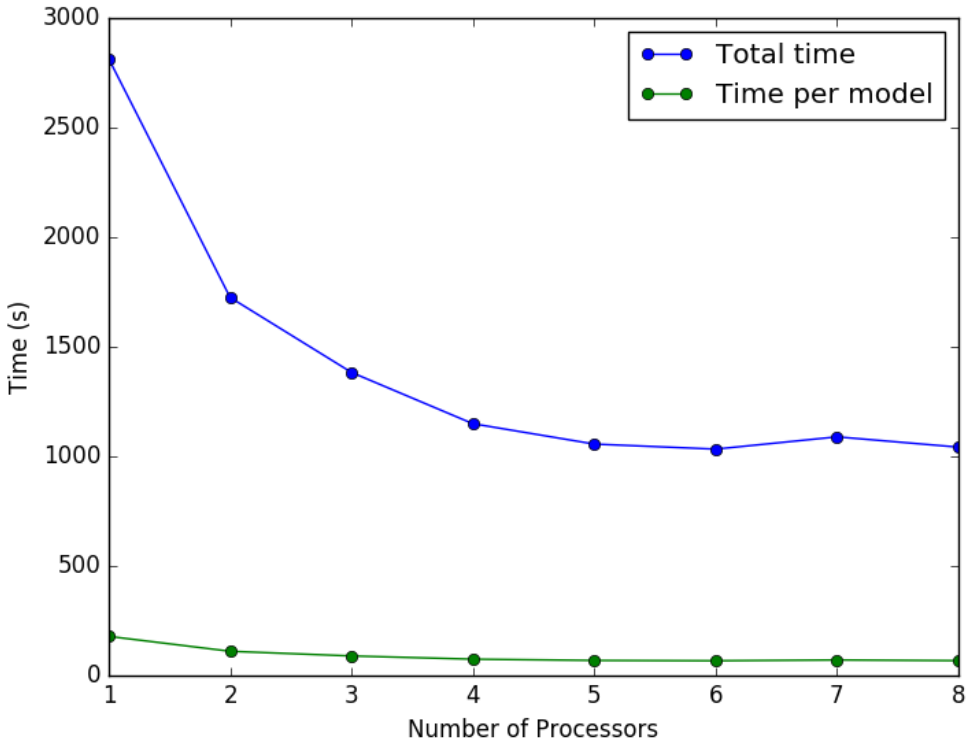

FIGURE 5.7: Result of changing the number of parallel processes used to run 16 identical Pogo jobs on an nVidia GeForce Titan X card.

number of CPU threads (one per model) and the available RAM on the GPU. Using a typical desktop CPU (an Intel Core i7), this can run 8 jobs in parallel. This was investigated using a simple small 2D model, chosen such that the memory footprint is small enough to fit 8 jobs on the GPU simultaneously. The result of this is shown in Fig. 5.7 which demonstrates that parallelisation can produce a significant acceleration of the process. Whilst individually each job takes longer to run due to fewer processors available to each job, overall the total time is reduced, effectively reducing the run time per model. A plateau in the run time is also present in Fig. 5.7 which may be attributed to the limitations on memory bandwidth, writing to disk limitations and the trade off in the number of GPU processors per job.

5.2.6 Experimental Validation

A key stage in the process of model assisted qualification is to demonstrate that the numerical model being used is an accurate representation of the real inspection. This is therefore something that must be achieved for this scenario to give credence to the modelled result. Two specimens of this scenario were

manufactured to provide experimental test pieces on which to perform experimental validation. One is defect free, that is it consists of only an aluminium plate with a hole in its centre, and one has two notches in it of different lengths, one on each face, created using electronic discharge machining (EDM).

The primary difficulty in matching experimental with simulated results is the matching of the properties of the probe. The probe has a rectangular footprint and is therefore defined in the model by two physical parameters, the length and width of the face of the transducer that is on the specimen. It was found through modelling trials that the two parameters are independent therefore could be tuned sequentially. This significantly reduces the complexity of matching these parameters. It is therefore possible to perform two scans of the defect free specimen, one away perpendicularly from the side wall (Scan 1) and one parallel to the hole at a distance that maximises the highest response (Scan 2) to tune the two parameters respectively. However, it was found that a constant force across the face of the transducer did not result in a match of the experimental and measured profiles. It was found that a two-dimensional Gaussian force profile across the transducer yielded good agreement between model and experiment and that the width of these distributions in each axis could be tuned independently. Physically, this is reasonable as the actual piezoelectric element that drives the transducer is significantly smaller than the footprint of the transducer and the beam of ultrasound will diverge before hitting the surface therefore causing a peak in the centre and lower amplitude at the edges. The properties of these distributions were therefore tuned alongside the length and width of the transducer to match experimental results.

There are two main challenges to performing these measurements: precisely knowing the location of the probe and maintaining good coupling between the probe and specimen. The former can be alleviated by co-locating the peak amplitudes of the experimental and measured scans. Therefore only the change in position is required rather than the position relative to the specimen. The latter can be controlled through careful application of the couplant and maintaining a constant force on the transducer to keep it in contact with the specimen. This was achieved through the use of a carefully constructed rig attached to a micrometer stage which allows precise positioning of the probe and through many averages the effects of small changes in couplant thickness was minimised. The results of this for Scan 1 and 2 are shown in Fig. 5.8 which largely shows good agreement between the model and experiment. The maximum amplitude of Scan 1 was used as a normalisation factor for the measurements and the

profiles agree well in both amplitude and shape. The exception to this is the presence in Scan 1 of some reflections at positions beyond 16 mm which are not present in the experiment. This is due to reflections from the edge of the plate in the FE model. This model can then be tested against a defect in the specimen (Scan 3). The result of this is shown in Fig. 5.8 which shows good agreement in both amplitude and shape of the profiles. Given this good agreement, these scans provide sufficient validation of the numerical model of the inspection and it can thus be used for mapping the parameter space.

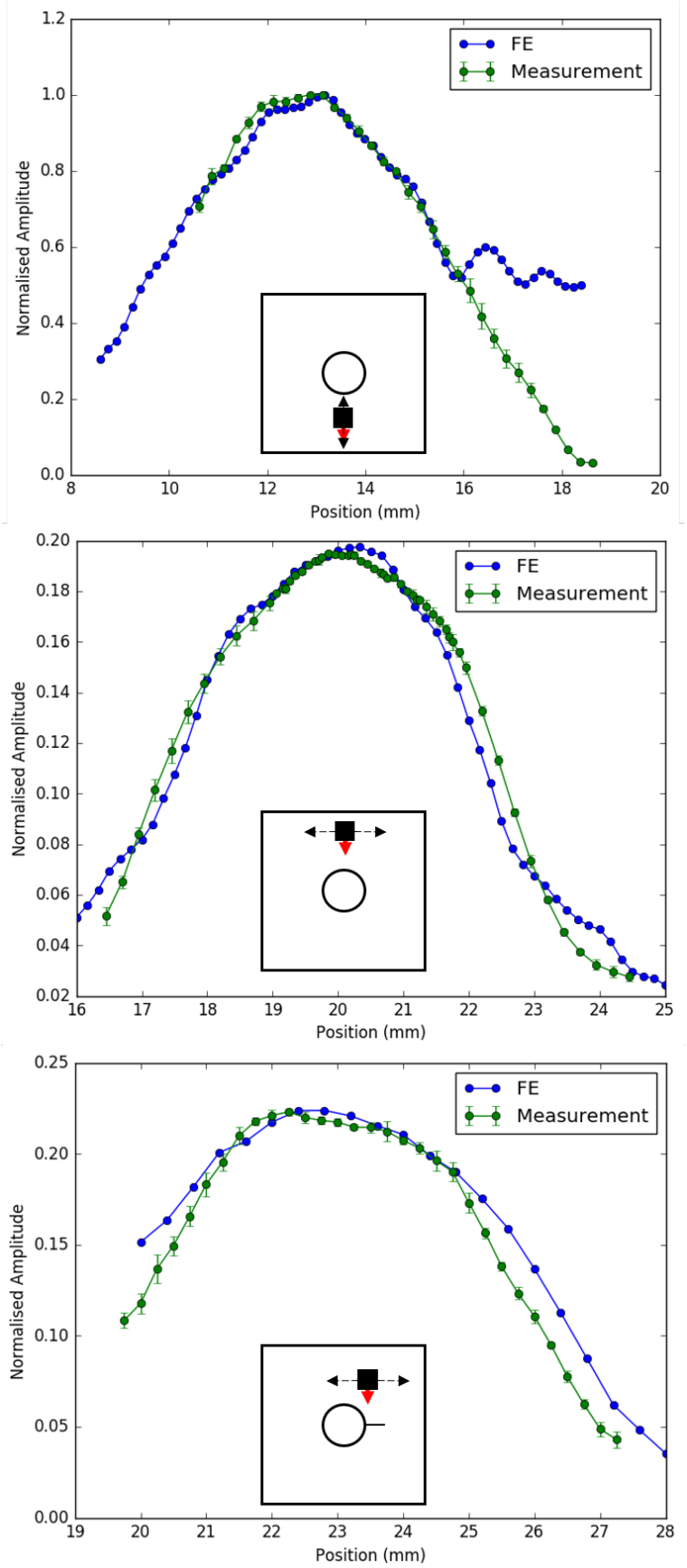

FIGURE 5.8: The modelled and experimentally measured scans for the validation of the numerical model. The three scans are Scan 1 (top), Scan 2 (middle) and Scan 3 (bottom). The inset diagrams show the probe (black square), the direction of the scans (black arrows) and the direction of the ultrasound beam (red).

5.3 Two Parameter Case

In the first instance, the properties of the crack are taken to be fixed with a height of 2.89 mm, length of 5 mm and an angle of 0^o. The rotation of the probe is taken to be fixed and the probe's lateral and perpendicular position were varied. Each of these parameters was assigned 11 values each, yielding a small parameter space of 121 points. Given this small volume, it was possible to fully map the response space, as shown in Fig. 5.9. This shows a response function which falls off in all directions away from an optimal position, qualitatively appearing to be the product of two Gaussian functions of each of the two parameters. This also appears to decrease faster as the position in the perpendicular axis changes than as the lateral position changes.

5.3.1 Response Function Mapping

Using the full data set it is possible to sub-sample and interpolate to test the predictive algorithm. Sampling was performed using the ESE LHD algorithm described in Chapter 4. As the full space was mapped, it is possible to repeat this process to obtain a measure of the repeatability of the process. An example of this process is shown in Fig. 5.10 which shows an example set of sample points chosen for both the error set and sets used to build the predictors. The sampling and interpolation algorithm was repeated 11 times by varying the choice of error set and initial sample set with all four interpolation algorithms applied each time. The MARS algorithm used a maximum degree of interaction of 4 and smoothing was not applied to the resulting interpolant. The Gaussian RBF function used the mean distance between sampled points as an approximation for the parameter β in Eqn. 4.18. The linear and cubic RBF functions require no specific parameter definitions.

The results of performing interpolation using these methods are shown in Fig. 5.11 for the comparison to both the independent error set and the full data set. The samples and the independent error set were the same for all of the methods thus providing an accurate comparison of the interpolation methods. Of these methods the cubic radial basis function and the cubic interpolation are shown to be the best methods as they produce the best interpolators, they have the smallest predictive errors, and are the most consistent, they have the smallest error bars showing the smallest variations between different sample sets.

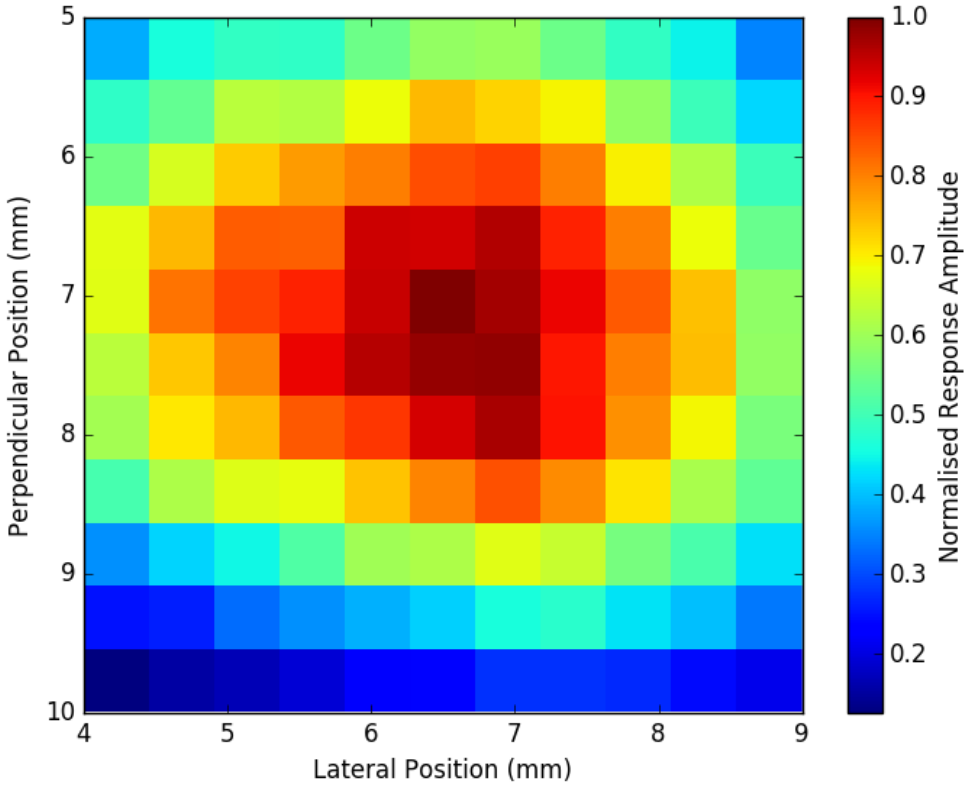

FIGURE 5.9: The response map for the 2 parameter case, mapped using a 3D finite element model in Pogo. The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o. The rotation of the probe is fixed at 0^o.

