Advanced statistical methods for analysis of NDE data

by

Yurong Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
William Q. Meeker, Jr., Major Professor
Bruce R. Thompson
Huaiqing Wu
Ranjan Maitra
Petrutza C. Caragea

Iowa State University

Ames, Iowa

2006

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Yurong Wang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my deep gratitude for my advisor, Professor William Q. Meeker, Jr., who has guided, encouraged and supported me in many different ways during my Ph.D. study. With his excellent guidance and helpful suggestions, my research has come so far. I am very grateful that I had the opportunity to study under his supervision. I would also like to thank Dr. Bruce R. Thompson, Dr. Huaiqing Wu, Dr. Ranjan Maitra, Dr. Petrutza C. Caragea for serving on my thesis committee and for many helpful comments and guidance on my research.

I am also grateful to many other people at the center for NDE, especially Ms. Lisa J. H. Brasche, Dr. Thomas Chiou, and Mr. Rick Lopez for their help during my research.

I want to take this opportunity to thank many other people I have worked with. They have extended my knowledge and kindly offered me various help. Only a few are named here: Ms. Angela Zuo, Mr. Chunwang Gao, Mr. Zhigang Zhou, Ms. Xia Xu and Mr. Gabriel Camano. I wish them all the best.

Last but not least, I would like to thank my husband, Linxiao Yu, and my daughter, Julia, for their continuous love and support. Without their support I would not have been able to complete this work. Very special thank you must go to my parents, my older sister for their utmost supports and encouragements that have made me go this far. Also I would like to thank my parents-in-law for their taking care of Julia during my thesis work.

ABSTRACT

Nondestructive Evaluation (NDE) uses noninvasive techniques to determine the integrity of a material, component or structure. In modern industry, NDE methods are often used in quality control and quality assurance. For example, ultrasonic inspection is a routine NDE method to detect flaws/defects in rotating components of jet engines. However, in any NDE system, there are random factors that can affect the performance and reliability of the system. Probability of detection (POD) is an important metric for quantifying NDE capability and reliability. The most commonly used POD assessment method is known as the $\hat{a}$ versus a method. However, the standard $\hat{a}$ versus a method can not be directly applied to some situations encountered in modern NDE operations. The objective of this research is to 1) extend the $\hat{a}$ versus a method to handle bivariate response allowing for data censoring and truncation. 2) extend the standard method to adjust for bias in POD estimates due to flaw sizing errors. 3) develop a more complete understanding of inspection variability by using statistical models to identify and quantify the variance components in NDE operations.

In Chapter 1, the standard $\hat{a}$ versus a method is extended to handle bivariate responses. The method of maximum likelihood (ML) estimation is used to deal with data censoring and truncation. To estimate the POD of a bivariate-response NDE system, a dual detection criterion is defined. The extended model is used to analyze two sets of available inspection data. In one set of inspection data, there were more flaw misses that could not be directly accounted for by the bivariate $\hat{a}$ versus a model. Extra modelling efforts were made to accommodate these flaw misses.

The standard $\hat{a}$ versus a method assumes that the flaw sizes are known without error. However, the true flaw size is usually not known exactly due to cost constraints. The measurement errors in flaw sizing will bias the POD estimates. In Chapter 2 of this thesis, we develop two

statistical models for adjusting for bias in POD estimates that is caused by flaw sizing errors. The models are fitted by using the ML estimation method. We present the results of simulation studies that show how the use of our models will reduce flaw-sizing bias and we demonstrate the use of the methods with simulated inspection data based on the collected real inspection data.

The model behind the standard $\hat{a}$ versus a method contains only one component of variance for the response. There are, however, many random factors introducing variability to NDE inspection. Excessive variability from various sources can degrade NDE inspection quality. There are strong needs to identify and quantify variability sources in NDE applications, as such information is needed to properly decide on strategies to reduce variability. In the Chapter 3 of this thesis, we develop the Bayesian hierarchial model to identify and quantify the variance components of inspection in the presence of data censoring. The Bayesian approach is demonstrated with simulated data and experimental data. The computations use MCMC simulation implemented in the in WinBUGS software.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

Nondestructive Evaluation (NDE) uses noninvasive techniques to determine the integrity of a material, component or structure or quantitatively measure some characteristic of an object. In modern industries, NDE is an important technique for quality control and assurance. It can be used to determine properties of materials, detect flaws in parts, or even classify flaws by size, shape and location (Olin and Meeker, 1996). Compared with the traditional destructive testing methods, it has the advantages of much lower cost and good repeatability. For example, to evaluate an anomaly within a critical component in a jet engine, the traditional metallographic examination method has to destroy the sample unit. Also, because of the destructive nature of this method, the destructive evaluation has no repeatability. NDE methods, on the other hand, do not destroy the unit, are relatively inexpensive, and are repeatable. With these advantages, NDE is widely used in the industries and plays an important role in:

- Process quality control
- Sample inspection of newly manufactured products to ensure quality
- Assuring that in-service parts are working safely and to increase system reliability
- Life extension of expensive components in systems such as aircraft and power generation equipment.

In any nondestructive inspection process, however, there are many factors that can affect the performance of inspection system. Examples of sources of variability include system alignment, material properties, flaw geometry, flaw orientation, operator differences, and so on. These

factors can be partitioned into three groups: factors relating to the inspection system $\underline{x}_{\mathrm{SYS}}$, factors relating to the material $\underline{x}_{\mathrm{PART}}$, and factors relating to the flaw itself ($\underline{x}_{\mathrm{FLAW}}$).

1. The factors $\underline{x}_{\mathrm{SYS}}$ relating to the NDE inspection system include the transducer parameters, scan resolution (mechanical increment of the scanning system in X and Y dimensions), system alignment/angulations, as well as operators in the experiment.

2. The factors $\underline{x}_{\mathrm{PART}}$ relating to the part to be inspected include part geometry (particularly the degree of curvature at the inspection location), material microstructure and anisotropy, surface roughness, etc.

3. The factors $\underline{x}_{\mathrm{FLAW}}$ characterizing a flaw include flaw size, shape, orientation, depth and density (e.g., percent nitrogen and degree of cracking and voiding for a hard alpha inclusions), etc.

All these factors contribute to inspection variability. This variability leads to the need to use a probabilistic characterization of NDE inspection capability. Probability of detection (POD) is an important metric for quantifying NDE capability caused by such uncertainty and it is an essential part of NDE applications.

POD related research is a relatively new topic in the NDE field. Berens (1989) described how methods for analyzing NDE inspection capability data have undergone a considerable evolution since the 1970's. Initially, a constant probability of detection of all flaws of a given size was postulated, and binomial distribution methods using hit-miss data were used to estimate the probability. "Hit" means that an NDE system response was interpreted as having detected a flaw while "Miss" means that an NDE system response was interpreted as not having detected a flaw. A more appropriate hit-miss analysis based on a binary regression versus flaw size was developed subsequently and was illustrated in various places including MIL-HDBK-1823 (1999). For a general treatment of binary regression, see Agresti (1990).

In the early 1980's, other methods of estimating POD were developed for surface defects with more general characteristics, using more advanced statistical methods. One widely-used method is known as the "$\hat{a}$ versus a" method. In NDE work, a is used to denote flaw size. The flaw-response signal is often translated into an estimate of flaw size and this leads to the

use of "$\hat{a}$" to denote the flaw-signal response, even for applications where such a translation is not used. Data with this nature are called "$\hat{a}$ versus a" data. Berens (1989) found that a natural logarithmic transformation on the response, $\hat{a}$ and the flaw size, a, data often obeys the commonly used normal distribution simple linear regression model.

Depending on the character of the NDE system, some data may be right censored due to saturation (e.g., a signal exceeding 100% full screen height on an oscilloscope). If the signal is below the noise level, it is a miss. If it is known that there was a particular miss (e.g., a seeded flaw that was not detected), the observation is left censored. Annis and others at Pratt and Whitney (private communication) extended the $\hat{a}$ versus a method to allow for censoring and this approach is also described in MIL-HDBK-1823 (1999). When there is a possibility that there are misses that are not recorded and remain unknown (as in field-find data), the finds can be viewed as having come from a left-truncated distribution and are said to be "left truncated data." Burkel, Sturges, Turker and Gilmore (1996) extended the $\hat{a}$ versus a method to allow for censoring and truncation. For truncated or censored data, ordinary least square are not appropriate. Maximum likelihood (ML) methods have capability of dealing with such data issues. Meeker and Escobar (1998) describe statistical methods to analyze censored or truncated data.

Burkel, Sturges, Tucker and Gilmore (1996) described the "effective reflectivity" or "Re" method. This method can be shown to be equivalent to $\hat{a}$ versus a method except that regression slope is constrained to be 1 (which has a physical basis for certain applications).

1.2 Motivation

The objective of this research is to provide statistical methods that can be used to help improve the NDE reliability by

- Extending the standard $\hat{a}$ versus a method for some advanced NDE applications;
- Identifying and quantifying the variance components of NDE inspection.

Specifically, this thesis covers the following three research topics: extending the standard "$\hat{a}$ versus a" method for bivariate responses encountered in modern NDE inspection systems;

developing statistical models to adjust the bias in POD estimates caused by flaw sizing errors; developing and illustrating statistical methods to identify and quantify the variance components of NDE process. The detailed discussion for each research topic is given below.

1.2.1 Bivariate $\hat{a}$ versus a Method Allowing Censoring and Truncation

Today, for the case of Hit/Miss data, POD is usually estimated by using binary regression. The $\hat{a}$ versus a method is used when data provide signal strength information. The standard $\hat{a}$ versus a method assumes a univariate response. In modern NDE operations, however, a bivariate response results from some inspection methods. To estimate the POD of these new NDE inspection methods, an extended model is needed to handle the bivariate responses. This part of our research is to extend the standard $\hat{a}$ versus a method to handle such bivariate responses.

The $\hat{a}$ versus a method can be extended to handle a bivariate response by using bivariate regression. The bivariate regression model assumes an underlying joint distribution in which the means of the marginal logarithm of bivariate signal values depend on flaw size, but the standard deviations and correlation of bivariate response do not depend on flaw size.

Data censoring and truncation also arise in NDE inspection systems having bivariate responses. Left-censored observations occur when a known flaw is missed in inspection. That is, if the existence of a flaw is known and the value of the measurement is below the threshold, the flaw is said to have been missed. Right censoring occurs in NDE applications due to saturation (i.e., observations that are so large that they exceed the upper bound of the measuring device). Truncation is similar to but different from censoring. In NDE applications, truncation usually is left truncation due to field flaw misses. The extended $\hat{a}$ versus a method allows for data censoring and truncation. The method of maximum likelihood (ML) estimation is used to handle data censoring and truncation. ML estimation is preferred over other methods because it has good statistical prosperities (e.g., the invariance property) and some asymptotic optimality properties (e.g. minimum variance).

As in the standard univariate method, detection criteria need to be specified to estimate the POD of a bivariate-response NDE system. We defined a dual detection criterion in the

extended model. Specifically, a detection can occur if either one of bivariate response exceeds its corresponding threshold.

As examples of application, the extended model was used to analyze two sets of available inspection data in our research. One data set came from a conventional inspection and the other came from a relatively new multi-zone method of inspection. In the conventional inspection data, there were more misses than could be directly accounted for by the extended $\hat{a}$ versus a method. Extra modelling efforts were made to accommodate those data misses.

1.2.2 $\hat{a}$ versus a Method Capable of Adjusting Flaw Sizing Errors

The $\hat{a}$ versus a method of analysis is widely used in the NDE field to estimate POD for various NDE inspection systems. The symbol "a" is used to denote flaw size. The flaw-response signal is often translated into an estimate of flaw size and this led to the use of the notation "$\hat{a}$" to denote the flaw-signal response. The basic idea behind the $\hat{a}$ versus a method is a simple linear regression with assumptions that the logarithm of $\hat{a}$ has a normal distribution with mean depending on flaw size and a constant standard deviation. If the $\hat{a}$ signal for a flaw is greater than the detection threshold, the flaw is detected, otherwise, it is missed. The standard $\hat{a}$ versus a method for POD computation assumes that the flaws in the available data have sizes that are known without error. Usually the true flaw size is not known exactly and must be inferred from some inexact method such as metallographic analysis based on only one or two slices through a flaw.

Results in the classical statistical literature indicate that such errors-in-variables (EV) will bias the estimated regression coefficients. The presence of measurement errors will also affect the linear regression in POD computations. Fuller (1987) introduced classical measurement error model, investigated the effects of measurement error on the ordinary least squares estimators, and provided methods to do correction. Carroll, Ruppert and Stefanski (1995) extended these ideas to cover the nonlinear measurement error model and provided general approaches to solve measurement error problem. The effect of EV issues has not been studied in NDE applications before. The objective of this part of our research is to adapt the classical measurement error model to NDE applications and extend the standard $\hat{a}$ versus a method to

adjust for measurement error in flaw sizing. Based on the knowledge of the flaw sizing process (metallographic study), we developed two measurement error models: the Burkel model and a geometrical model. The two models allow for data truncation and censoring using ML estimation.

1.2.3 Variance Components Analysis of NDE Inspections

As discussed before, there are several factors that can introduce variability into NDE inspection and thus affect the inspection performance. Improvement of inspection performance requires the identification and quantification of variability sources. This paper develops and illustrates the use of statistical methods that can be used to identify and quantify the variance components of NDE inspection in the presence of censoring and truncation.

The experimental data used in this research were taken from a manufactured "block" of material containing seeded defects of known size and character. This block was inspected according to an experimental design that will capture all different sources of variability.

We build the variability model for the experimental data. A Bayesian approach was used to analyze and quantify variability sources. The Bayesian approach can handle complicated problems of variance component analysis, even allowing for data censoring. Computations were done with the Winbugs software tools. Congdon (2003) illustrates the Bayesian approach to data analysis and modelling in various applications using WinBugs software.

1.3 Dissertation Organization

This dissertation consists of 3 main chapters, preceded by the present general introduction and followed by a general conclusion. Each chapter corresponds to a to-be-submitted journal article. Chapter 1 describes an extension of the standard $\hat{a}$ versus a method for bivariate response encountered in modern NDE inspection system. Chapter 2 describes the development of advanced statistical methods to adjust for bias caused by flaw sizing errors. Chapter 3 develops and illustrates statistical methods to identify and quantify the variance components in NDE processes.

References

Olin, B. D. and Meeker, W. Q. (1996), "Applications of Statistics in Nondestructive Evaluation," *Technometrics* 38, 95-112.

Berens, A. P. (1989), "NDE Reliability Data Analysis," *Metals Handbook* (9th ed., Vol. 17, *Nondestructive Evaluation and Quality Control*), Metals Park, OH: American Society for Metals pp. 689-701.

MIL-HDBK-1823 (1999), *Non-Destructive Evaluation System Reliability Assessment*, Standardization Order Desk, Building 4D, 700 Roberts Avenue, Philadephia, PA 19111-5094.

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley.

Burkel, R. H., Sturges, D. J., Tucker, W. T., and Gilmore, R. S. (1996), "Probability of Detection for Applied Ultrasonic Inspection," Review of Progress in Quantitative NDE, Vol. 15, edited by D. O. Thompson and D. E. Chimenti, Plenum Press, New York, NY, 1991-1998.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Carroll, R. J., Ruppert D. , and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London; New York, Chapman Hall.

Congdon, P. (2003), *Applied Bayesian Modelling*, New York: John Wiley & Sons.

CHAPTER 2. ASSESSMENT OF NONDESTRUCTIVE PROBABILITY OF DETECTION FOR INSPECTION WITH A BIVARIATE RESPONSE

A paper to be submitted to Technometrics

Yurong Wang and William Meeker

Department of Statistics

Iowa State University

Ames, IA 50011

Abstract

Nondestructive evaluation (NDE) methods are used widely in modern industry to assure the integrity of critical system components. Examples include rotating components of jet engines and heat-transfer tubes in nuclear power plants. There is an important need to quantify the probability of detection (POD) for NDE applications in both production quality control and in-service inspection for expensive components that degrade over time. The standard method of estimating POD, known as $\hat{a}$ versus a, uses a linear regression relating NDE signal response to flaw or defect size. This paper extends this standard one-dimensional POD estimation method for bivariate responses. The extended methods allow for truncation and censoring encountered in many NDE applications. Atypical flaw misses in one of our NDE application examples could not be directly accounted for by the the extended $\hat{a}$ versus a method. Extra modeling efforts were made to accommodate those flaw misses.

Key Words: Atypical flaw misses, Censoring, Nondestructive Evaluation (NDE), Probability of Detection (POD), Truncation

2.1 Introduction

2.1.1 Background

Nondestructive Evaluation (NDE) is an important technique for ensuring quality in certain industrial applications. It can be used to determine properties of material, detect flaws in components, or even classify flaws by size, shape and location (Olin and Meeker, 1996). Compared with the traditional destructive testing methods, it has the advantages of lower cost and good repeatability. For example, to evaluate an anomaly in a critical component in a jet engine, the traditional metallographic examination method has to destroy the component. NDE methods, on the other hand, do not destroy the unit, are relatively inexpensive, and are repeatable. Because of these advantages, NDE is widely used in certain industries and plays an important role in:

- Process quality control
- Sample inspection of newly manufactured products to ensure quality
- Assuring that in-service components are working safely and increase system reliability
- Life extension of expensive components in systems such as aircraft and power generation equipment.

In nondestructive inspection processes, however, there are many factors that can affect the performance of inspection system, such as material properties, flaw geometry, flaw orientation, operators and so on. All these factors contribute to inspection variability that requires a probabilistic characterization of NDE inspection capability. Probability of detection (POD) is an important metric for quantifying NDE inspection capability and is an essential part of NDE applications.

2.1.2 Related Work

Berens (1989) describes how methods for analyzing NDE reliability data underwent a considerable evolution between then and the 1970's. Initially, a constant probability of detection

of all flaws of a given size was postulated, and binomial distribution methods using hit-miss data were used to estimate the probability. A "hit" implies that an NDE system response was interpreted as having detected a flaw while a "miss" implies that an NDE system response was interpreted as not having detected a flaw. A more appropriate hit-miss analysis is based on a binary regression in which POD is modelled as a function of flaw size and is illustrated in various places including MIL-HDBK-1823 (1999). For a general treatment of binary regression, see Agresti (1990).

In the early 1980's, other methods of estimating POD were developed for surface defects with more general characteristics, using more advanced statistical methods. One widely-used method is known as the "$\hat{a}$ versus a" method. In NDE applications, a is used to denote flaw size. The flaw-response signal is often translated into an estimate of flaw size and this led to the use of "$\hat{a}$" to denote the flaw-signal response, even for applications where such a translation is not used.

Berens (1989) found that with a natural logarithmic transformation on the response $\hat{a}$ and the flaw size a, data can often be described by the commonly used normal distribution simple linear regression model. Burkel, Sturges, Tucker, and Gilmore (1996) describe the "effective reflectivity" or "Re" method. This method can be shown to be equivalent to $\hat{a}$ versus a method when the regression slope is constrained to be 1.

Depending on the character of the NDE system, some data may be right censored due to saturation (e.g., greater than 100% full screen height (FSH) on an oscilloscope). If the signal is below the noise level, it is a miss. If it is known that there was a particular miss (e.g., a seeded flaw that was not detected), the observation is left censored. Annis and others at Pratt and Whitney (private communication) extended the $\hat{a}$ versus a method to allow for censoring and this approach is also described in MIL-HDBK-1823 (1999). When there is a possibility that there are misses that are not recorded (as in field-find data), the finds can be viewed as having come from a left-truncated distribution and are said to be "left truncated data." Burkel, Sturges, Tucker, and Gilmore (1996) extended $\hat{a}$ versus a method to allow for truncation. For truncated or censored data, the ordinary least square method is not appropriate. Maximum likelihood (ML) methods have the capability of dealing with such data issues. Meeker and

Escobar (1998), for example, provided statistical methods to analyze data that are censored and truncated.

2.1.3 Motivation

The standard $\hat{a}$ versus a method assumes a univariate response. In modern NDE operations, however, some inspection methods provide a bivariate response. This paper develops and illustrates bivariate regression models and estimation methods that also allow for truncation and censoring. In one of our applications of the bivariate response NDE model, we encountered misses that could not be directly accounted for by the extended method. We show how to include model terms to accommodate those misses.

2.1.4 Overview

The remaining parts of this paper are organized as follows. Section 2 describes the FAA "Contaminated Billet Study (CBS)" that provided the data for our example and the motivation for this research. This section also describes the conventional and multizone inspection data that arose from these NDE ultrasonic studies. Section 3 reviews the univariate $\hat{a}$ versus a method. Section 4 presents the bivariate extension to the univariate $\hat{a}$ versus a method for estimating POD. Section 5 applies the bivariate $\hat{a}$ versus a model to the CBS multizone data. In the CBS conventional inspections, there were more misses than those could be accounted for directly by the $\hat{a}$ versus a method. In Section 6, we describe the use of accommodation model terms to handle the large number of misses. Section 7 applies the bivariate $\hat{a}$ versus a model with accommodation terms to conventional data. Section 8 gives conclusions and discussion about future research.

2.2 Experimental Data

2.2.1 Contaminated Billet Study

The safety of aircraft jet engines depends on the use of NDE inspection techniques for the detection of flaws in titanium alloys used in production of engine components. A major flaw

type of concern for titanium alloys used in production of jet engine fan discs is hard alpha defects. Hard alpha defects are more brittle than nominal material (CBS report, 2004) and can cause dangerous crack growth rates that have led to in-service disk failures. Because natural hard alpha flaws are very rare in rotor grade titanium, their availability for study and evaluation is limited. A melter, while producing titanium for a non-aerospace customer, found numerous natural hard alpha defects (a total 64 were detected) in 12 contaminated billets, from a single heat. The FAA purchased the 12 billets in 1994 in order to support NDE research efforts. FAA-funded inspection studies were conducted on the 12 contaminated billets in 1994 and 1995. In these studies, both conventional and multizone inspections were conducted.

2.2.2 Ultrasonic Inspection Methods

In the conventional study, inspections were made with a 5-MHz longitudinal, cylindrically focused transducer at normal incidence (i.e., incident angle of 0^o) and a 5-MHz refracted longitudinal, spherically focused transducer at angle incidence (i.e., incident angle of 9.6^o) to produce 45^o longitudinal wave. Bivariate responses were amplitude from the normal incidence transducer and the angle incidence transducer for each flaw.

In the multizone study, separate 5-MHz bi-cylindrical focused transducers were used to cover five different depth zones. For each flaw, the multizone data has a bivariate responses: ultrasonic signal amplitude (in % full screen height or FSH)and signal to noise ratio (SNR) for each flaw.

2.2.3 CBS Inspection Data

In the CBS study, to obtain the flaw morphology and area information, 10 of the 64 flaws were cut out and sectioned at 5-mil increments and studied metallographically. The flaw areas for the other flaws were estimated from multizone ultrasonic C-Scan images. In NDE applications involving actual field inspection, data can be censored (left or right) and/or left truncated. The multizone data in the CBS study are left truncated because there might have been missed flaws, but there is no information about the existence of such flaws is available. Based on the information from the multizone data, in the CBS conventional inspection, there was a large

number of flaw misses. The number of misses was so large that not all could be explained by the $\hat{a}$ versus a model. Such flaw misses might have different unknown causes. One possible cause is human factor mistakes. Because of this "referee" information provided in the multizone study, we assume that the conventional study data are not truncated. This is justified because the conventional truncation level, implied by the multizone truncation level, is small enough to ignore. The conventional data do have left censoring due to the misses and right censoring due to saturation.

2.3 The $\hat{a}$ versus a Method

2.3.1 Model

The $\hat{a}$ versus a model, introduced in Section 4.1.2, can be expressed as:

$$\log(\hat{a}) = \beta_0 + \beta_1 \log(a) + \epsilon, \tag{2.1}$$

where ϵ has a normal distribution with mean 0 and variance σ^2.