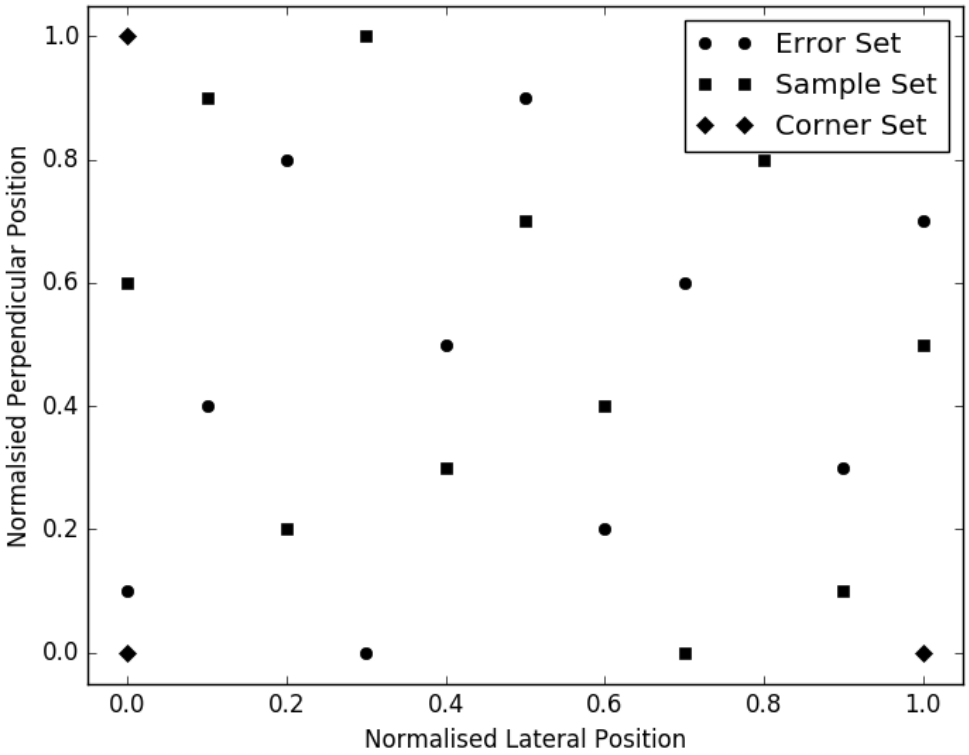

FIGURE 5.10: Example of points used to sample the parameter space. The error set is used to test the quality of the interpolation. The corner points and sample set are used to build the interpolator.

It should be noted that the cubic interpolator can only be applied to parameter spaces of one or two dimensions therefore can not be used for high dimension spaces. The Gaussian RBF function does not perform as well primarily due to this interpolation function being a poor representation of the underlying response function. Whilst it is unsurprising that, given the smooth nature of the response function, an appropriate interpolation function performs well, it is good demonstration of applying the mapping algorithm to simulated inspection data.

5.3.2 Inspection Metrics

The response map can be used to calculate metrics of the quality of the inspection. The combination of this data with an adjoint probability space allows a PoD curve to be calculated for each parameter. The probability space can be approximated by estimating the probability density functions for each parameter. The probability density function for the lateral position was modelled by a Gaussian of mean 6.5 mm and standard deviation 1 mm. Similarly, the

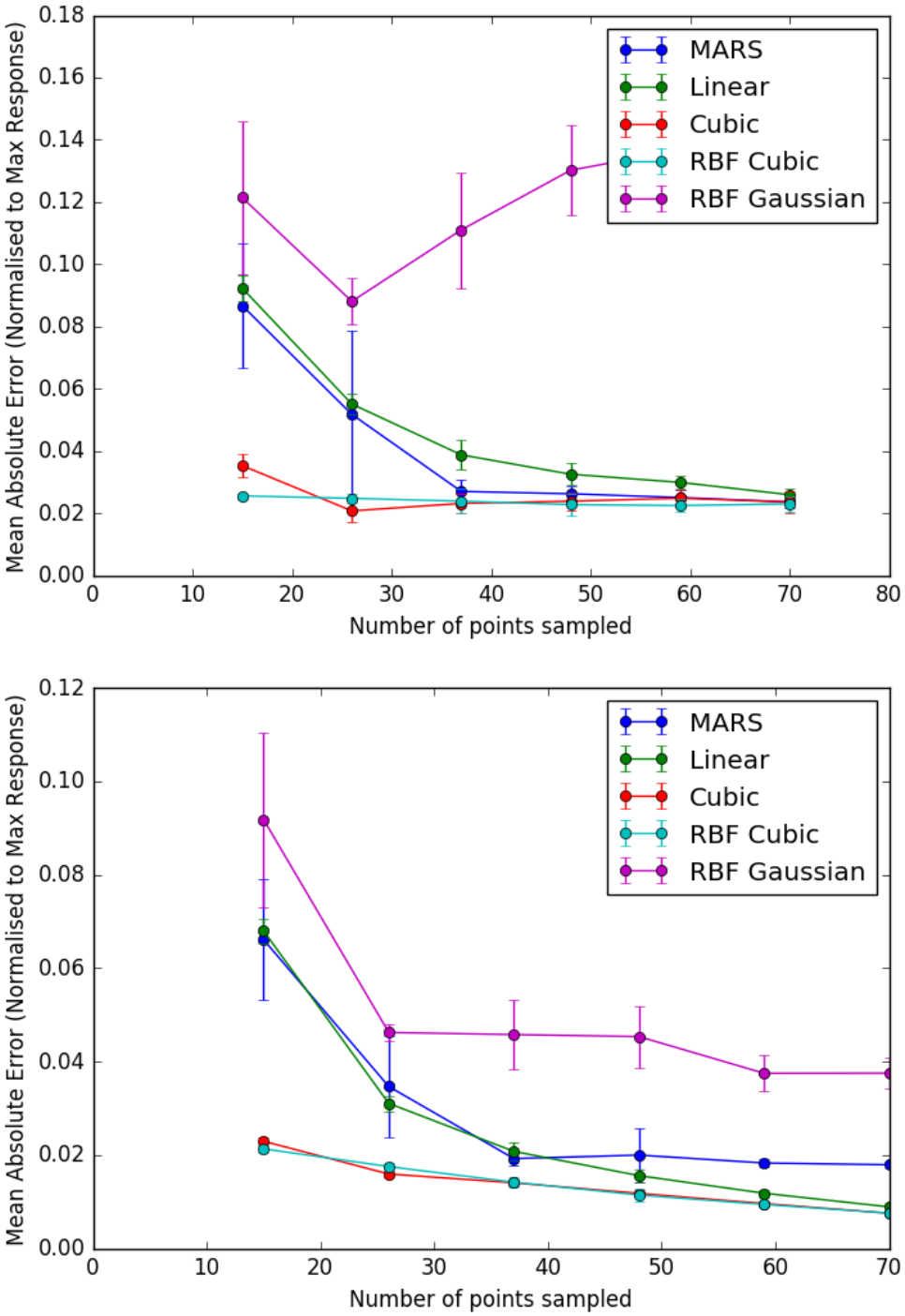

FIGURE 5.11: The error in the prediction of the response space when compared to an independent error set (top) and the full response map (bottom) using a range of interpolation methods: Multivariate Adaptive Regression Splines (MARS), linear interpolation, cubic interpolation, a cubic Radial Basis Function (RBF) and a Gaussian RBF. The error is calculated as a fraction of the maximum response.

probability density function for the perpendicular position was also modelled by a Gaussian with mean 7.5 mm and standard deviation 1 mm. It is assumed that the variation of parameters occur independently, that is the value of one parameter has no effect on the likelihood of a given value of another parameter. A decision threshold on the response of 0.5 of the maximum response amplitude was used. The PoD curves as a function of each of the two variable parameters are shown in Fig. 5.12, calculated using the fully mapped space. These plots demonstrate one advantage of this method over traditional qualification approaches, that it is possible to produce PoD curves for each parameter that has been varied in the mapping process from the same data set. The PoD curve for the perpendicular position vary far more across the parameter range than PoD curve for the lateral position. This suggests that the inspection is is more sensitive to the perpendicular position of the probe than the lateral position.

The relative importance of parameters can be calculated using Sobol sensitivity indices, as discussed in Chapter 4. The results of this process, as a function of the number of points sampled, is shown in Fig. 5.13. The cubic interpolation method is used to build the necessary interpolator. This shows that the perpendicular position of the probe is the most significant parameter, as suggested by the PoD curves in Fig. 5.12. There is also very weak interaction between the parameters, shown by the Sobol index for this interaction being approximately zero, suggesting that they are approximately independent. If this is true then the response function can be expressed as

$$R(\Omega) = f_0 + f_1(x_P) + f_2(x_L), \tag{5.3}$$

where f_0 is a constant intercept term, f_1 and f_2 are functions of the probe's perpendicular position x_P and lateral position x_L. The intercept term can be calculated by averaging the response over Ω. The functions f_1 and f_2 can be numerically evaluated by fixing the value of the other parameter and varying the parameter of the respective function. After two iterations of the sampling and interpolation algorithm, the model has been evaluated 26 times and the sensitivity indices suggest that the effect of the parameters on the response function are approximately independent, that is the response function can be approximated by Eqn. 5.3. A further 9 model evaluations per parameter are required to generate sufficient data to sample a complete slice across the parameter space (11 points) that is not on the edge, bringing the total number of models to 45 model evaluations. A cubic spline was used as an approximation of each of the functions in Eqn. 5.3 as a cubic interpolation has been shown to

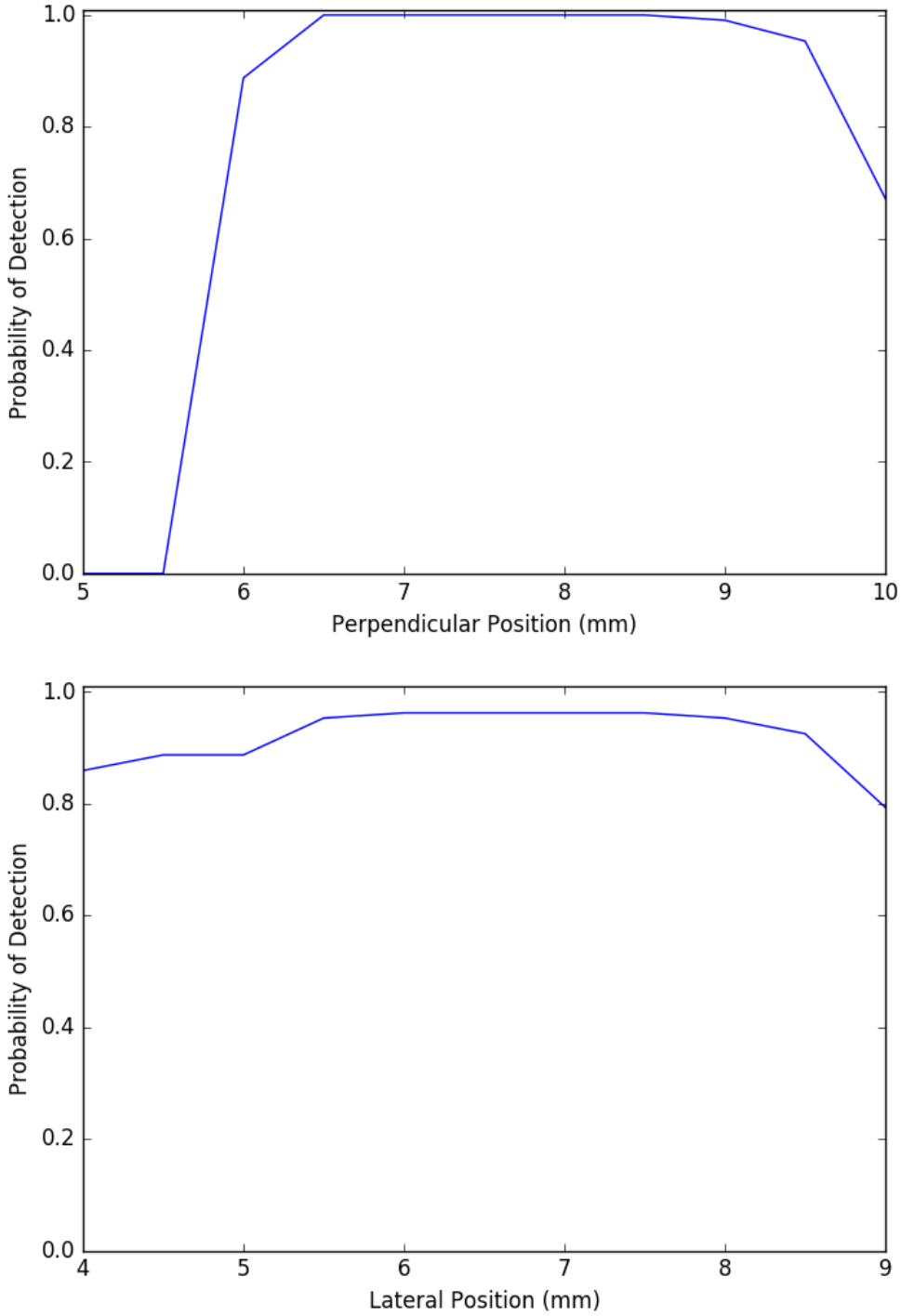

FIGURE 5.12: The Probability of Detection curves for the two variable parameters in the inspection, the perpendicular position of the probe (top) and the lateral position of the probe (bottom). The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o. The rotation of the probe is fixed at 0^o. A decision threshold on the response of 0.5 of the maximum response amplitude of the first reflection was used.

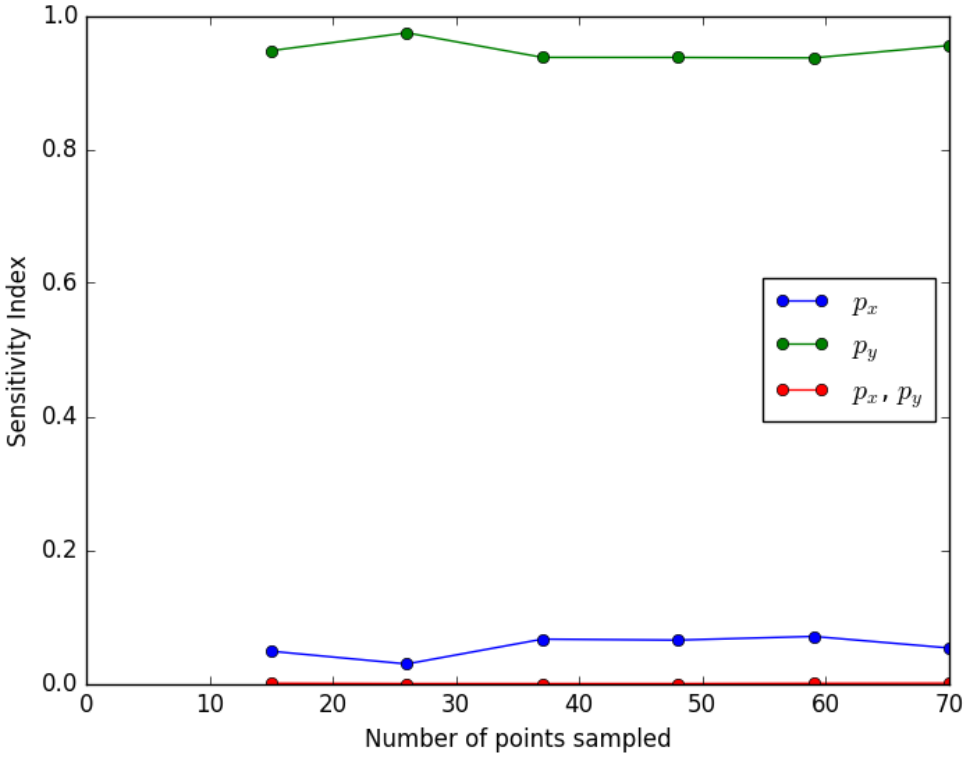

FIGURE 5.13: The sensitivity indices for each parameter and their interaction as a function of the number of sampled points.

be a good interpolator for this function. These functions are shown with the full data set in Fig. 5.14. The error in the prediction of this model is shown in Fig. 5.15 which has a mean absolute value of 4%. This value is skewed by the presence in some areas of a much greater error and the median value of the error at the 121 sampled points is 0.02%. When the independent functions are built using the full data set, the mean absolute error is 3%. This suggests that the use of only a single slice of data in each parameter attains close to the best possible approximation of the independent functions. It should be noted however that whilst this is a good demonstrator of this method, given the very good predictive nature of the interpolator, in this scenario the reduction of the parameter space through independence is unnecessary.

5.4 Three Parameter Case

In reality, the operator will rotate the probe as well as translating it on the surface, thereby introducing an additional degree of freedom into the inspection and creating a three dimensional parameter space. The rotation of the probe

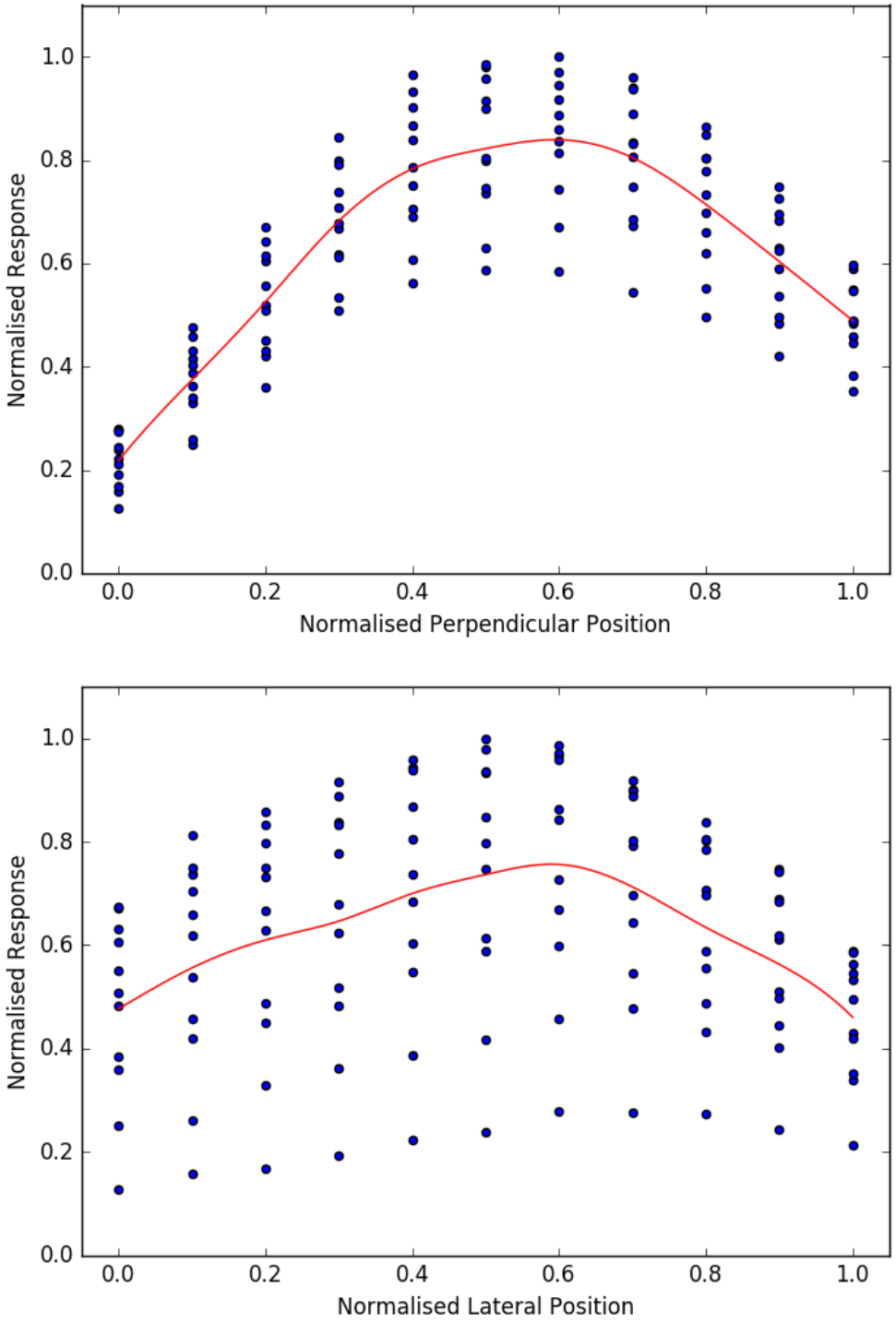

FIGURE 5.14: The data and fit for the first order functions when treating the two variable parameters as having an independent effect on the response of the function. The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o. The rotation of the probe is fixed at 0^o.

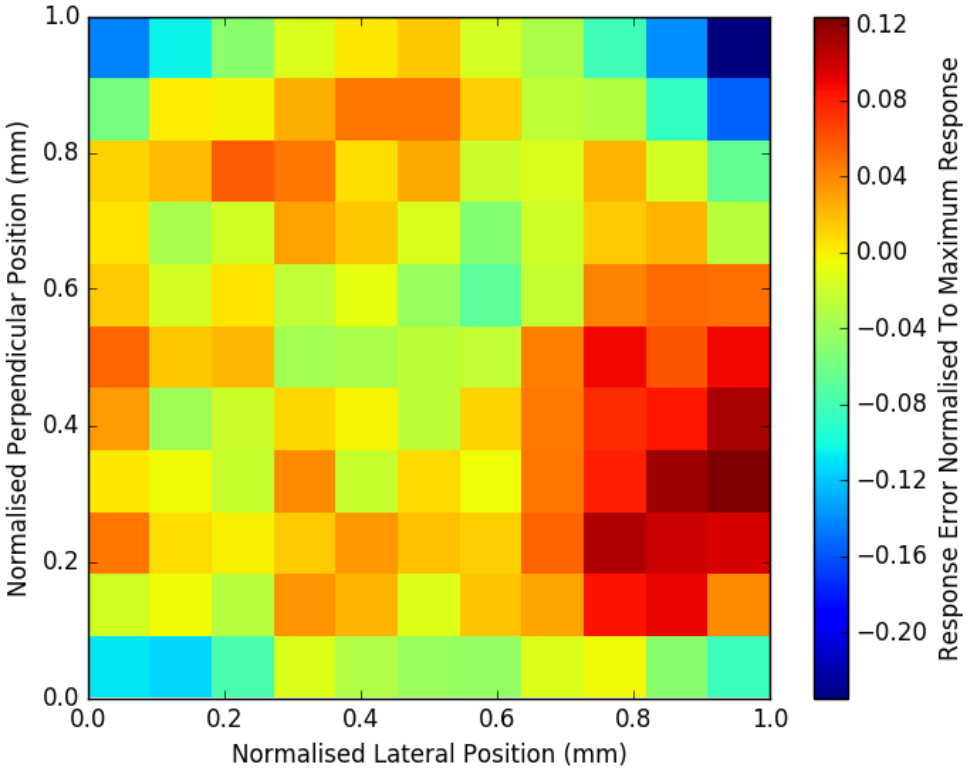

FIGURE 5.15: The error in the prediction when the parameters are treated as being independent, plotted as a fraction of the maximum response amplitude.

on the surface can be easily incorporated into the previous model and requires only a change of definition of which nodes on the surface constitute the probe. Appropriate modification of the calculation of the delay laws is also necessary to generate a beam at the desired angle in the specimen. This can be accomplished by rotating the positions of the nodes that constitute the probe onto a plane parallel to a single axis, in this case the axis for which the rotation of the probe is zero, using the two dimensional rotation matrix

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{5.4}$$

where x and y are the location coordinates of the nodes, x' and y' are the rotated coordinates and θ is the angle of rotation. In this scenario the range of possible rotations was $\pm 10^o$ from the probe being perpendicular to the crack.

This parameter was again assigned 11 values across this range, giving a total parameter space volume of 1331 points at this sampling resolution. This is a relatively large parameter space and given the appreciable model evaluation time of the order of 15 minutes, it was not fully mapped. The sampling and

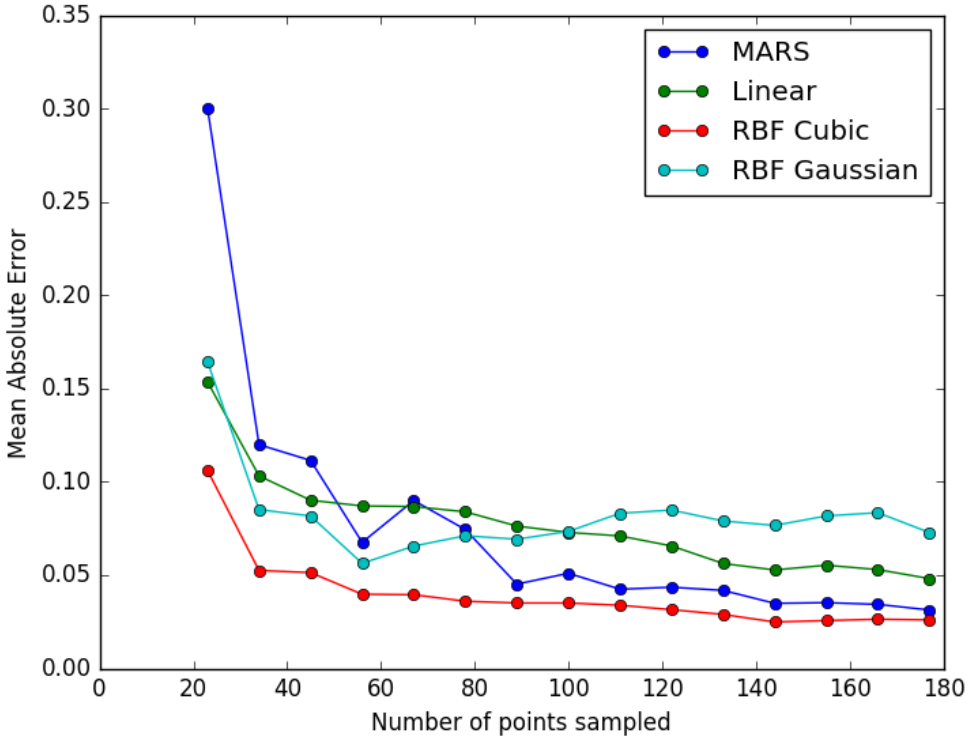

FIGURE 5.16: The error in the prediction using of the 3 parameter response space when compared to an independent error set. The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o.

interpolation algorithm using the ESE LHD algorithm, testing against an independent error set, as discussed in Chapter 4 was applied to map the response and build an approximation of the response function. The results of this process are shown in Fig. 5.16. The linear interpolator, MARS algorithm and the cubic RBF all perform well, giving a decreasing predictive error. Of these, the cubic RBF gives the lowest predictive error. The Gaussian RBF performs worst again, due to the basis function being a poor representation of the underlying response function. It also shows that it is possible to achieve a low predictive error after mapping a small proportion of the parameter space, in this case a smaller proportion of the parameter space than the two parameter case discussed previously. This is linked to a minimum density of points required to obtain a good prediction: as the number of parameters increases, the number of points required to obtain this density increases but at a slower rate than the increase in the volume of the parameter space. It is not possible to predict the required number of sampled points to map an unknown response function as it is highly dependent upon the smoothness of the function, in general a smoother function will require fewer sampled points.