According to Equation (2.1) and the linear property of normal distribution, $\log(\hat{a})$ has a normal distribution. Let $y = \log(\hat{a})$, then

$$y \sim N(\beta_0 + \beta_1 \log(a), \sigma^2). \tag{2.2}$$

The ML estimates of β_0, β_1, and σ are typically used today in practice when there are data censoring and/or truncation (Burkel, Sturges, Tucker, and Gilmore, 1996).

2.3.2 POD

POD is the probability that the signal response $\hat{a}$ exceeds the threshold $\widehat{a}^{\mathrm{TH}}$. That is,

$$\begin{aligned} \mathrm{POD}(a) &= \Pr(\hat{a} > \widehat{a}^{\mathrm{TH}}; a) \\ &= 1 - \Phi\left(\frac{\log(\widehat{a}^{\mathrm{TH}}) - \beta_0 - \beta_1 \log(a)}{\sigma}\right), \end{aligned} \tag{2.3}$$

where Φ is the cumulative probability function of standard normal distribution. From Equation (2.3), it is easy to see that

1. POD decreases as $\widehat{a}^{\mathrm{TH}}$ increases, for a given flaw size a.

2. POD increases as flaw size a increases, for a given fixed threshold $\widehat{a}^{\mathrm{TH}}$.

Based on the ML estimators of β_0, β_1, and σ, the estimated POD is:

$$\widehat{POD}(a) = 1 - \Phi\left(\frac{\log(\hat{a}^{\mathrm{TH}}) - \hat{\beta}_0 - \hat{\beta}_1 \log(a)}{\hat{\sigma}}\right), \tag{2.4}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}$ are the ML estimates of β_0, β_1, and σ, respectively.

2.4 The Extended $\hat{a}$ versus a Method

2.4.1 Extending the Classical $\hat{a}$ versus a Method

The $\hat{a}$ versus a method can be extended to handle a bivariate response using bivariate regression. The bivariate regression model assumes an underlying joint distribution in which the means of the marginal logarithm of bivariate responses depend on flaw size, but the standard deviations and correlation of the bivariate responses do not depend on flaw size. In this application, flaw size is flaw area. In the rest of this paper, flaw area will be used. This bivariate $\hat{a}$ versus a model can be written as:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \beta_0^{y_1} & \beta_1^{y_1} \\ \beta_0^{y_2} & \beta_1^{y_2} \end{pmatrix} \begin{pmatrix} 1 \\ X \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \tag{2.5}$$

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{y_1}^2 & \sigma_{y_1}\sigma_{y_2}\rho \\ \sigma_{y_1}\sigma_{y_2}\rho & \sigma_{y_2}^2 \end{bmatrix}\right)$$

where $\beta_0^{y_1}$, $\beta_1^{y_1}$, $\beta_0^{y_2}$, $\beta_1^{y_2}$, σ_{y_1}, σ_{y_2}, and ρ are unknown parameters that need to be estimated.

The random error term $(\epsilon_1, \epsilon_2)'$ is assumed to have a bivariate normal distribution with mean $\underline{0}$, standard deviation σ_{y_1}, σ_{y_2}, and correlation ρ. According to these assumptions, $(Y_1, Y_2)'$ has bivariate normal distribution with density function:

$$f(y_1, y_2; \mu_{y_1}, \mu_{y_2}, \sigma_{y_1}, \sigma_{y_2}, \rho) = \frac{1}{2\pi\sigma_{y_1}\sigma_{y_2}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\, Q\right)$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(y_1-\mu_{y_1})^2}{\sigma_{y_1}^2} - 2\rho\frac{(y_1-\mu_{y_1})(y_2-\mu_{y_2})}{\sigma_{y_1}\sigma_{y_2}} + \frac{(y_2-\mu_{y_2})^2}{\sigma_{y_2}^2}\right].$$

The regression relationships

$$\begin{aligned}\mu_{y_1} &= \beta_0^{y_1} + \beta_1^{y_1} x \\ \mu_{y_2} &= \beta_0^{y_2} + \beta_1^{y_2} x \\ x &= \log(a)\end{aligned}$$

express the dependency of the distribution means on flaw area.

The joint cumulative distribution function (CDF) of Y_1 and Y_2 can be written as:

$$F_{(Y_1,Y_2)}(y_1,y_2) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f(y_1,y_2;\mu_{y_1},\mu_{y_2},\sigma_{y_1},\sigma_{y_2},\rho)dy_1dy_2. \tag{2.6}$$

2.4.2 Maximum Likelihood Analysis with Censoring and Truncation

2.4.2.1 Censoring and truncation

A left-censored observation occurs when the exact value of the response has not been observed and we have, instead, an upper bound on the response (e.g., pages 34-35 of Meeker and Escobar, 1998). In NDE applications, a left-censored observation occurs when a known flaw is missed. That is, the existence of a flaw is known and only an upper bound on the signal value is available. Right censoring arises when the exact value of the response can not be observed and there is only a lower bound on the response. Right censoring occurs in NDE applications due to saturation (i.e., observations that are so large that they exceed the upper bound of the measuring device). Figure 2.1 uses simulated data to illustrate bivariate left censored data. The top graph shows data without censoring. The bottom graph shows the corresponding left censored data. Due to left censoring, the exact values of the data with the responses below the thresholds in the top graph become unknown except that the responses are below the thresholds.

Truncation is similar to but different from censoring. Truncation occurs when a response can be observed only when it falls in particular range, outside of which the existence of the flaw

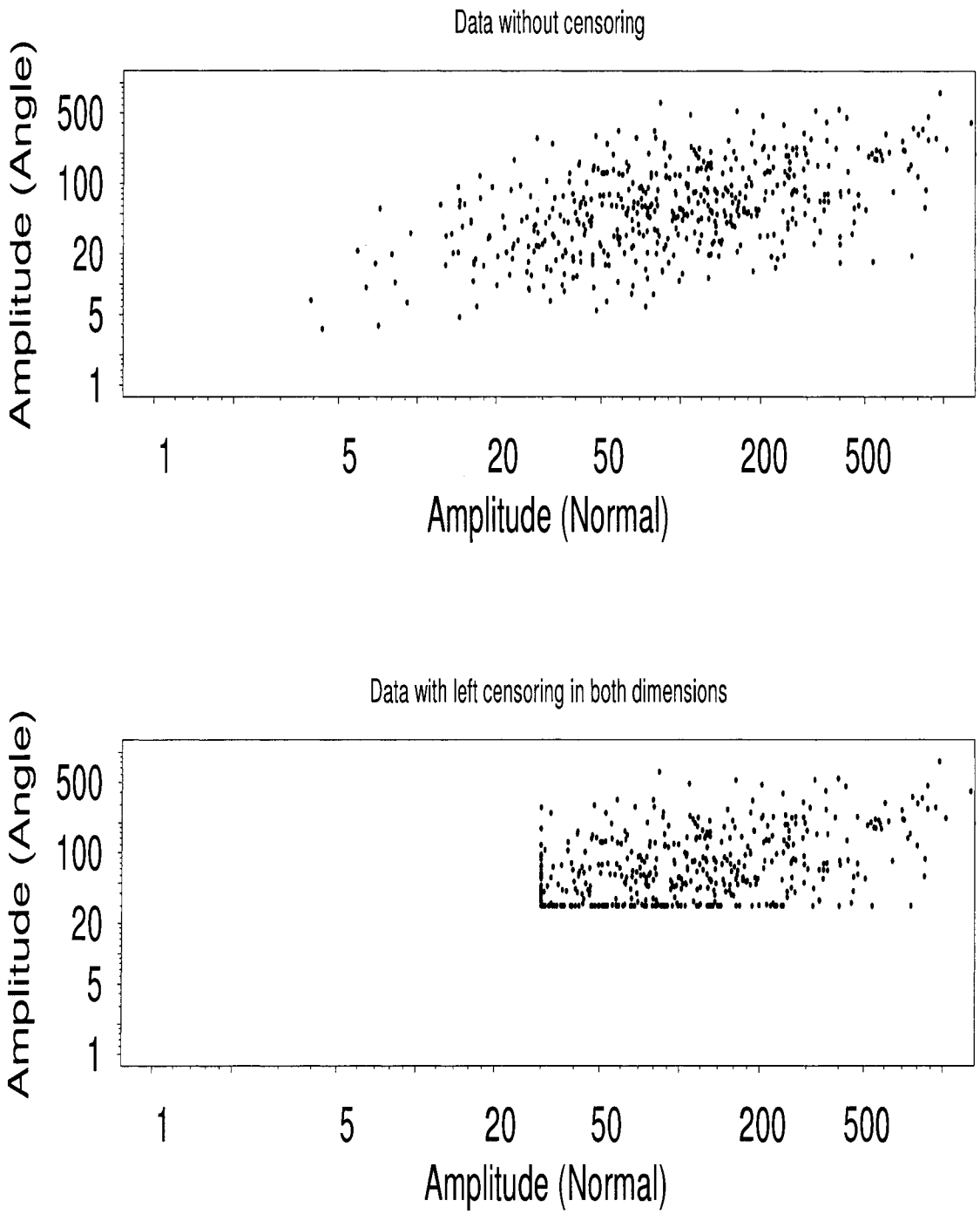

Figure 2.1 Simulated data used to illustrate bivariate left censoring.

is not known. Truncation usually is left truncation due to field flaw misses in NDE applications. The model for truncated data in the ML method is based on conditional probability.

2.4.2.2 Likelihood

The method of maximum likelihood (ML) estimation is one of the most versatile and popular statistical estimation techniques. Especially when there are complicating factors like censoring and truncation, ML estimation is preferred over other methods because it has good statistical properties (e.g., the invariance property) and some asymptotic optimality properties (e.g., minimum variance).

The likelihood function is central to estimation and inference. The natural log function is strictly increasing and thus the estimates maximizing log likelihood will also maximize the likelihood. Therefore, log likelihood is usually used because, numerically, it is much easier to maximize the log likelihood than likelihood. The log likelihood for the bivariate regression model defined in Equation (2.5) is

$$\mathcal{L}(\beta_0^{y_1}, \beta_1^{y_1}, \beta_0^{y_2}, \beta_1^{y_2}, \sigma_{y_1}, \sigma_{y_2}, \rho | y_1, y_2, x) \quad = \quad \sum_{i=1}^{n} \mathcal{L}_i \tag{2.7}$$

where $\mathcal{L}_i$ represents the contribution from the observation i.

When censoring and truncation mechanisms are active, the log likelihood function becomes more complicated, especially for the bivariate response. This is because the log likelihood needs to describe the probability behavior of parameters of statistical model for given complicated observed data. The contribution from each observation is defined in the next subsections.

The ML estimates are the values of $(\beta_0^{y_1}, \beta_1^{y_1}, \beta_0^{y_2}, \beta_1^{y_2}, \sigma_{y_1}, \sigma_{y_2}, \rho)$ that maximize the log likelihood in Equation (2.10). There are different ways to maximize the likelihood in Equation (2.10). For most practical problems involving complications like censoring or truncation, the likelihood must be maximized directly by using numerical methods, such as Newton's method.

2.4.2.3 Probability of the data contributions for different kinds of censoring

Because the combination of censoring and truncation can make the log likelihood function complicated, we first study the log likelihood contribution without truncation. For a univariate

response, there are three possibilities: left censoring, right censoring, and no censoring. For a bivariate response, we need to consider, potentially, 9 combinations of these observation types.

Figures 2.2 and 2.3 provide a visualization of the likelihood contributions from observations with different types of censoring, assuming no truncation. For example, with a doubly left censored observation, the bivariate response is below the given values in both dimensions. Thus the contribution to the likelihood is proportional to the probability of data falling into the rectangle of Area#1.1 in Figure 2.2. The contribution is computed as the cumulative probability, $F_{Y_1,Y_2}(y_1, y_2)$ defined in Equation (2.6). An uncensored response for Y_2 and a left censored response for Y_1 has a contribution equal to the probability of data falling into Slice#1.1 in Figure 2.2.