The probability distribution for the rotation of the probe is a Gaussian with

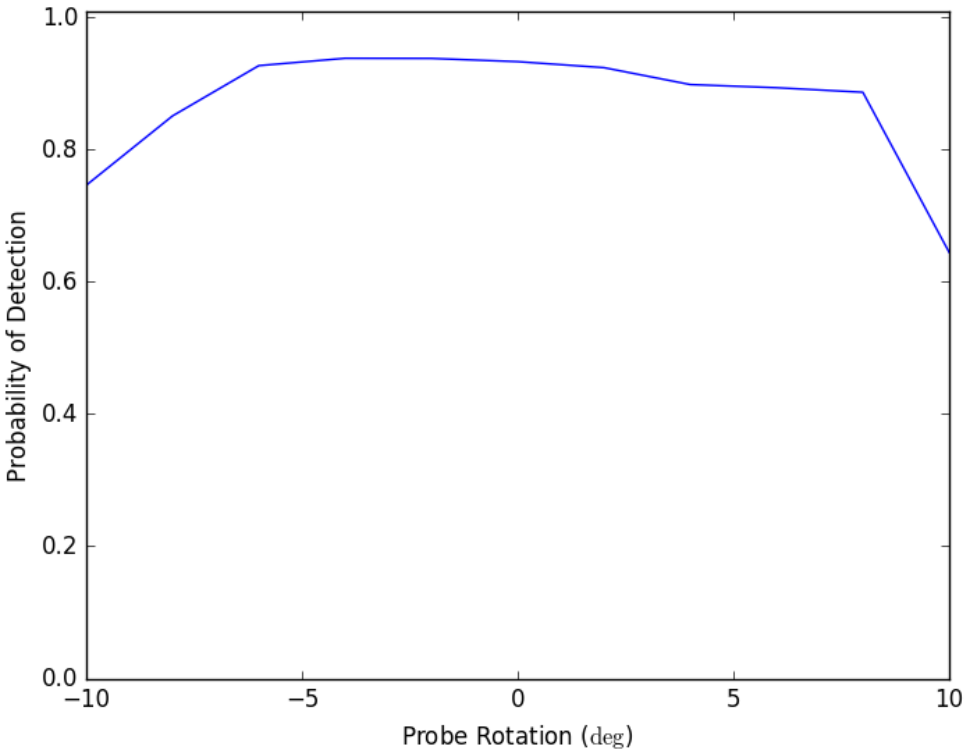

FIGURE 5.17: The Probability of Detection for the three parameter response space as a function of the rotation of the probe. The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o.

a mean of 0^o and a standard deviation of 3^o and the probability distributions for the perpendicular and lateral positions are the same as in the two parameter case. This information can be used to plot the PoD for the rotation of the probe as well as the PoD for the other parameters, as shown in Fig. 5.17, 5.18 and 5.19. A threshold on the response of half of the maximum was used for sentencing, the same as the two parameter example. Figures 5.18 and 5.19 show a significant change in the PoD curves for the two positional parameters when the rotation of the probe is considered. This suggests that the rotation of the probe is a significant parameter which has a noticeable effect on the response of the inspection. This can be quantified using Sobol indices and these, as a function of the number of sampled points, is shown in Fig. 5.20. These all converge after approximately 4 iterations of the sampling and interpolation algorithm and the minor changes as the number of iterations increase can be attributed to a combination of numerical error in their calculation, caused by the use of Monte Carlo integration, and small changes in the underlying map of the response function as more points are added. This shows that the perpendicular position of the probe remains the most significant parameter of the inspection and that the rotation of the probe has a greater impact on the response of the

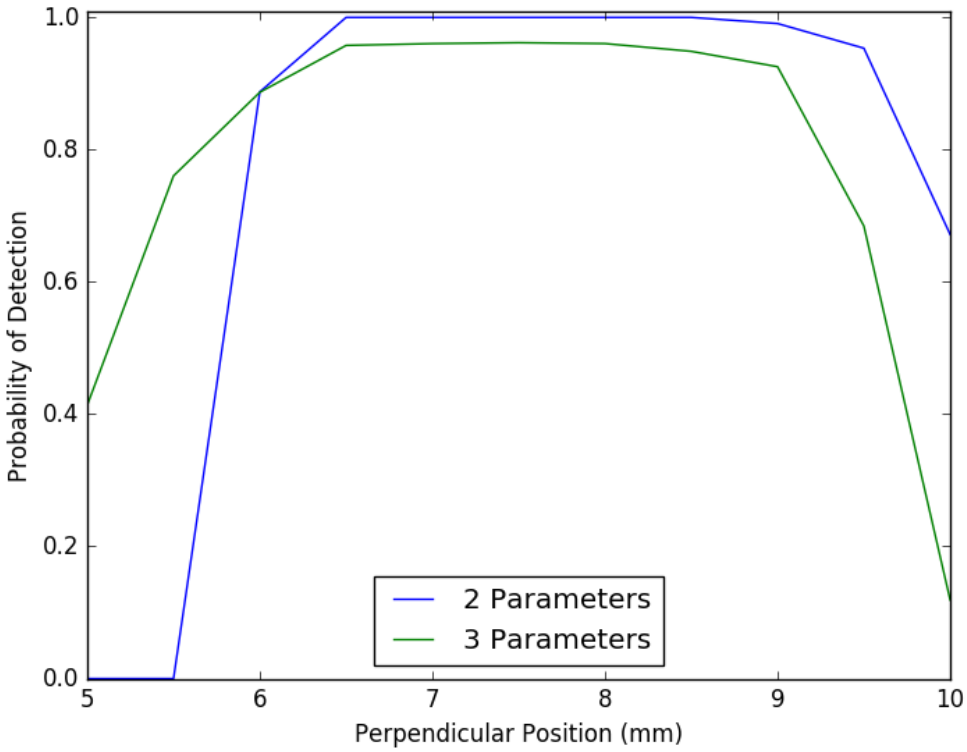

FIGURE 5.18: The Probability of Detection for the perpendicular position of the probe as function of two parameters and three parameters. The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o.

inspection than the lateral position of the probe. However, for all parameters the total sensitivity indices are not approximately equal to their total sensitivity indices which demonstrates appreciable interaction between them. Based on this, it is not possible to discount any parameters as they all have a significant total order sensitivity index. As they are clearly not independent, it is also not possible to reduce the dimensionality of the parameter space through treating some parameters as being independent. This example demonstrates the need to consider all parameters which may have a non-trivial impact on the response as the ignoring of parameters which have an appreciable impact can lead to an incorrect measure of the reliability.

In practice, the properties of the defect, such as its length, will be the parameter of interest, therefore variations in its properties must be accounted for. This larger parameter space is investigated in the following section.

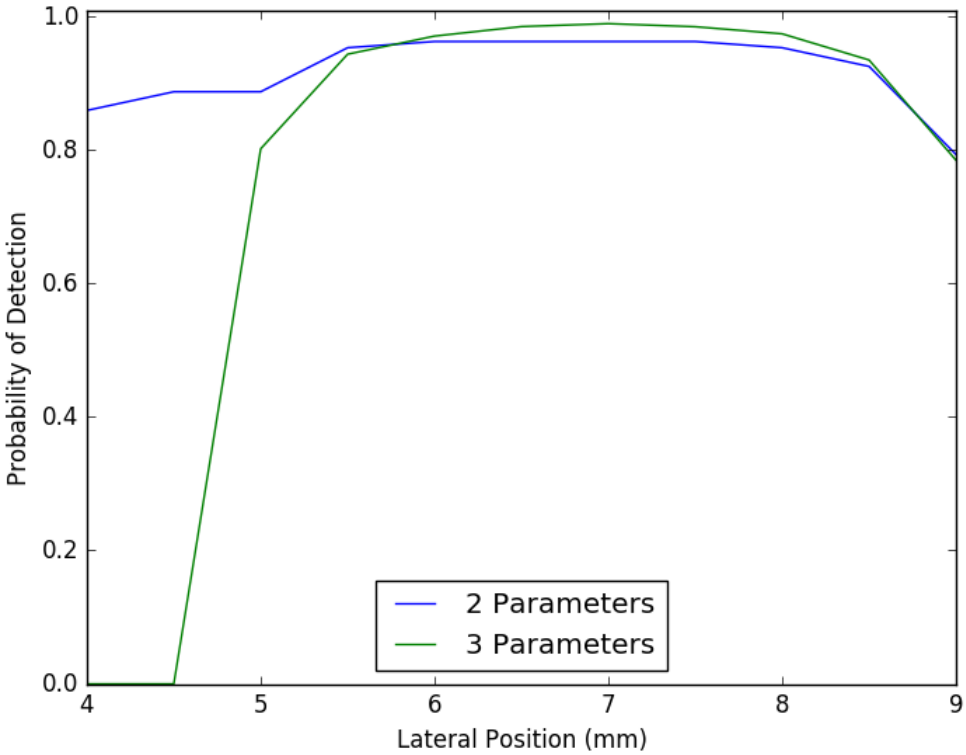

FIGURE 5.19: The Probability of Detection for the lateral position of the probe as function of two parameters and three parameters. The crack has a fixed length of 5 mm, height of 2.89 mm and rotation of 0^o.

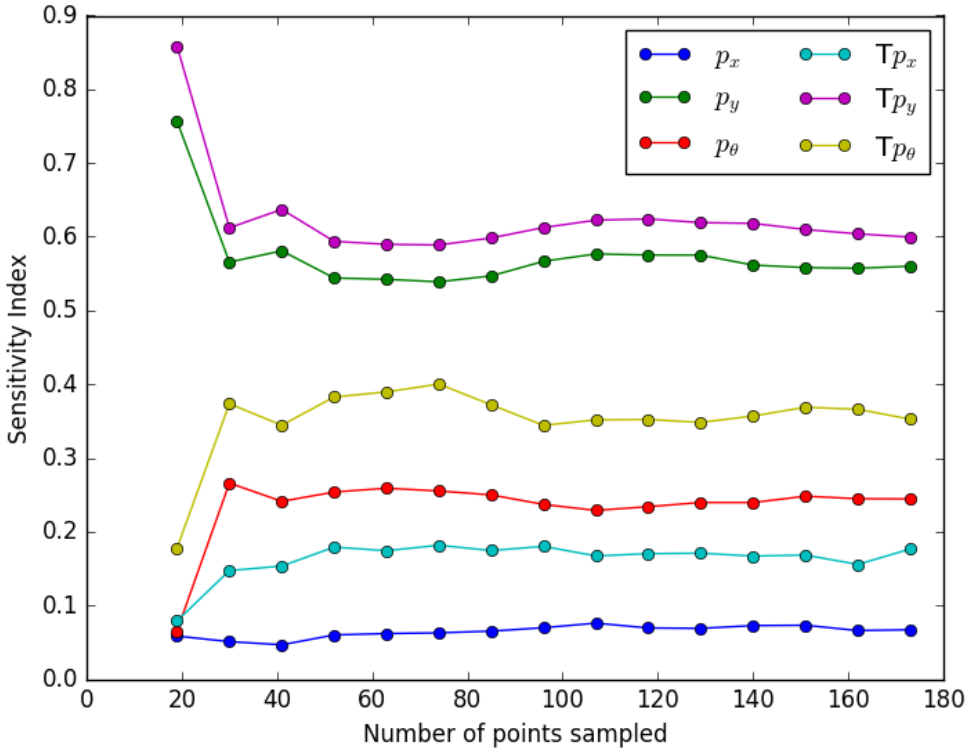

FIGURE 5.20: The first order sensitivity indices (p_x, p_y and p_θ) and the total order sensitivity indices (Tp_x, Tp_y and Tp_θ) as a function of the number of sampled points. The total sensitivity indices incorporate the interaction between parameters.

5.5 Eight Parameter Mapping

The previous qualification assumed that the crack was of a fixed size and orientation and that the only variations present were due to the operator moving and rotating the probe. In practice, the size and rotation of the crack can vary, therefore introducing a further three parameters into Ω. As these are the parameters of interest in this inspection it is essential to include these. Equipment factors may also be important, therefore the electrical noise in the system and the effects of varying couplant thickness will have an impact on the response and must be considered. This brings the total dimensionality of Ω to eight parameters which is evidently too large to evaluate the model at every coordinate.

If the inspector is properly trained, they should be able to apply couplant consistently in order to perform a repeatable measurement no matter the position or orientation of the probe or the properties of the defect, thus it can be considered independent of the other parameters. Similarly, the electrical noise present in the system is not affected by the actions of the operator or defect properties and can thus also be considered independent. These parameters can therefore be incorporated into the calculations of inspection metrics through the methods described in Chapter 4, reducing the dimensionality of the numerically modelled parameter space to six parameters.

5.5.1 Modelling Independent Parameters

The two independent parameters are better suited to being experimentally measured than being numerically modelled. In the case of electrical noise, the transducer was placed on the surface of the specimen with appropriate coupling applied and left in place for an hour. The use of an automated data collection method, in this case by programming data acquisition through the interface of the oscilloscope with a PC, minimises human impact on the inspection. A fixed time gate was applied to each of the 10,000 collected time traces where no reflections were present, thus in theory having a zero amplitude. The measured amplitudes in this window were recorded for each trace and this data was used to generate a distribution, found to be approximately a normal distribution, of the noise in the measured signals. The mean of this distribution is 0.0 and the standard deviation is 1.0% of the maximum response amplitude.

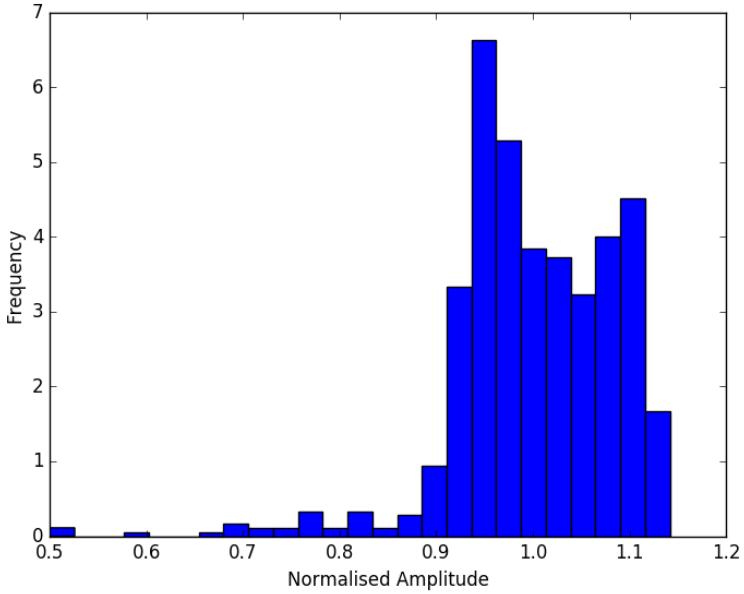

FIGURE 5.21: The distribution of response amplitude caused by changes in the thickness of couplant.

The effect of coupling thickness can be quantified experimentally through performing a measurement consistently in which the only parameter that may vary is the thickness of the applied couplant. This is achieved through performing an inspection of the back wall of a specimen repeatedly by removing the probe, reapplying the couplant and performing the measurement on the same location on the specimen. The amplitude of the first reflection is used as a normalisation factor for subsequent reflections. The histogram of this data is shown in Fig. 5.21, from which a numerical probability density function can be derived for use in the calculation of inspection metrics.

5.5.2 Automated Model Generation

The variability of the properties of the crack introduces significantly greater complexity into the automated generation of models. The allowance of variation in rotation, length and height of the crack introduces more failure points in the automated work flow. One major challenge is that the meshing algorithm is capable of producing a mesh with too many nodes and elements to fit within the available GPU RAM, which was found to occur in approximately 30-40% of models. The sampling and interpolation algorithm requires a finite result to be obtained for as many sampled points as possible and the omission of data points will degrade the quality of the interpolator and therefore the accuracy of

the response map. Thus a failure rate of 30% is not acceptable and will result in significant wasted effort.