The likelihood contributions for doubly left censoring, left censoring only in Y_1, and other censoring types are defined as follows. Let S_1 and S_2 denote the data status (i.e., left censoring (L), no censoring (E), and right censoring (R)) with respect to y_1 and y_2, respectively. The probability of observation i with response (y_1, y_2) can be expressed as

$$\begin{aligned} \mathcal{L}_i &= \Pr(y_1^l \le y_1 \le y_1^u \quad and \quad y_2^l \le y_2 \le y_2^u) \\ &= \text{Prob}(y_1^l, y_2^l, y_1^u, y_2^u, S_1, S_2) \end{aligned}$$

- If $S_i = \text{L}$, then $y_i^l = -\infty$, $y_i^u = y_i^{CL}$, where y_i^{CL} is the left censoring level.
- If $S_i = \text{E}$, then $y_i^l = y_i - \delta$, $y_i^u = y_i + \delta$.
- If $S_i = \text{R}$, then $y_i^l = y_i^{CR}$, $y_i^u = \infty$, where y_i^{CR} is the right censoring level.

2.4.2.4 Density approximations for observation reported as exact values

Typically there is enough precision in measured response values (amplitude and SNR in the current context) that they are recorded as exact values. Referring to the definitions of $\text{Prob}(y_1^l, y_2^l, y_1^u, y_2^u, S_1, S_2)$ in the subsection 2.4.2.3, when δ approaches 0, the limit of $\text{Prob}(y_1^l, y_2^l, y_1^u, y_2^u, S_1,$ can be approximated by using a density function that is probability of the data. For example, when the width of the Slice#1.1, δ, approaches 0, the contribution can be written in probabil-

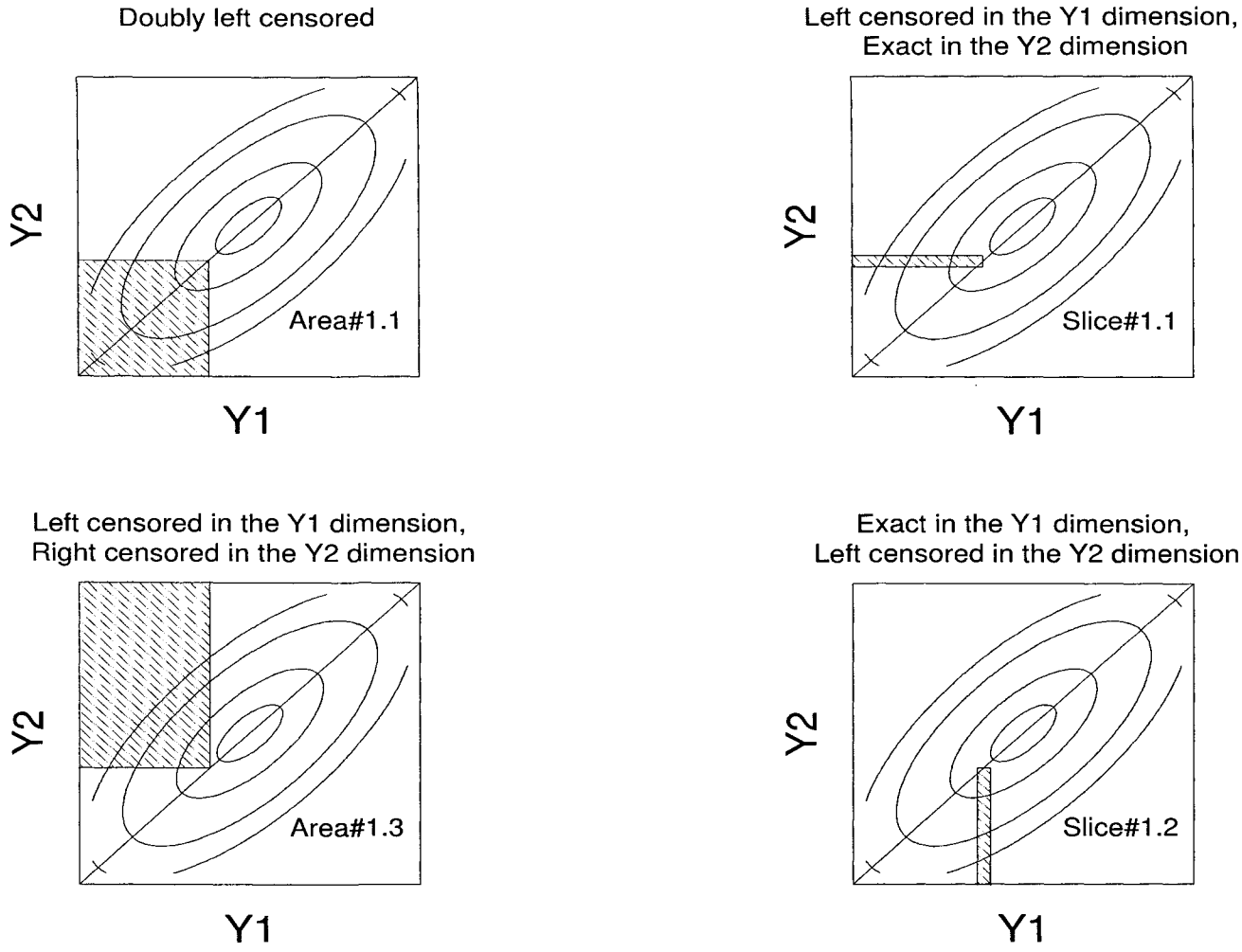

Figure 2.2 Contribution of censored data to likelihood in bivariate model (Continued).

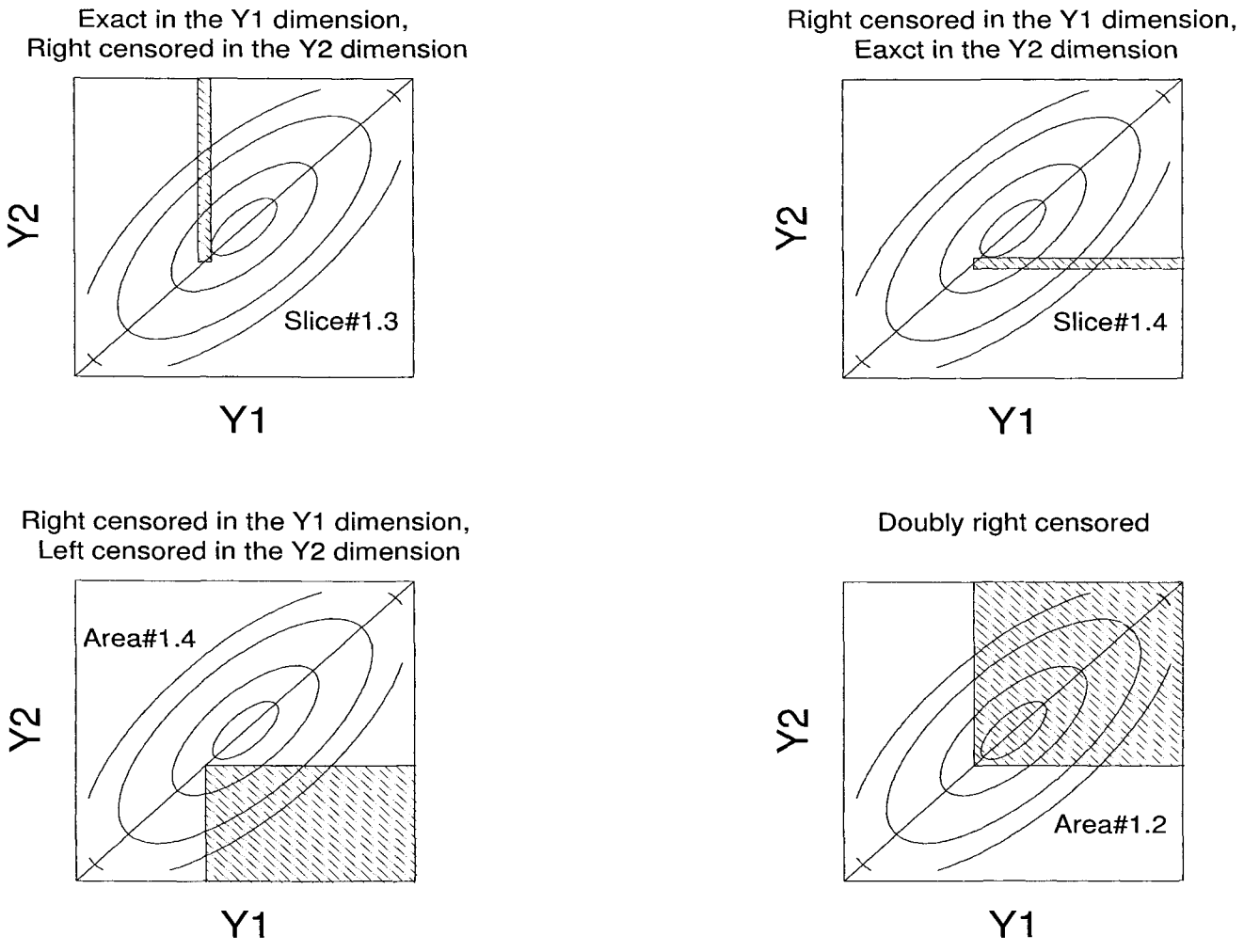

Figure 2.3 Contribution of censored data to likelihood in bivariate model.

ity density form as $F_{Y_1|Y_2}(y_1) \times f_{Y_2}(y_2)$. Using the density approximations provide definitions of the likelihood contributions that are easier to specify and compute. In particular, the log likelihood contribution is $\mathcal{L}_i = \log\left(\text{Contr}_i(y_1, y_2)\right)$, where $\text{Contr}_{(i)}(y_1, y_2)$ can be expressed as:

$$\text{Contr}_{(i)}(y_1, y_2) = \begin{cases} F_{Y_1,Y_2}(y_1, y_2) & S_1 = \text{L}, S_2 = \text{L} \\ F_{Y_1|Y_2}(y_1) \times f_{Y_2}(y_2) & S_1 = \text{L}, S_2 = \text{E} \\ F_{Y_1}(y_1) - F_{Y_1,Y_2}(y_1, y_2) & S_1 = \text{L}, S_2 = \text{R} \\ F_{Y_2|Y_1}(y_2) \times f_{Y_1}(y_1) & S_1 = \text{E}, S_2 = \text{L} \\ f_{Y_1,Y_2}(y_1, y_2) & S_1 = \text{E}, S_2 = \text{E} \\ (1 - F_{Y_2|Y_1}(y_2)) \times f_{Y_1}(y_1) & S_1 = \text{E}, S_2 = \text{R} \\ F_{Y_2}(y_2) - F_{Y_1,Y_2}(y_1, y_2) & S_1 = \text{R}, S_2 = \text{L} \\ (1 - F_{Y_1|Y_2}(y_1)) \times f_{Y_2}(y_2) & S_1 = \text{R}, S_2 = \text{E} \\ 1 - F_{Y_1}(y_1) - F_{Y_2}(y_2) + F_{Y_1,Y_2}(y_1, y_2) & S_1 = \text{R}, S_2 = \text{R}. \end{cases} \tag{2.8}$$

Note that,

- Some of the $\text{Contr}_{(i)}(y_1, y_2)$ contributions (e.g., $f_{Y_1,Y_2}(y_1, y_2)$) are in density form.
- $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ denote the CDF of the marginal distribution of Y_1 and Y_2, respectively.
- $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ denote the PDF of the marginal distribution of Y_1 and Y_2, respectively.
- $F_{Y_2|Y_1}(y_2)$ denotes the CDF of the conditional distribution of Y_2 given a particular value of Y_1. $F_{Y_1|Y_2}(y_1)$ denotes the CDF of the conditional distribution of Y_1 given a particular value of Y_2.

2.4.2.5 Contributions with censoring and truncation

When considering data truncation, we assume that $(Y_1, Y_2)'$ has a doubly truncated distribution with truncation levels y_1^{TL} and y_2^{TL}. This truncated distribution is adapted from the untruncated distribution whose CDF is denoted by $F_{(Y_1,Y_2)}(y_1, y_2)$. The CDF of the doubly

truncated bivariate distribution can be written as

$$G_{(Y_1,Y_2)}(y_1,y_2) = \begin{cases} \frac{F_{(Y_1,Y_2)}(y_1,y_2)-F_{(Y_1,Y_2)}(y_1^{lb},y_2^{lb})}{1-F_{(Y_1,Y_2)}(y_1^{\mathrm{TL}},y_2^{\mathrm{TL}})} & y_1 > y_1^{\mathrm{TL}} \quad \text{or} \quad y_2 > y_2^{\mathrm{TL}} \\ 0 & \text{Otherwise,} \end{cases} \tag{2.9}$$

where $y_1^{lb} = y_1^{\mathrm{TL}} \wedge y_1$ and $y_2^{lb} = y_2^{\mathrm{TL}} \wedge y_2$ ("$\wedge$" denotes minimum function). The term in the denominator of Equation (2.9) accounts for truncation and makes total probability under truncated density function equal to 1.

In some applications, there is truncation in only one dimension or there is no truncation. Left one-dimensional truncation is just a special case of left doubly truncation where either $y_1^{\mathrm{TL}} = -\infty$ or $y_2^{\mathrm{TL}} = -\infty$. The no truncation case (i.e., $y_1^{\mathrm{TL}} = -\infty$ and $y_2^{\mathrm{TL}} = -\infty$) was described in subsection 2.4.2.3.

For the case with both censoring and truncation, we can write the contribution for observation i with response (y_1, y_2) as

$$\mathcal{L}_i = \log\left(\frac{\mathrm{Contr}_i(y_1,y_2)}{1-F_{(Y_1,Y_2)}(y_1^{\mathrm{TL}},y_2^{\mathrm{TL}})}\right), \tag{2.10}$$

where $\mathrm{Contr}_i(y_1,y_2)$ is the same as described in Equation (2.8), except when either or both of Y_1 and Y_2 is left censored. When one or both of the responses are left censored, $\mathrm{Contr}_i(y_1,y_2)$ is defined as:

$$\mathrm{Contr}_{(i)}(y_1,y_2) = \begin{cases} F_{Y_1,Y_2}(y_1,y_2) - F_{Y_1,Y_2}(y_1^l,y_2^l) & S_1 = \mathrm{L}, S_2 = \mathrm{L} \\ (F_{Y_1|Y_2}(y_1) - F_{Y_1|Y_2}(y_1^l)) \times f_{Y_2}(y_2) & S_1 = \mathrm{L}, S_2 = \mathrm{E} \\ (F_{Y_1}(y_1) - F_{Y_1,Y_2}(y_1,y_2)) - (F_{Y_1}(y_1^l) - F_{Y_1,Y_2}(y_1^l,y_2^l)) & S_1 = \mathrm{L}, S_2 = \mathrm{R} \\ (F_{Y_2|Y_1}(y_2) - F_{Y_2|Y_1}(y_2^l)) \times f_{Y_1}(y_1) & S_1 = \mathrm{E}, S_2 = \mathrm{L} \\ (F_{Y_2}(y_2) - F_{Y_1,Y_2}(y_1,y_2)) - (F_{Y_2}(y_2^l) - F_{Y_1,Y_2}(y_1^l,y_2^l)) & S_1 = \mathrm{R}, S_2 = \mathrm{L} \end{cases} \tag{2.11}$$

where the definitions of y_1^l and y_2^l depend on the type of censoring and on which of Y_1 and Y_2 is truncated.

- When both Y_1 and Y_2 are truncated
  - If $S_1 = \mathrm{L}$ and $S_2 = \mathrm{L}$, then $y_1^l = y_1^{TL}$ and $y_2^l = y_2^{TL}$
  - Otherwise, $y_1^l = -\infty$ and $y_2^l = -\infty$
- When Y_i is truncated but Y_j is not $(i = 1, 2, \quad i \neq j)$
  - If $S_i = \mathrm{L}$, and $S_j = \mathrm{L}$, then $y_i^l = y_i^{TL}$ and $y_j^l = y_j$.
  - If $S_i = \mathrm{L}$, and $S_j \neq \mathrm{L}$, then $y_i^l = y_i^{TL}$ and $y_j^l = y_j$.
  - If $S_i \neq \mathrm{L}$, but $S_j = \mathrm{L}$, then $y_i^l = y_i^{TL}$ and $y_j^l = -\infty$.

2.4.3 POD for the Bivariate $\hat{a}$ versus a Model

The bivariate response $\hat{a}$ versus a method uses a dual detection criterion. A detection can occur if either of the bivariate responses exceeds its corresponding threshold. Thus, the POD can be written as

$$\begin{aligned} POD(y_1^{\mathrm{TH}}, y_2^{\mathrm{TH}}) &= \Pr(Y_1 > y_1^{\mathrm{TH}} \quad \text{or} \quad Y_2 > y_2^{\mathrm{TH}}) \\ &= 1 - F_{(Y_1,Y2)}(y_1^{\mathrm{TH}}, y_2^{\mathrm{TH}}), \end{aligned} \tag{2.12}$$

where, y_1^{TH} and y_2^{TH} are the thresholds for two dimensions, respectively.

2.5 Fitting the Bivariate $\hat{a}$ versus a Model to the CBS Multizone Data

2.5.1 The CBS Multizone Inspection Data

Section 2.2 describes the CBS study including the multizone and conventional study. The multizone study was conducted in 1994 and 1995. Because the 1994 inspection was incomplete, only the data from the multizone study in 1995 were used in our analysis in this section. The CBS multizone 1994 data are summarized in Table B.1. We dropped 4 of the 64 observations from the analysis because the flaw area was not available. In the multizone data, there are two measurements for each flaw: amplitude in percent of FSH and SNR. There were 43 of the 60 amplitudes (70%) that had amplitude greater than 100 percent FSH. Readings above 100

percent FSH are obtained by attenuating the signal by a known amount, taking the reading, and then translating back to the original scale.

The bivariate response for the multizone study is defined as

$$\begin{aligned} Y_1 &= \log(\hat{a}_1) = \log(\text{Amplitude}_{\text{multizone}}) \\ Y_2 &= \log(\hat{a}_2) = \log(\text{SNR}_{\text{multizone}}). \end{aligned}$$

Following the approach used by Burkle, Sturges, Tucker and Gilmore (1996), the truncation level for multizone is obtained by adding 10% FSH to the observed noise level. For the CBS data, the noise level was taken to be 20% of FSH for all observations, corresponding to the average noise level in the CBS billets.

2.5.2 ML Estimates for the Bivariate Model Parameters

The ML estimates of the model parameters are given in Table 2.1. Figure 2.4 shows the part of the bivariate regression model for the multizone data in which both amplitude and SNR depend on flaw area. The ellipse depicts a contour of the bivariate normal density function. The contour moves toward the northeast (stronger signal) as the flaw area increases.

Table 2.1 1995 Multi Data ML Estimation Results under Bivariate $\hat{a}$ versus a Model

Parameters	*MLE*	*Std.Err*	*95% Lower*	*95% Upper*
β_0^1	3.86	0.37	3.14	4.57
β_1^1	0.09	0.04	0.03	0.16
β_0^2	0.65	0.43	-0.20	1.49
β_1^2	0.12	0.04	0.04	0.20
σ_{ϵ_1}	0.43	0.04	0.36	0.51
σ_{ϵ_2}	0.51	0.05	0.42	0.61
ρ	0.77	0.05	0.66	0.87

2.5.3 Multizone POD

The southwest rectangle represents the no-detect region. This region is bounded by the thresholds for bivariate response:

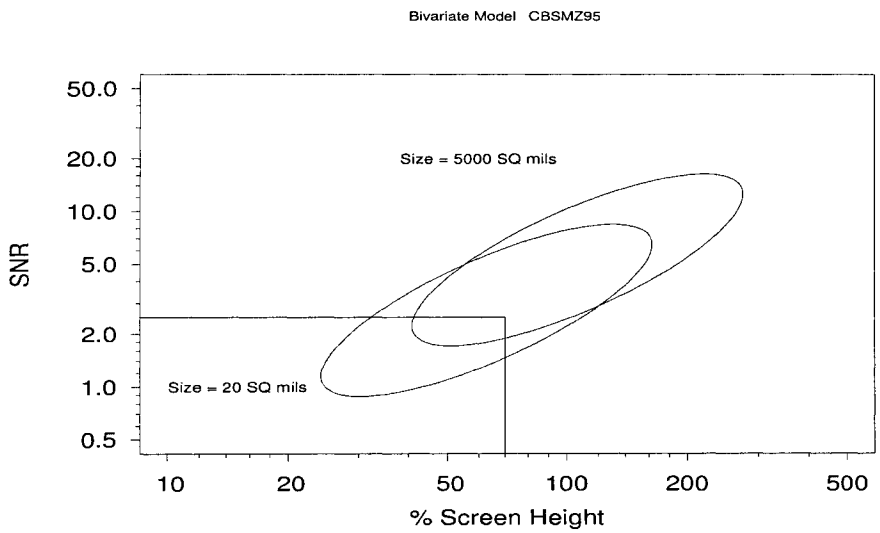

Figure 2.4 Plot illustrating the bivariate $\hat{a}$ vs. a model for CBS multizone 1995 data.

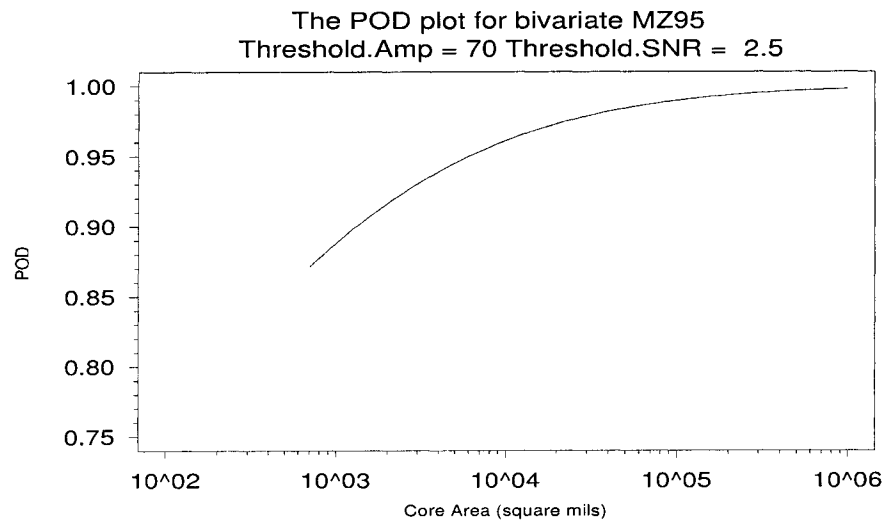

Figure 2.5 POD plot of bivariate $\hat{a}$ vs. a model for CBS multizone 1995 data.

- 70% FSH with calibration such that 80% FSH corresponds to the signal from a #2 flat bottom hole
- 2.5 SNR where, in the multizone detection rule, SNR is defined as

$$\text{SNR} = \frac{P_s - \mu_n}{P_n - \mu_n}.$$

Here P_s, P_n, and μ_n are the peak signal, the peak noise, and the mean noise, respectively and the noise distribution is taken pixels over a small square surrounding the signal.

A detection occurs if the signal amplitude exceeds 70% FSH or if the SNR exceeds 2.5. If an observation falls into the southwest rectangle, there is no detection. The probability of data falling into southwest rectangle (e.g., 1 - POD) will decrease as the flaw area increases. An estimate of POD curve for the CBS multizone inspection, computed by substituting the ML estimates into Equation (2.12), is shown in Figure 2.5. The curve is computed for core area greater than 600 square mils because there were no flaws on the data set with area less than 600 square mils.

2.6 The Bivariate $\hat{a}$ versus a Model for Atypical Misses with Accommodation

2.6.1 Typical and Atypical Misses

The CBS conventional data collected in 1994 were used as the basis for the work in this section. The bivariate responses are amplitudes from both the normal and the angle inspections. Based on the information from the multizone inspection, however, there were a substantial number of known misses in the conventional 1994 data. In particular, 44.3% of the known flaws were missed by both the normal and the angle inspections. Figure 2.6 shows the Hit/Miss data for CBS conventional normal and angle 1994 inspection data. Here a "hit" represents flaw detection in either normal or angle inspection. A "miss" represents flaw missing in both normal and angle inspection. The S-shape curve is the estimated POD using hit-miss analysis which is based on a binary logistic regression relating hit/miss to flaw size. In this binary regression

method, POD is modelled as a function of flaw size. Note that even some large flaws were missed. The conventional inspection data are summarized in Table B.1.