As the amount of required RAM is dependent upon more than just the number of nodes and elements, it is not trivial to calculate the amount of required memory and whether this will exceed available resources. Therefore the simplest way to test if the model will be evaluated successfully is simply to attempt to evaluate it and then test whether a finite result is obtained, rather than an error of some nature in the case of a failed job. The built-in exception handling functions of Python make this process easy to perform, simplifying the check. In the case that a model is unsuccessful, the maximum element size constraint is relaxed by a small percentage and the model rebuilt. This process continues until a finite result is obtained. In practice, it has been found that the model typically exceeds the available resources by only a small percentage therefore only a small increase in size of the elements is required. This was not found to significantly degrade the accuracy of the model. This ensures that almost, if not, all the sampled data points return a finite result and thus can be used in the construction of the interpolator.

5.5.3 Parameter Space Mapping

The results of applying the sampling and interpolation algorithm are shown in Fig. 5.22. All interpolation methods are shown to work well with no one interpolation method being clearly best. This is due to all of the interpolation functions being a good approximation of the underlying response function on the length scale of of each interpolator. They all produce a small predictive error of approximately 4% of the maximum response amplitude after 600 model evaluations, calculated by comparing the interpolated result to an independent error set, which required approximately a week of simulation time. A slice of the response map is illustrated in Fig. 5.23 which shows a slice of the data for a fixed set of crack properties and a fixed probe rotation. This shows a realistic result; there is a strong reflection that corresponds to the reflection from the base of the hole and a weaker reflection from the crack which tails off as the crack decreases in height. This method provides a visual check that the response map produced is realistic.

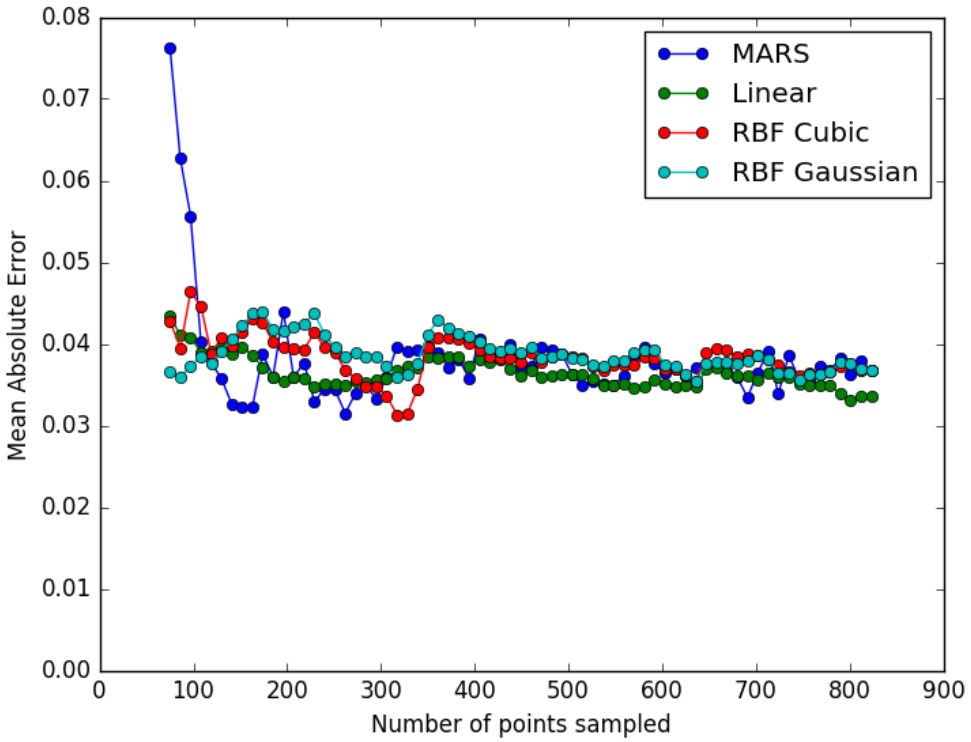

FIGURE 5.22: The mean absolute error in the prediction of the response space when compared to an independent error set for the six parameter inspection mapped using a finite element model. A range of interpolation methods are used: Multivariate Adaptive Regression Splines (MARS), linear interpolation, cubic interpolation, a cubic Radial Basis Function (RBF) and a Gaussian RBF. The error is calculated by comparison of interpolated results to an independet error set, normalised to the maximum response.

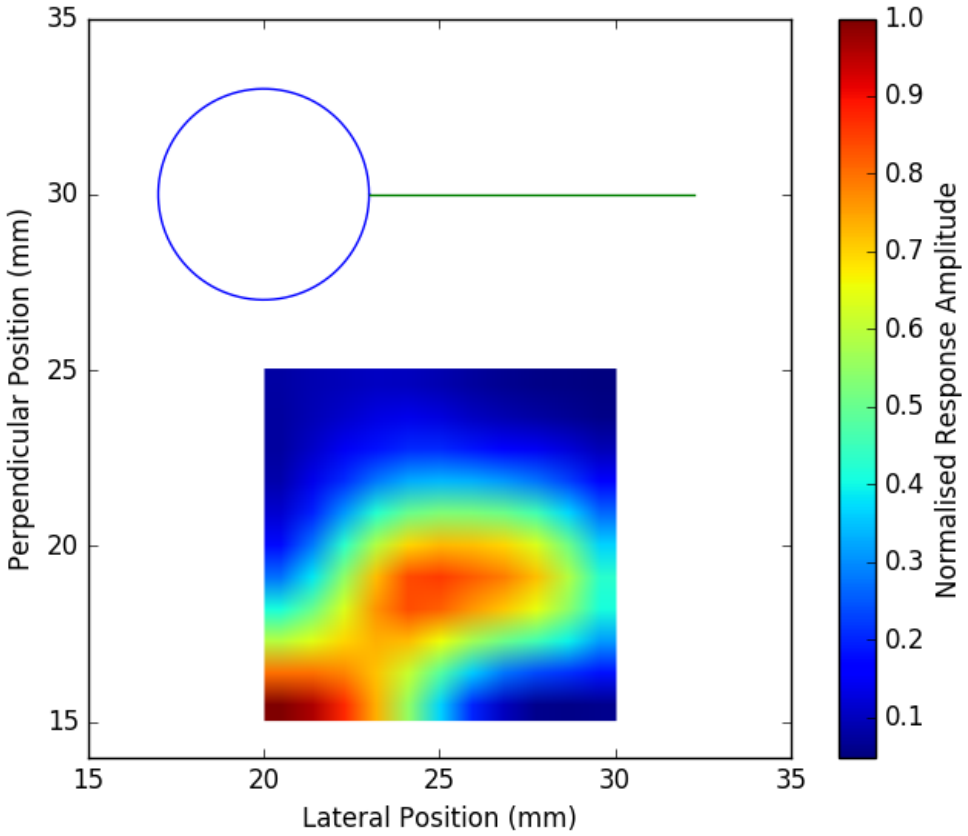

FIGURE 5.23: A slice of the interpolated data, plotting the response as a function of probe position. This can be used as a qualitative check of the quality of the interpolation. The response decreases with increasing lateral position due to the triangular profile of the crack.

5.5.4 Calculation of Reliability Metrics

In this inspection, the crack length is the parameter of interest. An initial estimate of the probability distribution is that all of the parameters are independent and are estimated as follows. The position of the probe in the lateral direction is a normal distribution with a mean position of 10 mm away from the root of the crack and a standard deviation of 4 mm. The position of the probe in the perpendicular direction is also a normal distribution with a mean of 2 mm from the root of the crack and a standard deviation of 4 mm. The rotation of the probe is estimated to be uniformly distributed. It is more likely that the crack grows radially out from the hole rather than at high angles therefore the defect angle follows a normal distribution with a mean of 0^o and a standard deviation of 15^o. All values of the length and height of the defect are assumed equally likely therefore these both follow a uniform distribution. It is also assumed that all combinations of all parameters are possible. The resulting PoD curve is shown in Fig. 5.25. This shows that the PoD increases as the length of the crack increases. This physically is reasonable as the size of the reflecting area of the defect increases as the length of the crack increases, therefore presenting a larger target for reflection. However, there is only a relatively small increase in the PoD from small crack sizes to large crack sizes, suggesting that the inspection is not particularly sensitive to this parameter.

The Sobol indices for the inspection parameters are shown in Fig. 5.24 as a function of the number of samples, calculated using the linear interpolator. The indices for the probe parameters converge as the number of samples increases however some indices show a sudden change at approximately 600 samples, most noticeably in some of the total sensitivity indices. As the Sobol indices are a metric of global sensitivity, these variations are caused by changes in the underlying response function, specifically the addition of more sampled points changes the interpolator sufficiently to induce a significant change in the response function. This is manifesting in increased interaction between the parameters which is causing greater changes in the total sensitivity indices compared to the first order indices.s

Figure 5.24 indicates that the perpendicular position of the probe is the most significant parameter in terms of direct, first order effect however the lateral position of the probe has the greatest total effect. This is due to greater interaction between the lateral position and the variation in the other parameters. These probe parameters are more significant than the crack parameters

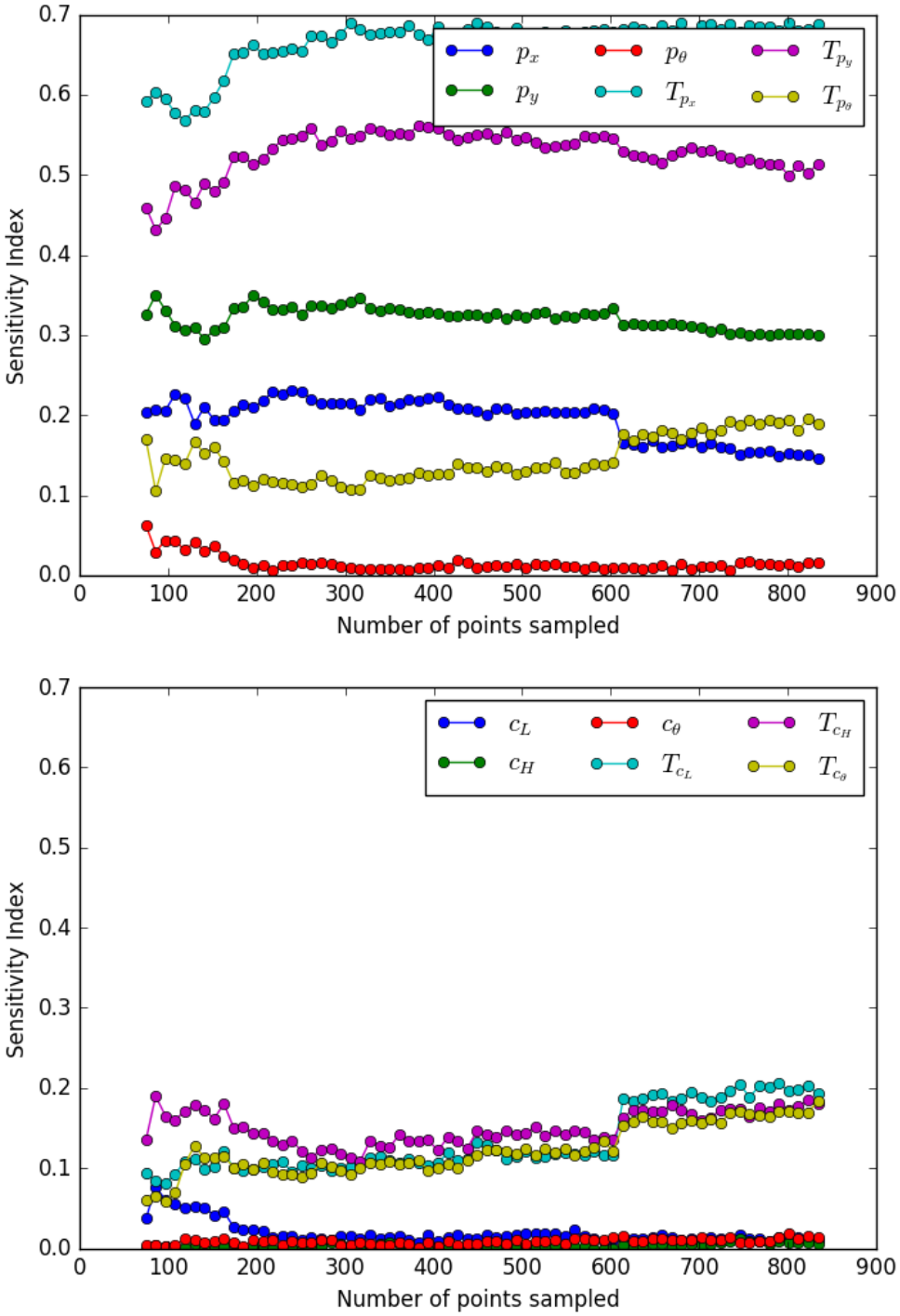

FIGURE 5.24: The first order and total order sensitivity indices, denoted by the prefix T, for the probe parameters (top) and the crack parameters (bottom). The sudden change in indices at approximately 600 samples is caused by a change in the underlying response function caused by the addition of more sampled points.