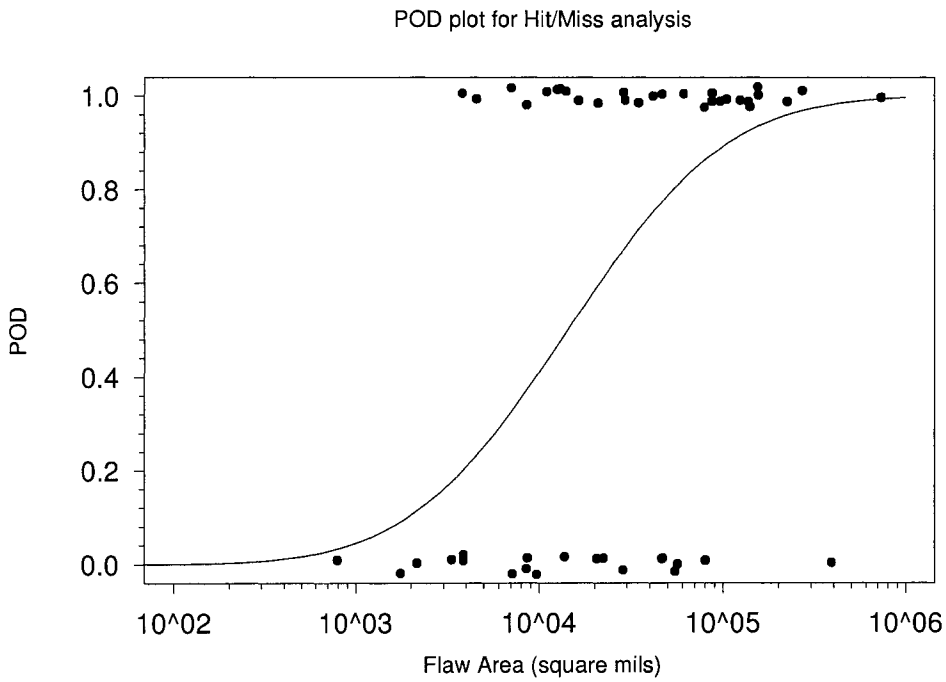

Figure 2.6 POD plot of hit/miss method for CBS conventional normal and angle 1994 data and POD estimated from a logistic regression model.

Because of these misses, the bivariate $\hat{a}$ versus a model does not provide an adequate description of the data (i.e., the $\hat{a}$ versus a model does not fit the data well). In order to develop a model to accommodate the large number of misses, we assume that there are two types of misses:

- Type I (typical) misses have responses that follow the standard $\hat{a}$ versus a model, but due to chance, had a signal below the threshold.
- Type II (atypical) misses that might have had a signal above the threshold, but was missed due to some other cause or causes.

We also assume that there is no information available that would allow us to precisely assign the known misses into these categories.

2.6.2 Likelihood for the Bivariate Response Model with Accommodation Terms

In order to have a regression model that provides an adequate fit to the conventional data, an accommodation term is used in the model. This term says there is a probability of an atypical miss that depends on flaw area. This term, for example, might account for flaw misses due to serious human factors errors. The two plots in Figure 2.7 show observations with only typical misses on the top and with both typical and atypical misses on the bottom. Comparing these two graphs, we can see that many data points that were in the upper right corner in the top plot move to the threshold boundaries in the bottom graph, indicating atypical misses in one dimension or the other. Such a reduction in data points in the upper right corner is caused by atypical flaw misses. In order to write the log likelihood function clearly, we break it into four parts corresponding to the two possible outcomes in the two different dimensions: HH, MM, HM and MH. Here H represents a "hit" response corresponding to a signal above threshold, while M represents a "miss" for which the response is below threshold. The first (second) position reflects the normal (angle) response. Let p_1, p_2, and p_3 be the probability of an atypical miss in normal only, angle only, or both normal and angle inspections, respectively. Then the log likelihood is

$$\begin{aligned} \mathcal{L} = \mathbf{C} \quad &+ \sum_{HH} \log\left[\Pr(HH)\right] + \sum_{MM} \log\left[\Pr(MM)\right] \\ &+ \sum_{MH} \log\left[\Pr(MH)\right] + \sum_{HM} \log\left[\Pr(HM)\right] \end{aligned} \tag{2.13}$$

where, C does not depend on any unknown parameters, the summations are over all flaws in the respective categories, and

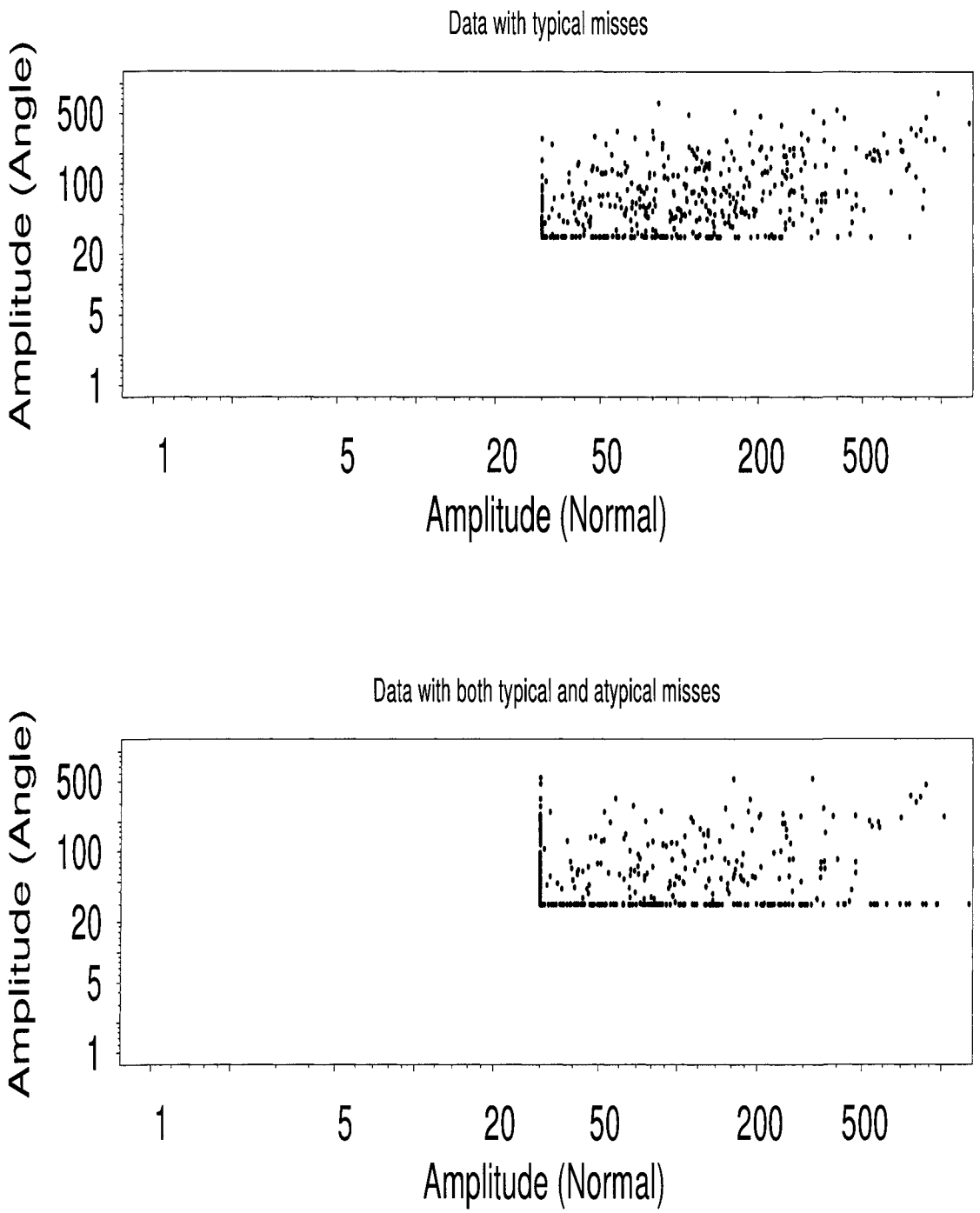

Figure 2.7 Censored data due to typical and atypical misses.

$$\begin{aligned}\Pr(MM) = \quad & (1-p_1-p_2-p_3)\times(Area\#1.1)+p_1\times(Area\#2.1)+p_3 \\ + \quad & p_2\times(Area\#2.2)\end{aligned} \tag{2.14}$$

$$\begin{aligned}\Pr(MH) = \quad & (1-p_1-p_2-p_3)\times\left[(Slice\#1.1)^{\delta_A}+(Area\#1.3)^{1-\delta_A})\right] \\ + \quad & p_1\times\left[(Slice\#2.1)^{\delta_A}+(Area\#2.3)^{1-\delta_A}\right]\end{aligned} \tag{2.15}$$

$$\begin{aligned}\Pr(HM) = \quad & (1-p_1-p_2-p_3)\times\left[(Slice\#1.2)^{\delta_N}+(Area\#1.4)^{1-\delta_N})\right] \\ + \quad & p_2\times\left[(Slice\#2.2)^{\delta_N}+(Area\#2.4)^{1-\delta_N}\right]\end{aligned} \tag{2.16}$$

$$\begin{aligned}\Pr(HH) \propto \quad & (1-p_1-p_2-p_3)\times\left[f_{(Y_N,Y_A)}(y_N,y_A)\right]^{\delta_1}\times(Slice\#1.4)^{\delta_2} \\ & \times(Slice\#1.3)^{\delta_3}\times(Area\#1.2)^{1-\delta_1-\delta_2-\delta_3},\end{aligned} \tag{2.17}$$

where p_1, p_2 and p_3 are the probabilities of an atypical miss in normal only, angle only, or both normal and angle inspections, respectively and are modelled in the different accommodation models in Section 2.6.3. The indicator functions are defined as:

- δ_A is 1 if the response in the angle inspection is saturated. Otherwise, it is 0. δ_N is the same as δ_A except corresponding to normal inspection.
- δ_1 is 1 if a flaw is detected without saturation in both inspections. Otherwise, it is 0.
- δ_2 is 1 if a flaw is detected in both inspections, but with saturation in the normal inspection. Otherwise, it is 0.
- δ_3 is 1 if a flaw is detected in both inspections, but with saturation in angle inspection, Otherwise, it is 0.

For a flaw that is missed in both inspections (MM), there are four probabilities in Equation (2.14): atypical miss in both inspections; typical miss in both inspections; typical miss in

one inspection and atypical miss in the other inspection. For a flaw that is missed in the normal inspection but detected in the angle inspection (MH), there are also four possibilities: typical flaw miss in the normal dimension and flaw detection in the angle dimension; typical flaw miss in the normal dimension and flaw detection and response saturation in the angle dimension; atypical flaw miss in the normal dimension and flaw detection in the angle dimension; atypical flaw miss in the normal dimension and flaw detection and response saturation in the angle dimension; HM and HH are similar.

Figure 2.8 illustrates the contributions to the log likelihood for observations with atypical misses. In this figure, we use Y_1 to represent Y_N and Y_2 represent Y_A. For example, an observation with atypical miss in Y_1 and typical miss in Y_2, means that the value of Y_1 might have been any positive number but its value is unknown, while the value of Y_2 can be any positive number below the corresponding censoring level. Thus the contributions of such observations are proportional to the probability of data falling into Area#2.1.

The log likelihood contributions (Contr_i) for observations without atypical misses can be found in Equation (2.8). For example, $Area\#1.1$ in Equation (2.14) can be represented as $F_{Y_1,Y_2}(y_1, y_2)$ in Equation (2.8). Let S_1 and S_2 denote the possible data status [i.e., left censoring, including left censoring due to typical flaw misses (L) and left censoring due to atypical flaw misses (LA), no censoring (E) and right censoring (R)] with respect to y_1 and y_2, respectively. The log likelihood contribution for the observations with atypical flaw misses, corresponding to Figure 2.8, can be written as:

$$\text{Contr}_{(i)}(y_1, y_2) = \begin{cases} F_{Y_2}(y_2) & S_1 = \text{LA}, \quad S_2 = \text{L}, \quad \text{(Area\#2.1)} \\ F_{Y_1}(y_1) & S_1 = \text{L}, \quad S_2 = \text{LA}, \quad \text{(Area\#2.2)} \\ f_{Y_2}(y_2) & S_1 = \text{LA}, \quad S_2 = \text{E}, \quad \text{(Slice\#2.1)} \\ f_{Y_1}(y_1) & S_1 = \text{E}, \quad S_2 = \text{LA}, \quad \text{(Slice\#2.2)} \\ (1 - F_{Y_2}(y_2)) & S_1 = \text{LA}, \quad S_2 = \text{R}, \quad \text{(Area\#2.3)} \\ (1 - F_{Y_1}(y_1)) & S_1 = \text{R}, \quad S_2 = \text{LA}, \quad \text{(Area\#2.4)} \end{cases} \tag{2.18}$$

where $F_{Y_1}(y_1)$ and $f_{Y_1}(y_1)$ are the CDF and PDF, respectively, of the marginal distribution of Y_1. $F_{Y_2}(y_2)$ and $f_{Y_2}(y_2)$ are the CDF and PDF, respectively, of the marginal distribution of Y_2.

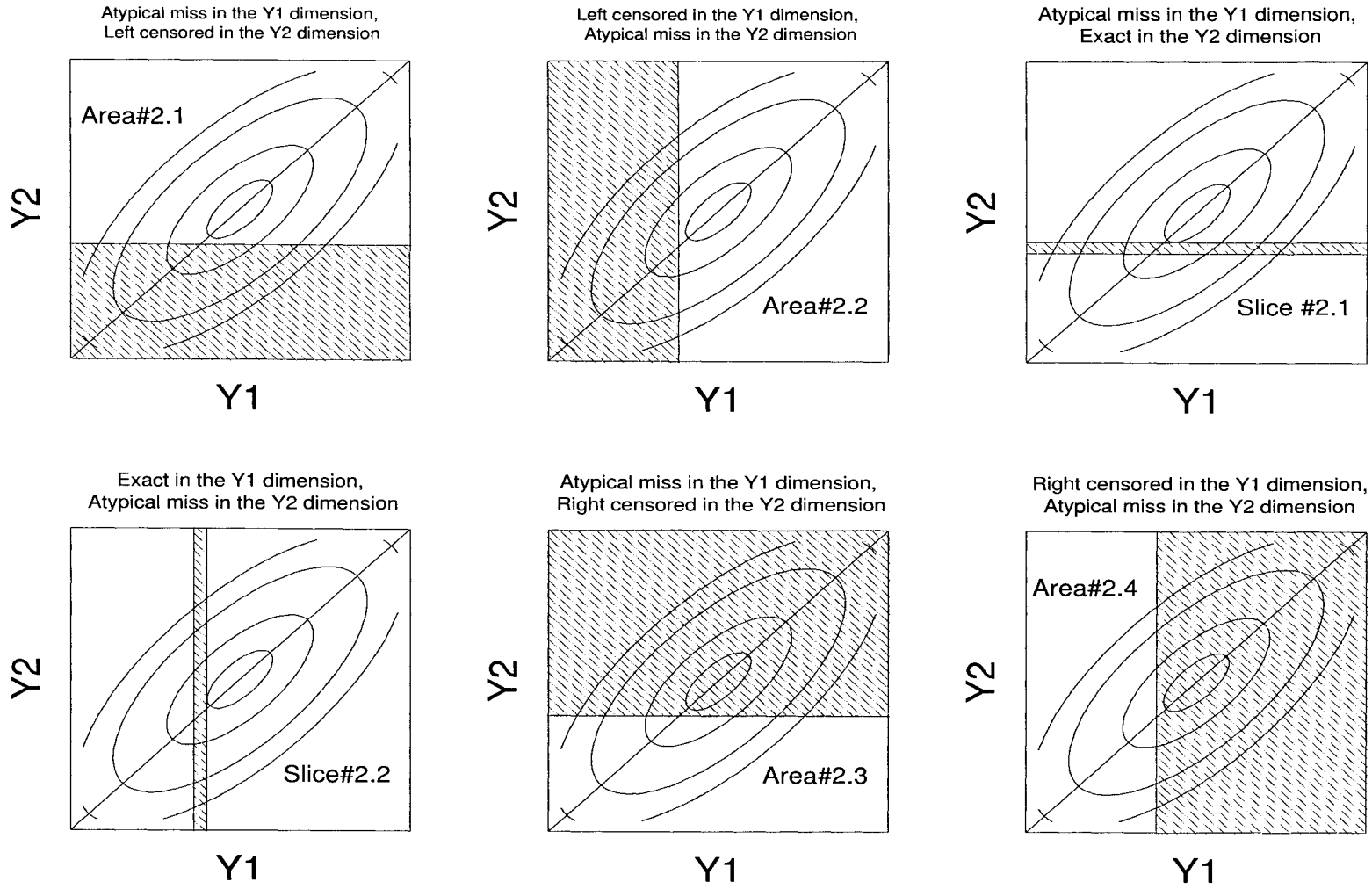

Figure 2.8 Contribution of atypical miss data to likelihood in bivariate model.

2.6.3 Accommodation Models

Using the terms in Equations (2.14), (2.15), (2.16), and (2.17), we consider the following accommodation models

- Model 1: Atypical miss probabilities $p_1 = p_2 = p_3 = 0$, (i.e., no accommodation).
- Model 2: Atypical miss probabilities p_1, p_2, and p_3 are constants that do not depend on flaw area.
- Model 3: Atypical miss probabilities p_1, p_2, and p_3 depend on flaw area through a multiple logistic regression model.

We express Model 3 as

$$\begin{aligned} \log\left(\frac{p_3}{1-p_3}\right) &= \beta_{p30} + \beta_{p31} \times x \\ \log\left(\frac{p_1}{1-p_1-p_3}\right) &= \beta_{p10} + \beta_{p11} \times x \\ \log\left(\frac{p_2}{1-p_1-p_2-p_3}\right) &= \beta_{p20} + \beta_{p21} \times x, \end{aligned} \tag{2.19}$$

where $x = \log(\text{FlawSize})$. In this regression model, having a non-negative β_{p31} makes the probability of an atypical miss in both inspections a non-decreasing function of flaw area. Model 2 is nested in Model 3. Model 3 reduces to Model 2 when $\beta_{p31} = 0$, $\beta_{p11} = 0$ and $\beta_{p21} = 0$. Model 1 is nested in Model 2. We will use likelihood ratio tests to compare these models.

2.6.4 POD for the Bivariate $\hat{a}$ versus a Model with Atypical Miss Model Accommodation Terms

For the bivariate response model with accommodation for misses, a detection can arise only when there is no atypical miss in either dimension and a response in one dimension or the other dimension exceeds its corresponding threshold. Thus,

$$
\begin{aligned}
& POD(y_1^{\mathrm{TH}}, y_2^{\mathrm{TH}}) \\
& = \Pr\left[\text{atypical misses in at most one dimension and } \quad (Y_1 > y_1^{\mathrm{TH}} \quad \text{or} \quad Y_2 > y_2^{\mathrm{TH}})\right] \\
& = (1 - p_3) \times \left[1 - F_{(Y_1,Y_2)}(y_1^{\mathrm{TH}}, y_2^{\mathrm{TH}})\right], \qquad (2.20)
\end{aligned}
$$

where y_1^{TH} and y_2^{TH} are the thresholds for two dimensions, respectively.

2.7 Analyzing the Conventional CBS Data Using the Extended $\hat{a}$ versus a Model with Atypical Miss Model Accommodation Terms

2.7.1 Results from Fitting the Models

In this section, we fit the bivariate $\hat{a}$ versus a model with Models 1, 2, and 3 using the CBS conventional 1994 data. As described in Subsection 2.6.3, Models 1, 2, and 3 are used to describe atypical miss accommodation terms. The ML estimation results are summarized in Tables 2.2 - 2.4.

Table 2.2 1994 Conventional Data ML Estimation Results under the Bivariate $\hat{a}$ versus a Model with Accommodation Model 1.

Log likelihood at maximum point: -484.3				
Parameters	*MLE*	*Std.Err*	*95% Lower*	*95% Upper*
β_{01}	0.08	1.12	-2.1	-2.27
β_{11}	0.36	0.10	0.15	0.56
β_{02}	1.72	1.13	-0.49	3.94
β_{12}	0.18	0.11	-0.03	0.39
σ_{ϵ_1}	0.86	0.15	0.57	1.16
σ_{ϵ_2}	0.95	0.17	0.63	1.28
ρ	0.65	0.11	0.43	0.86

2.7.2 Comparison of the Models

The hypothesis tests of null hypothesis about accommodation terms accounting for atypical misses:

- Test I:

Table 2.3 1994 Conventional Data ML Estimation Results under the Bivariate $\hat{a}$ versus a Model with Accommodation Model 2.

Log likelihood at maximum point: -463.5				
Parameters	*MLE*	*Std.Err*	*95% Lower*	*95% Upper*
β_{01}	3.42	0.51	2.42	5.42
β_{11}	0.09	0.05	0.00	0.18
β_{02}	4.33	0.55	3.25	5.4
β_{12}	0.002	0.05	-0.09	0.10
σ_{ϵ_1}	0.29	0.05	0.30	0.39
σ_{ϵ_2}	0.32	0.05	0.23	0.41
ρ	0.24	0.2	-0.14	0.63
$p1$	0.04	0.03	-0.01	0.09
$p2$	0.10	0.04	0.02	0.18
$p3$	0.36	0.07	0.23	0.49

Table 2.4 1994 Conventional Data ML Estimation Results under the Bivariate $\hat{a}$ versus a Model with Accommodation Model 3.

Log likelihood at maximum point: -457.5				
Parameters	*MLE*	*Std.Err*	*95% Lower*	*95% Upper*
β_{01}	3.43	0.5	2.46	4.40
β_{11}	0.09	0.05	0.001	0.18
β_{02}	4.33	0.55	3.25	5.40
β_{12}	0.002	0.05	-0.09	0.10
σ_{ϵ_1}	0.29	0.05	0.20	0.38
σ_{ϵ_2}	0.32	0.05	0.23	0.41
ρ	0.25	0.19	-0.13	0.63
β_{p10}	14.79	10.78	-6.34	35.92
β_{p11}	-1.80	1.18	-4.10	0.51
β_{p20}	-3.42	4.55	-12.34	5.50
β_{p21}	0.17	0.41	-0.64	0.97
β_{p30}	5.33	2.38	0.66	9.99
β_{p31}	-0.58	0.24	-1.05	-0.12

H_0: Model 1 is true.

H_a: Model 2 is true.

Using a likelihood ratio test, p value = 4.8e-009.

- Test II:

  H_0: Model 2 is true.

  H_a: Model 3 is true.

  Using a likelihood ratio test, p value = 0.007.

Here Models 1, 2 and 3 refer to the accommodation models used in Section 2.6.3. From likelihood ratio tests, we have strong evidence for Model 3 (probability of an atypical miss, depends on flaw area), relative to Models 1 and 2.

2.7.3 The POD Estimation for the CBS Conventional Study

For a conventional inspection, a detection occurs if the signal amplitude in the normal inspection exceeds 60% FSH or if the signal amplitude in angle inspection exceeds 60% FSH and there is no atypical miss either in the normal or the angle inspection.

Figure 2.9 shows the different components related to POD computation in Equation (2.20). In this figure, "A" represents atypical misses and "N" represents no atypical misses. And "norm" represents the normal inspection and "angle" represents the angle inspection. In this analysis, we modelled the probability of atypical misses under the four different explanations with the corresponding probabilities: atypical misses in both the normal and angle inspections (A.norm-A.angle); atypical misses only in the normal inspection (A.norm-N.angle); atypical misses only in the angle inspection (N.norm-A.angle); no atypical misses either in the normal or angle inspection (N.norm-N.angle). The multiple logistic regression model in Equation (2.19) was used to relate these probabilities to flaw area in Model 3. The probability of atypical misses in both inspections is modelled as a non-increasing function of flaw area, which accounts for the fact that POD approaches to 100% even though flaws with larger flaw sizes are easier to detect.