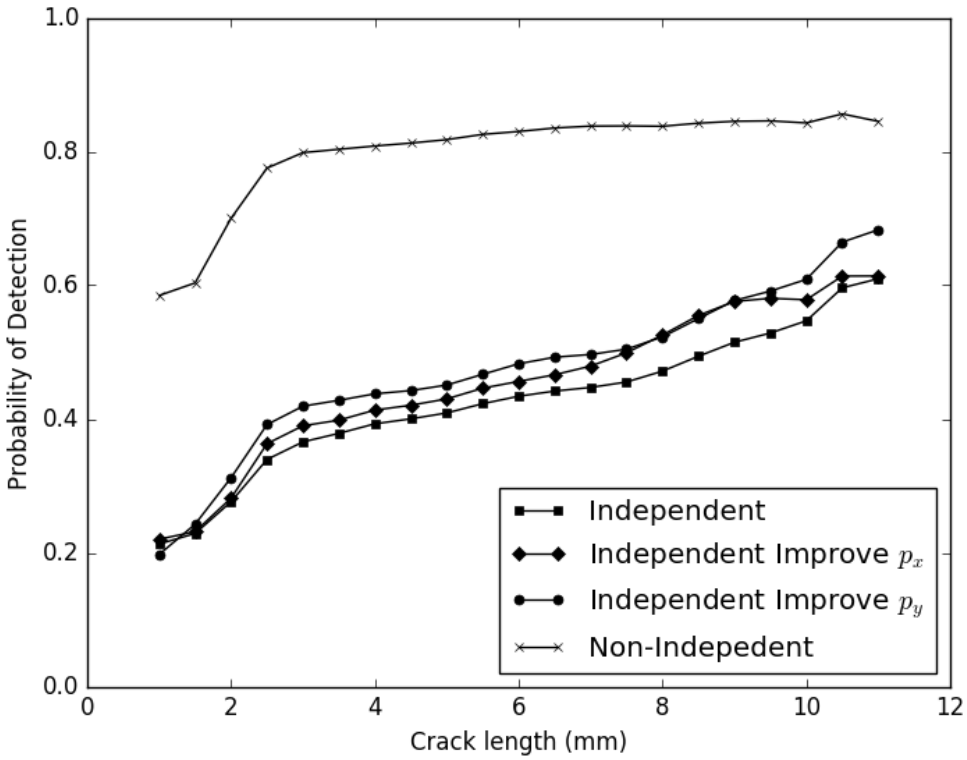

FIGURE 5.25: The probability of detection curve calculated using different probability distributions. The independent curve uses a probability function based on the assumption that all parameters are independent. The non-independent curve uses a probability function based on a interdependency of parameters. Using the independent probability function, the effect of halving the variance of the lateral probe position probability function is shown to have a less significant effect than halving the variance of the perpendicular probe position.

which suggests that the result of the inspection is dominated by human factors in this case. It is notable that the first order index for the crack length is approximately 2%. This shows that the inspection is very insensitive to this parameter which explains the very small variation across the PoD curve as the response is dominated by the other parameters. As none of the parameters have a small total sensitivity index, none can be discounted due to insignificance. Similarly, as none have a total sensitivity index approximately equal to the first order index, no parameters can be treated as independent and thus reduce the dimensionality of the mapped response space.

The perpendicular position of the probe has the greatest direct impact on the response, therefore this suggests that the operator should most focus their effort on this parameter and secondly on the lateral position of the probe. The effect of the operator focussing more on these parameters, and therefore halving the standard deviation of the variability in these axes, is shown in Fig. 5.25. This shows that focussing more upon the perpendicular position leads to a greater increase in the PoD than focussing upon the lateral position, which is expected given their relative importance. This information is potentially very useful to the developer of an inspection and can lead to more reliable assessments of structural integrity.

In reality, it is more likely that the human controlled parameters are not independent and that the operator will optimise all three degrees of freedom simultaneously to attempt to maximise the response which has some dependence on the properties of the crack. This can be estimated as follows. The position of the probe (x, y) is transformed into a perpendicular distance r from the crack and a distance t along the crack from its root at (x_0, y_0) by calculating

$$r = (y - y_0)\cos(c_\theta) + (x - x_0)\sin(c_\theta), \tag{5.5}$$

and

$$t = (y - y_0)\sin(c_\theta) - (x - x_0)\cos(c_\theta). \tag{5.6}$$

The probability distribution for the r coordinate is a normal distribution with mean of 10.5 mm and a variance which is itself a normal distribution function of the height of the crack. The mean is chosen as that is the optimal distance away from the crack to inspect it, given the angle of the beam and the thickness of the specimen. The variance is chosen to reflect the notion that cracks with shorter heights have a smaller distance from the optimal position in which they can be detected and the operator is more likely to have a final probe position

in this more narrow range. This distribution has a mean of the maximum crack height and a variance of 2 mm. The probability distribution for the t coordinate is again a normal distribution with a mean of 0 and a variance which is itself a normal distribution function of the length of the crack. The mean is chosen as it is the optimal location to inspect the crack. The variance is chosen to reflect the fact that shorter cracks have a smaller range around the optimal position in which they can be detected and the operator is more likely to finish with the probe in that range. This distribution has a mean of the maximum length of the crack and a variance of 3 mm. The rotation of the probe is also dependent on the rotation of the crack therefore the probability distribution for this parameter is a normal distribution with a mean of the rotation of the crack and a variance of 2^o.

Using this probability function results in a significant change in the PoD curve, as shown in Fig. 5.25. This suggests that the metrics are very sensitive to the choice of probability distribution. This significant variation in apparent capability given the same response function suggests that these probability distributions need careful, accurate definition to gain an accurate measure of inspection capability using metrics that are dependent on probability of variations. This also highlights that methods that make implicit assumptions to the nature of these distributions, such as the $\hat{a}$ vs a method, should be used with caution and evidence of the satisfaction of these assumptions should be provided alongside any results. This suggests that metrics which do not require definition of the probability functions, such as Sobol indices, may provide a more reliable measure of a technique's capability as they are not sensitive to changes in the likelihood of variations occurring.

With either probability distribution, the PoD never reaches 1. This is primarily caused by the possibility of cracks with very small heights which present a very small reflecting area and are therefore very difficult to detect by this method. It may be more realistic that a crack of a given length may have a minimum height in which case a subset of this data may be used with different probability distributions. This highlights the flexibility of this approach to calculating inspection metrics; as this will require a subset of the data, the probability distributions can be changed with no further model evaluations required.

The PoD may be improved by decreasing the response decision threshold however this will come at the cost of a greater false call rate. The effect of this

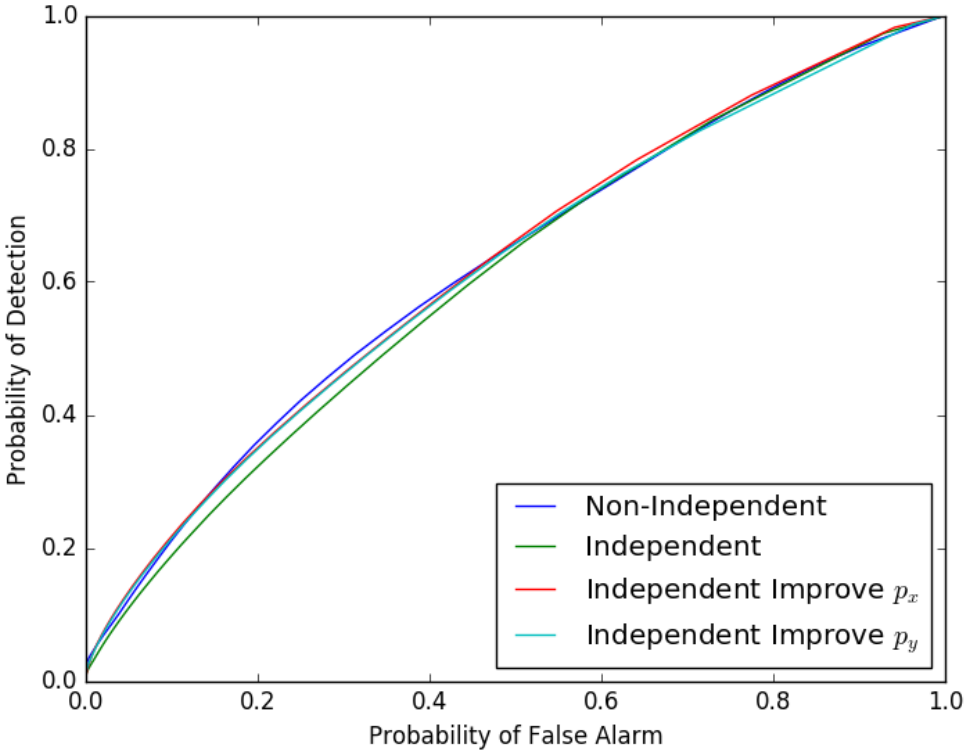

FIGURE 5.26: The Receiver Operator Characteristic (ROC) curve for a crack of 4 mm length. The curve is calculated by varying the response decision threshold.

can be investigated through a Receiver Operator Characteristic (ROC) curve calculated for a crack of length 4 mm as shown in Fig. 5.26. The ROC curves for all of the probability distributions are similar, they are all very close to the diagonal line of gradient 1 which would be the result of random guessing. This provides further evidence that this is an ineffective inspection as changing the threshold is causing similar changes in both the PoD and probability of false calls, showing that it is ineffective at distinguishing between false calls and true detections. The primary reason for this is that there is a coherent noise source in this inspection that has a response of magnitude of at least the magnitude of the response from the defect with a non-trivial probability of occurring. In this case, it is the strong reflection from the hole which is present independently of the geometry of the defect therefore it is always possible that a reflection from the hole may be misconstrued as a signal from a defect. This may be due to the incorrect choice of probability distributions and in practice the operator may be better at distinguishing the reflections from the hole rather than defect. This would result in a lower probability of these reflections being misconstrued as a reflection from a defect which would require a different definition of the probability function.

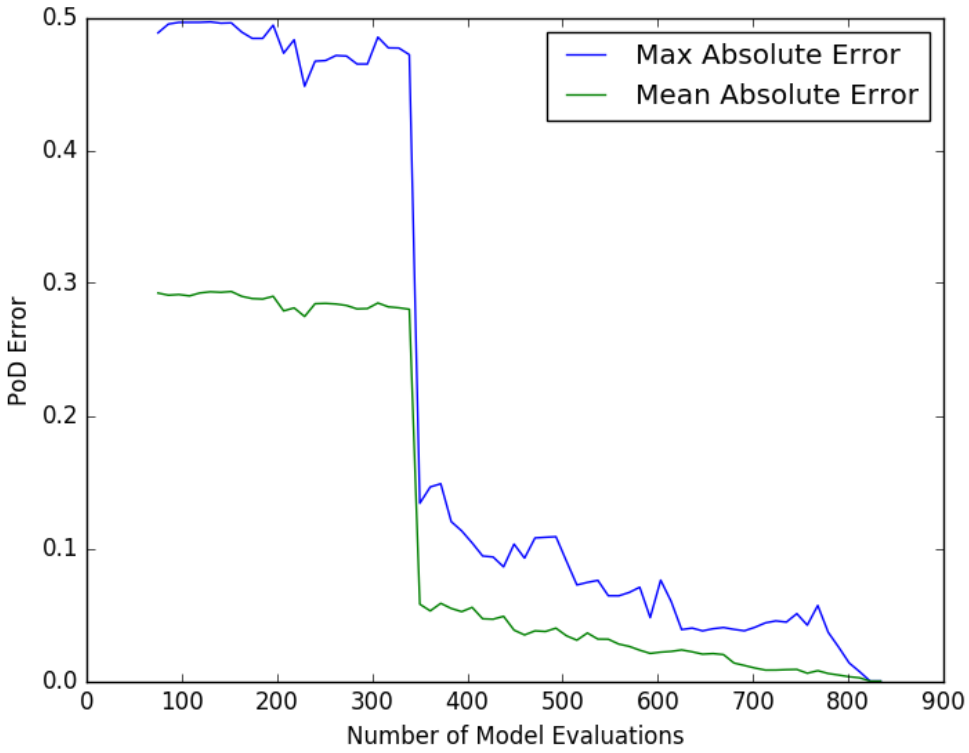

FIGURE 5.27: The mean and maximum absolute error in the calculation of the probability of detection curve for the eight parameter example, calculated using the linear interpolator and the non-independent probability distribution.

The convergence of the PoD curve can also be investigated. The result of this calculation for both the maximum absolute and mean errors in the calculation are shown in Fig. 5.27. These were calculated using the linear interpolator and the non-independent probability distribution with the final generated PoD curve taken as the best estimate of the true result, thus deviations were calculated relative to this. The results show a dramatic improvement after approximately 260 samples which can be attributed to the response map for the areas of highest probability becoming more accurate at this point. Following this, the maximum error and mean error both follow a downward trend, converging to the final PoD curve. This example demonstrates that is is possible to obtain a precise estimate of the PoD using a reasonable number of model evaluations. As discussed, this result will only be accurate if the probability distributions are representative of the reality of the inspection.

This process has demonstrated that it is possible to map the response function for several parameters in a reasonable time using numerical models coupled with an appropriate sampling and interpolation methodology. It is evident that more work is required to better estimate the probability distributions as these

have a significant impact on the outcome of the calculation of inspection metrics. Nonetheless, the process is capable of providing significant information on the relative importance of parameters and how the inspection can be optimised.

5.6 Summary

This chapter has demonstrated the applicability of the general method of calculating reliability metrics to the ultrasonic inspection of cracks emanating from fastener holes in wing skins. The single element, angle probe inspection was modelled using the finite element code Pogo and the necessary tools for automated model generation developed. The methodology was shown to work for increasing numbers of parameters and required only a relatively small number of model evaluations to achieve an accurate result. This was feasible using desktop hardware and even the most complex, eight parameter example required less than two weeks of simulation time, meeting the MoD's relatively short time scales and not requiring large scale computing resources. This process highlighted the need for accurate definitions of probability distributions to describe the likelihood of variations occurring as it is possible to obtain significantly different measures of probabilistic reliability, such as PoD, given different probability functions. This highlights the need to clearly validate and present evidence of the validation of the choice of probability distributions, whether this general metric calculation method, a Monte Carlo method or the $\hat{a}$ vs a method is used. Given this, the use of sensitivity indices as a metric of inspection capability, which are not dependent on the definition of probability functions, may present a more accurate measure of inspection capability. The sensitivity indices demonstrated that this inspection is very insensitive to the parameter of interest, the crack length, and is in fact dominated by the human factors present, specifically the location of the probe. This suggests that this inspection is relatively poor at detecting changes in crack length and in reality an alternative method should be sought. The sensitivity index therefore provides a good metric for directly comparing inspections without having to quantify their probabilities of variations occurring. This process has thus achieved the aims of this project and demonstrated a method of quantifying reliability of inspections primarily using numerical models of inspections rather than burdensome experimental trials.

Chapter 6

Conclusions and Future Work

6.1 Summary of Key Findings

1. A generalised approach to calculating metrics of the quality of an inspection, such as Probability of Detection (PoD), has been derived which makes no assumptions of the nature of the response of the inspection or probability of variations occurring.

2. A model based qualification approach has been demonstrated as a feasible method of reducing the burden upon experimental trials to demonstrate the reliability of an inspection technique, assessed by any desired quantitative metric, reducing the time and cost of qualification. This has been shown to be achievable for a relatively complex inspection using less than two weeks of simulation effort on desktop computing hardware, a much lower burden than performing experimental trials.

3. The use of appropriate sampling and interpolation algorithms has been shown to accurately map the response function using a relatively small number of model evaluations. This process is independent of the choice of probability distributions, therefore these can be changed in the calculation of inspection metrics without requiring further model evaluations.

4. The use of sensitivity analysis, specifically Sobol indices, has been demonstrated to allow the relative importance of parameters to be assessed, allowing some parameters to be discounted due to insignificance and others to be treated as independent parameters. This can accelerate the response mapping process and reduce the required simulation effort.

5. Sensitivity indices have also been shown to be a useful quantitative metric of the quality of an inspection as well as a tool to provide insight into the relative importance of parameters, highlighting which parameters should be heeded most effort to optimise. In the case of Sobol indices, as they are not reliant upon the definition of probability distributions, they can potentially provide a more accurate measure of the capability of an inspection technique.

6. The choice of probability distributions has been shown to have a significant impact on the calculated PoD curve. Further work is required to determine an accurate method of establishing these distributions.

6.2 Conclusions

The need for a fast and cost effective qualification methodology for inspections is well documented and this thesis has presented a model assisted methodology that can be used to demonstrate the capability of a technique. A general approach to calculating metrics of the quality of an inspection, making no assumptions as to the probability of variations occurring or the nature of variations in the response of the inspection, allows a more accurate measure of a technique's capability to be obtained.

This process requires knowledge of every possible outcome of an inspection and methods of achieving this using numerical models have been presented. The use of an appropriate sampling and interpolation methodology allows the response to be accurately quantified using only a smaller sub set of all possible inspections. The use of quantitative sensitivity analysis, specifically Sobol indices, allows the relative importance of parameters to be assessed and those with minimal impact ignored. Sobol indices also provide a novel metric of inspection capability, providing a single, unit-less metric of the capability of the technique. As it incorporates information of the Probability of Detection (PoD) as well as the Probability of False Alarm (PFA), it is a more complete metric than either of these in isolation. It is also independent of the choice of probability distributions, unlike other methods that attempt to combine both PoD and PFA such as the area under the ROC curve. As the choice of probability distributions has been shown to have a significant impact on the PoD, not being dependent upon this choice is a significant advantage.

These methods have been applied to an example ultrasonic inspection, developing the model to incorporate increasing numbers of parameters, initially investigating two parameter and increasing to eight. This process demonstrated that it is possible to efficiently map the response of the inspection using a reasonable number of model evaluations, importantly being feasible on desktop hardware rather than requiring the use of a high performance cluster. A 3D finite element model was used to map the response of the inspection and it has been demonstrated that this process can be completed in a matter of weeks, making it compatible with the MOD's requirement to qualify an inspection in under two months. The information generated in this process can be used to calculate metrics of inspection capability, including the PoD and PFA, which allows the quality of the inspection to be assessed. The Sobol index for the parameter of interest, the crack length, was found to be very low, suggesting that this is a poor inspection for detecting changes in the length of the crack. This was borne out in the calculation of the PoD curve for the crack length which showed a relatively small variation in the PoD over the range of crack lengths. The PoD was found to be very sensitive to the choice of probability distributions and therefore these require accurate estimation. The information this process generates allows the relative importance of parameters to be assessed, providing insight into how the inspection can be optimised. It was shown that focussing on optimising the most significant parameters, thus reducing their variability, allows the PoD to be increased. In this case, the human factors, specifically the location of the probe, were found to be the most significant parameters. This potentially guides operators to better focus their effort during an inspection on the parameters which have the greatest impact on the response, and thus improving the performance of the inspection.

A significant challenge to this methodology becoming widely used within the MOD is the need for a numerical modelling capability for the inspection of interest. In the example used in this work, the 3D finite element code Pogo FE was used. Significant work was invested in developing the necessary tool chain to allow this model to be used for qualification and obtaining the appropriate hardware. Therefore for this model to be used in the MOD, the technical expertise and resources need to be in place before qualification can commence. This is true for any model that is to be used, an optimised tool chain for the model needs to be established within the organisation to allow model assisted qualification to be performed. Should these be put in place, regular efficient qualifications of inspections should be possible. It should be noted that the time

and cost of achieving this may be cheaper and faster than obtaining the necessary samples for an experimental qualification, especially if the specimens are complex or rare. Furthermore, given the large numbers of inspection techniques that are developed numerically, often a numerical model exists that could be used to perform qualification therefore the obtaining of the model may not be such a significant barrier.

6.3 Future Work

The large amount of resources required to perform a rigorous, blind experimental PoD trial was a major barrier in this project to gaining experimental validation of the PoD curves generated in the example qualification and it was not possible to perform in the given time frame. This would therefore constitute the largest body of future work with the goal of gaining validation of the PoD curves. This will be performed by TWI with qualified operators as part of the larger Dstl program. This will form the validation case for the qualification protocol they are writing. Originally, that part of the project was due to run alongside this research however bureaucratic delays hindered progress. Another key avenue of future work is to determine accurate and efficient ways of determining the probability function for an inspection. This is a sufficiently broad area that it could be covered by several future theses and time did not allow for a rigorous investigation of this part of the methodology however there is already a significant body of literature on human factors in NDT that could leveraged. One avenue to assess this is to record operators performing inspections and build probability distributions from this. One key challenge of this will be to ensure that the inspection is being performed in as realistic a manner as possible and at least some part of this should be performed as part of the PoD trial performed by TWI. Ideally, these probability distributions will be reusable, that is they should be transferable between qualifications to a high degree. This ties in with the notion of a modular approach to qualification and that the time it takes to qualify a technique will be minimised if as many results of previous qualifications can be reused.

There are various improvements that could be made to the Pogo tool chain to accelerate the pre and post processing of models. The main improvement is to change the meshing output file into a binary format. This would eliminate the need to parse text files and massively accelerate the pre-processing of the model.

One further improvement, although it would require much greater alterations to the meshing program, would be to pass a pointer to the memory location of the nodes and elements once they are generated to the input file write, thereby removing the need to write out a node and element files. Various other tweaks to the tool chain could be implemented, primarily involving optimisations of the Python code used although the 80/20 rule certainly applies here, the majority of the time spent writing an input file is creating the mesh so this should be optimised first.

The sampling and interpolation methodologies presented here are shown to be capable of mapping the response function in a reasonable time however there are likely to be better methods. More effort should be expended on the sparse grid method and whether this can be tweaked to be more efficient as the underlying idea of the method is very appealing. There are also a wide range of underlying basis functions that could be used to form the sparse grid and if time allowed these would have been further investigated. An alternative approach would be the use of quasi-random low discrepancy sequences, such as Sobol sequences [88], and applying selective criteria to whether these points are sampled based on the magnitude of the local predictive error. Further tweaks could also be applied to the Enhanced Stochastic Evolution algorithm that could accelerate the generation of Latin Hypercube Designs (LHDs). An alternative use of these could be to generate smaller local LHDs in regions of rapidly changing gradient of the response function which could allow them to be used in a more adaptive manner.

As this work is feeding into the Dstl protocol, some future work will be to present these methods in a very accessible manner for practitioners of qualification. The hope is that this work, coupled with a protocol that captures the key details of implementing the methodology, will allow it to be widely used in industry in the future. Ideally, model assisted qualification will become a widely used method of qualifying techniques in anger, allowing the MOD and others to utilise novel inspection technologies faster and more efficiently.

Appendix A

Qualification Protocol Overview

This work is feeding into a protocol being written for Dstl. For completeness, the latest outline flowchart of the protocol is provided in Fig. A.1 below. This provides a high level overview of how the methods described in this thesis can be applied in practice. It is in essence a flow chart representation of the methodology described in Chapter 3.1.

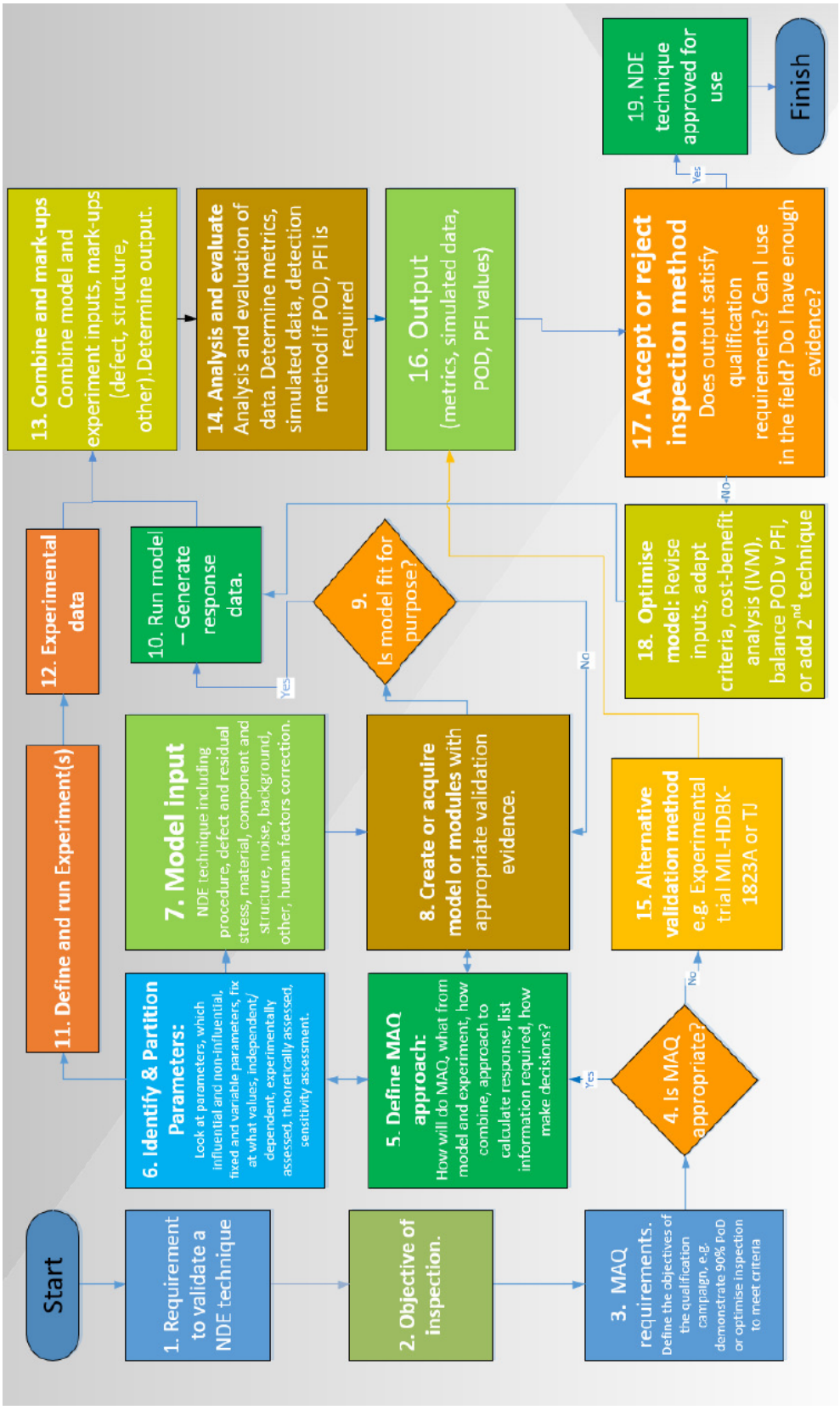

FIGURE A.1: An overview flowchart of the qualification protocol. Credit: Martin Wall, ESR Techology, Oxford, UK

Bibliography

[1] R. A. Smith and et al. *Paper 119 - Mechanism For Introduction of New Non-Destructive Testing Capability*. Tech. rep. Military Aircraft Structures Airworthiness Advisory Group, 2012.

[2] European Network for Inspection and Qualification (ENIQ). *European Methodology For Qualification of Non-Destructive Testing Third Issue*. Tech. rep. ENIQ, 2007.

[3] British Standards Institute. *BS EN 4179:2009 Aerospace Series. Qualification and Approval Of Personnel For Non-Destructive Testing*.

[4] ASC/ENSI. *MIL-HDBK-1823 Nondestructive Evaluation System Reliability Assessment*. Tech. rep. Department of Defense, 2009.

[5] R. A. Smith. *Paper 122 - Development of a Protocol for Acceptance of New NDT Capability in the Air Domain*. Tech. rep. Military Aircraft Structures Airworthiness Advisory Group, 2014.

[6] Health and Safety Executive (HSE). *Information for the Procurement and Conduct of NDT Part 4: Ultrasonic Sizing Errors and Their Implication for Defect Assessment*. Tech. rep. HSE, 2008.

[7] Amusement Device Safety Council. *Safety of Amusement Devices: Non-Destructive Testing*. Tech. rep. Amusement Device Safety Council, 2012.

[8] TWI and Royal & SunAlliance Engineering. *Best Practice for Risk Based Inspection as a Part of Plant Integrity Management*. Tech. rep. Health & Safety Executive, 2000.

[9] British Standards Institute. *PD CEN/TR 14748:2004. Non-destructive testing. Methodology for qualification of non-destructive tests*. Tech. rep. BSI, 2004.

[10] Alan P Berens. "NDE reliability data analysis". In: *ASM Handbook.* 17 (1989), pp. 689–701.

[11] Russel CH Cheng and TC Iles. "Confidence bands for cumulative distribution functions of continuous random variables". In: *Technometrics* 25.1 (1983), pp. 77–86.

[12] B. Thompson. "A Unified Approach to the Model Assisted Determination of Probbility of Detection". In: *AIP Conference Proceedings* 975 (2007), pp. 1685–1692.

[13] European Network for Inspection and Qualification (ENIQ). *ENIQ Recommmended Practice 6 - The Use of Modelling in Inspection Qualification*. Tech. rep. ENIQ, 2011.

[14] C.A. Harding, G.R. Hugo, and S. J. Bowles. "Application Of Model Assisted POD Using A Transfer Function Approach". In: *AIP Conference Proceedings* 1096 (2009), pp. 1792–1799.

[15] British Standards Institute. *General Requirements For Qualification And PCN Certification Of NDT Personnel*. Tech. rep. BSI, 2018.

[16] European Network for Inspection and Qualification (ENIQ). *Recommended general requirements for a body operating qualification of non-destructive tests*. Tech. rep. ENIQ, 2010.

[17] L. J. Nelson et al. "Ultrasonic Detectability of Potentially Closed Cracks from Cold-Worked Holes Under Loaded Conditions". In: *NDT 2007*. 2007.

[18] R. B. Thompson, W.Q. Meeker, and L. J. H. Brasche. "POD of Ultrasonic Detection of Synthetic Hard Alpha Inclusions in Titanium Aircraft Engine Forgings". In: *AIP Conference Proceedings* 1335 (2011), pp. 1533–1540.

[19] J. H. Kurz et al. "Reliability Considerations of NDT by Probability of Detection (POD) Determination Using Ultrasound Phased Array". In: *Engineering Failure Analysis* 35 (2013), pp. 609–617.

[20] A. A. Carvalho and et al. "Reliability of Non-Destructive Test Techniques in the Inspection of Pipelines Used in the Oil Industry". In: *International Journal of Pressure Vessels and Piping* 85 (2008), pp. 745–751.

[21] R. A. Smith et al. "An Ultrasonic Solution For Second-Layer Crack Detection". In: *Proc. NDT2004 - Annual Conf. of BInstNDT* (2004), pp. 223–228.

[22] S. Demeyer et al. "Transfer Function Approach Based on Simulation Results for the Determination of POD Curves". In: *AIP Conference Proceedings* 1430 (2012), pp. 1757–1764.

[23] Michel D Bode, Justin Newcomer, and Stephanie Fitchett. "Transfer function model-assisted probability of detection for lap joint multi site damage detection". In: *AIP Conference Proceedings*. Vol. 1430. 1. AIP. 2012, pp. 1749–1756.

[24] M Wall, SF Burch, and J Lilley. "Review of models and simulators for NDT reliability (POD)". In: *Insight-Non-Destructive Testing and Condition Monitoring* 51.11 (2009), pp. 612–619.

[25] C.A. Harding, G.R. Hugo, and S. J. Bowles. "Model Assisted POD for Ultrasonic Detection of Cracks at Fastener Holes". In: *AIP Conference Proceedings* 820 (2006), pp. 1862–1869.

[26] Rollo Jarvis, Peter Cawley, and Peter B Nagy. "Performance evaluation of a magnetic field measurement NDE technique using a model assisted Probability of Detection framework". In: *NDT & E International* 91 (2017), pp. 61–70.

[27] K. Smith et al. "Model Assisted Probability of Detection Validation for Immersion Ultrasonic Application". In: *AIP Conference Proceedings* 894 (2007), pp. 1816–1822.

[28] Dooyoul Lee et al. "Investigation of Detectable Crack Length in a Bolt Hole Using Eddy Current Inspection". In: *TRANSACTIONS OF THE KOREAN SOCIETY OF MECHANICAL ENGINEERS A* 41.8 (2017), pp. 729–736.

[29] Richard Howard and Frederic Cegla. "Detectability of corrosion damage with circumferential guided waves in reflection and transmission". In: *NDT & E International* 91 (2017), pp. 108–119.

[30] V Memmolo et al. "Model assisted probability of detection for a guided waves based SHM technique". In: *Health Monitoring of Structural and Biological Systems 2016*. Vol. 9805. International Society for Optics and Photonics. 2016, p. 980504.

[31] M. Carboni and S. Cantini. "A Model Assisted Probability of Detection approach for ultrasonic inspection of railway axles". In: *18th World Conference on Nondestructive Testing*. 2012.

[32] M. Carboni, S. Cantini, and C. Gilardoni. "Validation of the Rotating UT Probe for In-Service Inspection of Freight Solid Axles by Means of the MAPOD Approach". In: *5th European-American Workshop on Reliability of NDE*. 2014.

[33] J. S. Knopp et al. "Investigation of a Model Assisted Approach to Probability of Detection Evaluation". In: *AIP Conference Proceedings* 894 (2007), pp. 1775–1782.

[34] J.C. Aldrin and J. S. Knopp. "Case Study For New Feature Extraction Algorithms, Automated Data Classification, And Model Assisted Probability Of Detection Evaluation". In: *AIP Conference Proceedings* 894 (2007), pp. 257–264.

[35] J. C. Aldrin et al. "Model Assisted Probability of Detection Evaluation For Eddy Current Inspection Of Fastener Sites". In: *AIP Conference Proceedings* 1096 (2009), pp. 1784–1791.

[36] J. C. Aldrin and J. S. et al. Knopp. "Demonstration of model-assisted probability of detection evaluation methodology for eddy current nondestructive evaluation". In: *AIP Conference Proceedings* 1430 (2012), pp. 1733–1740.

[37] N. Maleo. "PICASSO imProved reliabIlity inspeCtion of Aeronautic structure through Simulation Supported POD". In: *4th International Symposium on NDT in Aerospace*. 2012.

[38] N. Dominguez et al. "Progress in POD Estimation: Methods and Tools". In: *4th International Symposium on NDT in Aerospace*. 2012.

[39] M. Li, R. B. Thompson, and W. Q. Meeker. "Physical Model-Assisted Probability Of Detection Of Flaws In Titanium Forgings Using Ultrasonic Nondestructive Evaluation". In: *Technometrics* 56 (2014), pp. 78–91.

[40] C. Mandache et al. "Numerical Modelling as a Cost-Reduction Tool for Probability of Detection of bolt hole eddy current testing". In: *Nondestructive Testing and Evaluation* 26 (2011), pp. 57–66.

[41] J. Enkvist. *A Human Factors Perspective On Non-Destructive Testing (NDT): Detection and Identification of Cracks*. Department of Psychology, Stockholm Univ., 2003.

[42] Health and Safety Executive (HSE). *Programme For The Assessment Of NDT In Industry (PANI3)*. Tech. rep. HSE, 2008.

[43] M Wall. "Modelling of inspection reliability". In: *Engineering Science & Education Journal* 6.2 (1997), pp. 63–72.

[44] Plate Inspection Steering Committee. *Evaluation of the PISC trials results*. Tech. rep. PISC, 1979.

[45] Plate Inspection Steering Committee. *Evaluation of the PISC-2 trial results*. Tech. rep. PISC, 1986.

[46] Plate Inspection Steering Committee. *Third programme for the inspection of steel components (PISC III): an introduction*. Tech. rep. PISC, 1988.

[47] *The Oxford English Dictionary*. Oxford University Press, 2018.

[48] W. D. Rummel. "Probability of detection (POD) is not NDT/E reliability". In: *AIP Conference Proceedings* 1511 (2013), pp. 1809–1816.

[49] L. Gandossi and K. Simola. "Framework for the Quantitative modelling of the European Methodology for Qualification of Non-Destructive Testing". In: *International Journal of Pressure Vessels and Piping* 82 (2005), pp. 814–824.

[50] L. Gandossi and K. Simola. *EUR 22675 EN: A Bayesian Framework for the Quantitative Modelling of the ENIQ Methodology for Qualification of*

Non-Destructive Testing. Tech. rep. European Commission Directorate-General Joint Research Centre Institue for Energy, 2007.

[51] L. Gandossi and K. Simola. "Application of a Bayesian Model for the Quantification of the European Methodology for Qualification of Non-Destructive Testing". In: *International Journal of Pressure Vessels and Piping* 87 (2010), pp. 111–116.

[52] Health and Safety Executive (HSE). *Tech Report 454 Probability of Detection (PoD) Curves: Derivation, Applications and Limitations*. Tech. rep. HSE, 2006.

[53] E. Anderson et al. *LAPACK Users' Guide*. Third. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1999. ISBN: 0-89871-447-8 (paperback).

[54] A.P. Berens. "Reliability Data Analysis". In: *Metals Handbook* 17 (1989), pp. 689–701.

[55] Forman S Acton. *Analysis of straight-line data*. Wiley New York, 1959.

[56] Calyampudi Radhakrishna Rao et al. *Linear statistical inference and its applications*. Vol. 2. Wiley New York, 1973.

[57] M Pavlovi, Kazunori Takahashi, and Christina Müller. "Probability of detection as a function of multiple influencing parameters". In: *Insight-Non-Destructive Testing and Condition Monitoring* 54.11 (2012), pp. 606–611.

[58] John F Monahan. *Numerical methods of statistics*. Cambridge University Press, 2011.

[59] Ilya M Sobol. "On quasi-monte carlo integrations". In: *Mathematics and computers in simulation* 47.2-5 (1998), pp. 103–112.

[60] Il'ya Meerovich Sobol'. "On the distribution of points in a cube and the approximate evaluation of integrals". In: *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 7.4 (1967), pp. 784–802.

[61] John H Halton. "Algorithm 247: Radical-inverse quasi-random point sequence". In: *Communications of the ACM* 7.12 (1964), pp. 701–702.

[62] Robert Piessens et al. *QUADPACK: a subroutine package for automatic integration*. Vol. 1. Springer Science & Business Media, 2012.

[63] Gene H Golub, Michael Heath, and Grace Wahba. "Generalized cross-validation as a method for choosing a good ridge parameter". In: *Technometrics* 21.2 (1979), pp. 215–223.

[64] Ron Kohavi et al. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. 2. Montreal, Canada. 1995, pp. 1137–1145.

[65] M. D. McKay, R. J. Beckman, and W. J. Conover. "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". In: *Technometrics* 21.2 (1979), pp. 239–245.

[66] M. Stein. "Large Sample Properties of Simulations Using Latin Hypercube Sampling". In: *Technometrics* 29.2 (1987), pp. 143–151.

[67] Ruichen Jin, Wei Chen, and Agus Sudjianto. "An efficient algorithm for constructing optimal design of computer experiments". In: *Journal of Statistical Planning and Inference* 134.1 (2005), pp. 268–287.

[68] *Numerical Recipes in C: The Art of Scientific Computing, author=Press, William H and Teukolsky, Saul A and Vetterling, William T and Flannery, Brian P, year=1992, publisher=Cambridge Univ. Press.*

[69] Youssef G Saab and Vasant B Rao. "Combinatorial optimization by stochastic evolution". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 10.4 (1991), pp. 525–535.

[70] Lawrence Davis. "Handbook of genetic algorithms". In: (1991).

[71] F. A. C. Viana, G. Venter, and V. Balabanov. "An algorithm for fast optimal Latin hypercube design of experiments". In: *International Journal for Numerical Methods in Engineering* 82.2 (2010), pp. 135–156.

[72] Atsuyuki Okabe. *Spatial tessellations*. Wiley Online Library, 1992.

[73] William H Press et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[74] Boris Delaunay. "Sur la sphere vide". In: *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* 7.793-800 (1934), pp. 1–2.

[75] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. "The quickhull algorithm for convex hulls". In: *ACM Transactions on Mathematical Software (TOMS)* 22.4 (1996), pp. 469–483.

[76] J. H. Friedman. "Multivariate adaptive regression splines". In: *The annals of statistics* 19.1 (1991), pp. 1–67.

[77] G. Jekabsons. *Adaptive Regression Splines toolbox for Matlab/Octave*. http://www.cs.rtu.lv/jekabsons/. 2015.

[78] Sergey Smolyak. "Quadrature and interpolation formulas for tensor products of certain classes of functions". In: *Soviet Math. Dokl.* Vol. 4. 1963, pp. 240–243.

[79] Michael Griebel, Michael Schneider, and Christoph Zenger. *A combination technique for the solution of sparse grid problems*. Citeseer, 1990.

[80] Volker Barthelmann, Erich Novak, and Klaus Ritter. "High dimensional polynomial interpolation on sparse grids". In: *Advances in Computational Mathematics* 12.4 (2000), pp. 273–288.

[81] Jochen Garcke. "Sparse grids in a nutshell". In: *Sparse grids and applications*. Springer, 2012, pp. 57–80.

[82] Dirk Pflüger. "Spatially adaptive refinement". In: *Sparse grids and applications*. Springer, 2012, pp. 243–262.

[83] Dirk Pflüger, Benjamin Peherstorfer, and Hans-Joachim Bungartz. "Spatially adaptive sparse grids for high-dimensional data-driven problems". In: *Journal of Complexity* 26.5 (Oct. 2010). published online April 2010, pp. 508–522. ISSN: 0885-064X.

[84] Miroslav K Stoyanov and Clayton G Webster. "A dynamically adaptive sparse grids method for quasi-optimal interpolation of multidimensional functions". In: *Computers & Mathematics with Applications* 71.11 (2016), pp. 2449–2465.

[85] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. Wiley, 2008.

[86] Toshimitsu Homma and Andrea Saltelli. "Importance measures in global sensitivity analysis of nonlinear models". In: *Reliability Engineering & System Safety* 52.1 (1996), pp. 1–17.

[87] Ilya M Sobol. "Sensitivity estimates for nonlinear mathematical models". In: *Mathematical Modelling and Computational Experiments* 1.4 (1993), pp. 407–414.

[88] Paul Bratley and Bennett L Fox. "Algorithm 659: Implementing Sobol's quasirandom sequence generator". In: *ACM Transactions on Mathematical Software (TOMS)* 14.1 (1988), pp. 88–100.

[89] P. Huthwaite. "Accelerated finite element elastodynamic simulations using the GPU". In: *Journal of Computational Physics* 257, Part A.0 (2014), pp. 687 –707.

[90] Jan Achenbach. *Wave propagation in elastic solids*. Vol. 16. Elsevier, 2012.

[91] Robert D Cook et al. *Concepts and applications of finite element analysis*. John Wiley & Sons, 2007.

[92] Anton Van Pamel et al. "Finite element modelling of elastic wave scattering within a polycrystalline material in two and three dimensions". In: *The Journal of the Acoustical Society of America* 138.4 (2015), pp. 2326–2336.

[93] Mickael Brice Drozdz. "Efficient finite element modelling of ultrasound waves in elastic media". PhD thesis. Imperial College London, 2008.

[94] Hang Si. "TetGen, a Delaunay-Based Quality Tetrahedral Mesh Generator". In: *ACM Trans. Math. Softw.* 41 (2015), 11:1–11:36.

[95] *Pogo Tool Library, howpublished = https://github.com/ab9621/PogoLibrary, note = Accessed: 2018-03-01 author = Ballisat, A.*

[96] Robin M Betz, Nathan A DeBardeleben, and Ross C Walker. "An investigation of the effects of hard and soft errors on graphics processing unit-accelerated molecular dynamics simulations". In: *Concurrency and Computation: Practice and Experience* 26.13 (2014), pp. 2134–2140.

[97] Bianca Schroeder, Eduardo Pinheiro, and Wolf-Dietrich Weber. "DRAM Errors in the Wild: A Large-Scale Field Study". In: *SIGMETRICS*. 2009.