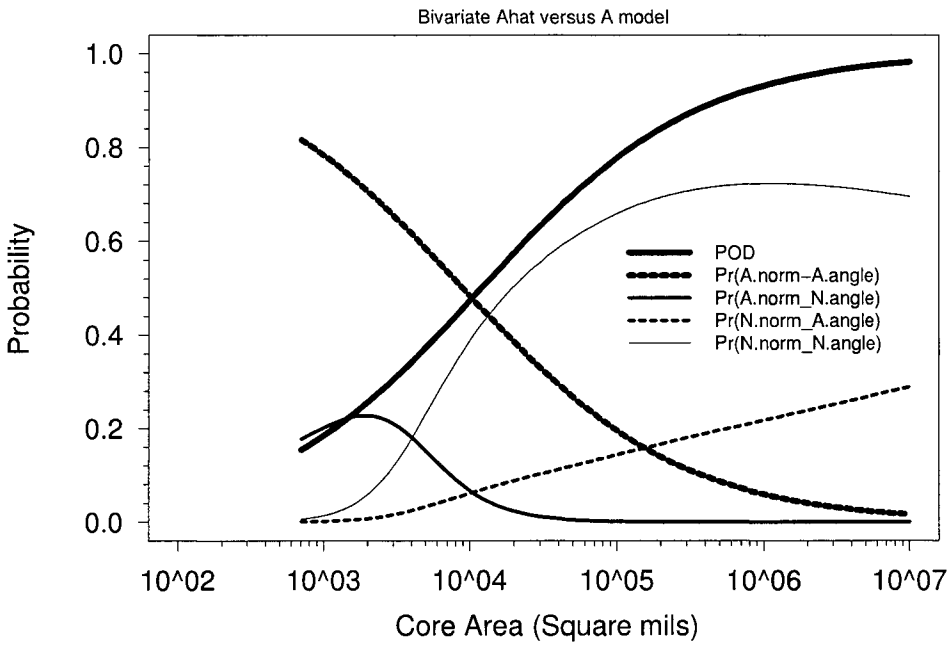

Figure 2.9 POD and probability of atypical misses for bivariate model for CBS 1994 conventional inspection data and POD when atypical misses are not eliminated.

The thin solid line labelled "N.Norm_N.angle" gives the probability of no atypical flaw misses in either dimension while the thick dash line labelled "A.Norm_A.angle" gives the probability of atypical flaw misses in both dimensions. The other two lines describe the probabilities of atypical flaw misses in only one dimension, but not the other. The solid thickest line is the POD curve computed using Equation (2.20). This curve is S-shaped. The POD approaches 1 as flaw area becomes large enough, corresponding to the fact that flaws with larger sizes are easier to detect and that the probability of an atypical flaw diminishes with increasing flaw size.

The smallest flaw area in the CBS inspection data was 600 square mils. The probability curves were estimated starting from this minimum flaw area because our analysis is based on the available data, and extrapolation to smaller flaws, without a physical basis, could be seriously misleading (Thompson, Gray, and Meeker, 2006). Indeed, a physical model suggest a change

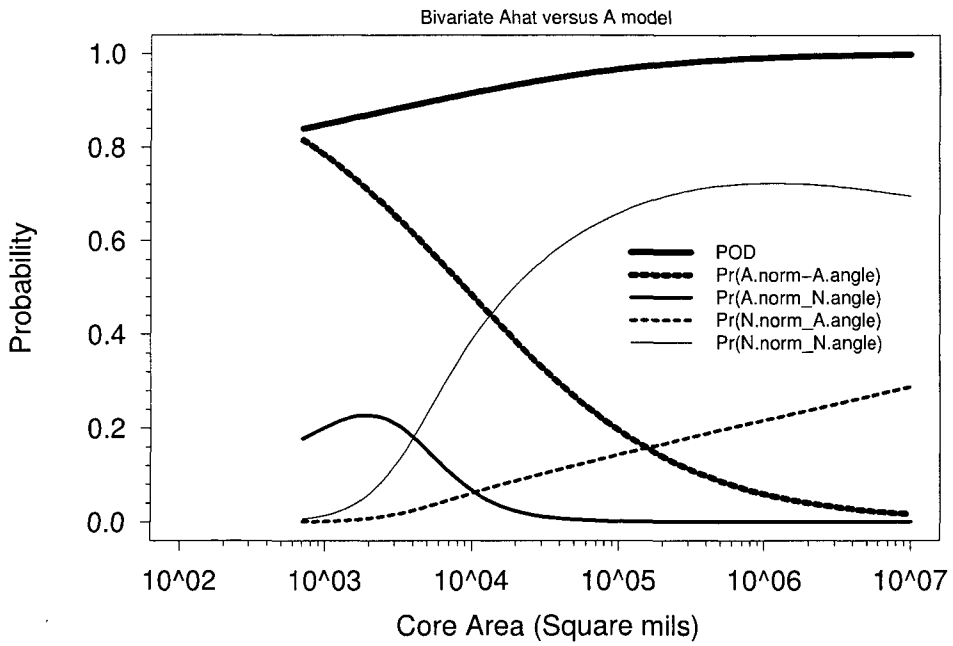

Figure 2.10 Probability of atypical miss for bivariate model for CBS 1994 conventional inspection data and POD when atypical misses are eliminated.

in the relationship for small flaws.

Figure 2.10 is similar to Figure 2.9 except that the POD curve was computed under the assumption that atypical flaw misses could be eliminated. If the atypical flaw misses could be eliminated, the POD curve would be much higher. The POD curves in Figures 2.9 and 2.10 are similar when the flaw area is very large.

2.8 Concluding Remarks and Areas for Future Work

This paper extended standard univariate $\hat{a}$ versus a model to bivariate responses, allowing for truncation and censoring. We used the extended $\hat{a}$ versus a method to analyze multizone data and provided the POD curve using a dual detection criterion. Motivated by the need to analyze conventional data with atypical misses, we also developed a model with accommodation terms

that will allow for atypical misses. The POD function was then computed. For the conventional inspection, the model includes terms to accommodate the large number of misses. In this case the POD was compared with the POD that could be achieved if the case of atypical misses could be eliminated.

The bivariate $\hat{a}$ versus a method could be extended to higher dimensions. For example, in some conventional studies, there is both amplitude and SNR information for both normal and angle inspection. The responses are in four dimensions. Analysis of these data would require multivariate regression with truncation and censoring.

2.9 Acknowledgements

This material is based upon work supported by the Federal Aviation Administration under Contract #DTFA03-98-D-00008, Delivery Order # 0034 and performed at Iowa State University's Center for NDE as part of the Engine Titanium Consortium program, through the Airworthiness Assurance Center of Excellence. We would like to express special thanks to R. Bruce Thompson and Floyd Spencer for their helpful suggestions relating to this research. We also acknowledge help comments on the CBS data and on our modelling methods, as they evaluated, from Jon Bartos, Richard Burkel, Waled Hassan, and Tim Mouzakis.

References

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley.

Berens, A. P., (1989), "NDE Reliability Data Analysis," *Metals Handbook* (9th ed., Vol. 17, *Nondestructive Evaluation and Quality Control*), Metals Park, OH: American Society for Metals 689-701.

Brasche, L., Chiou, C. P., Thompson, R. B., Smith, K., Meeker, W. Q., Margetan, F., Panetta, P., Chenail, R., Galli, F., Umbach, J., Raulerson, D., Degtyar, A., Bartos, J., Copley, D., McElligott, R., Howard, P., Bashyam, M., Contaminated Billet Study, DOT/FAA/AR-xx/xx to be published in 2005.

Burkel, R. H., Sturges, D. J., Tucker, W. T., and Gilmore, R. S. (1996), "Probability of Detection for Applied Ultrasonic Inspection," Review of Progress in Quantitative NDE, Vol. 15, edited by D. O. Thompson and D. E. Chimenti, Plenum Press, New York, NY, 1991-1998.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

MIL-HDBK-1823 (1999), *Non-Destructive Evaluation System Reliability Assessment*, Standardization Order Desk, Building 4D, 700 Roberts Avenue, Philadephia, PA 19111-5094.

Olin, B.D. and Meeker, W. Q. (1996), "Applications of Statistics in Nondestructive Evaluation," *Technometrics* Vol. 38, 95-112.

Thompson R. B., Gray T. A., and Meeker W. Q. (2006), "Use of Physics-based Models to Guide The Extrapolation of Aircraft Engine Ultrasonic POD Data to Small Flaw Sizes," to appear in Review of Progress in Quantitative NDE, Vol. 25, edited by D. O. Thompson and D. E. Chimenti, Plenum Press, New York, NY.

CHAPTER 3. A STATISTICAL MODEL TO ADJUST FOR FLAW-SIZING ERRORS IN THE ESTIMATION OF PROBABILITY OF DETECTION

A paper to be submitted to Quality Engineering

Yurong Wang and William Meeker

Department of Statistics

Iowa State University

Ames, IA 50011

Abstract

There is an important need to quantify the probability of detection (POD) in both production quality control and in-service reliability for parts that degrade over time. The standard assessment method, known as $\hat{a}$ versus a, uses a linear regression model to relate Nondestructive Evaluation (NDE) signal response to flaw or defect area. Bias in flaw sizing will, however, cause bias in estimates of POD. This paper describes two statistical models for adjusting for bias in POD estimates that is caused by flaw sizing errors. The models are fitted by using the method of maximum likelihood. We present the results of simulation studies that show how the use of our models will eliminate flaw-sizing bias. We illustrate the methods with simulated inspection data based on the collected real inspection data.

Key words: Errors in variables, Nondestructive Evaluation, Probability of Detection, Regression analysis

3.1 Introduction

3.1.1 Background

Nondestructive Evaluation (NDE) is an important tool for quality assurance. Compared with the destructive methods, it has advantages of lower cost and repeatability. NDE methods are commonly used, for example, to inspect newly manufactured safety-critical jet engine components. In addition, NDE also plays an important role in ensuring that in-service components are safe and in extending the life of some expensive system components for aircraft and power generation equipment. After a system has been in service for some time, certain critical components may have to be examined by using NDE techniques in order to ensure safety. If a component passes examination, it can continue in service. Otherwise, if a weakness (e.g., a crack) is detected, the component may have to be replaced with a new one. Especially for an expensive system, it is economically efficient to extend the system life by replacing old components with new ones when an inspection suggests need for this. In any nondestructive inspection system, however, there are random factors which can affect the performance of the system and may contribute to inspection uncertainty. These characteristics necessitate probabilistic characterization of inspection capability. Probability of detection (POD) is a commonly-used metric for this purpose.

3.1.2 Related Work and Motivation

In NDE applications, POD is often estimated by using the "$\hat{a}$ versus a" method (Berens, 1989). In NDE work, a is usually used to denote flaw area. The flaw-response signal is often translated into an estimate of flaw area and this leads to the use of "$\hat{a}$" to denote the flaw-signal response, even for applications where such a translation is not used. The basic idea behind the "$\hat{a}$ versus a" method is simple linear regression with assumptions that the $\log(\hat{a})$ has a normal distribution with mean depending on flaw area and constant standard deviation. If $\hat{a}$ from a flaw is so small that it is not discernible from noise, (i.e., less than the specified detection threshold), the flaw would be "missed". The $\hat{a}$ versus a method for POD computation assumes that the flaw sizes in the available data are known without error. However, the true

flaw area is usually not known exactly and must be inferred from some inexact method such as metallographic analysis. For example, the Jet Engine Titanium Quality Committee (JETQC) coordinates the collection of data of titanium alloy molt-related inclusions. The data consist of inspection results (e.g., signal amplitude for "conventional" inspection and both amplitude and signal-to-noise ratio for "multizone" inspection) and flaw characteristics. In the data collected under JETQC guidelines, flaw area is measured by a simple metallographic analysis, involving one or two cuts through the flaw and which generally causes the measured flaw area to be smaller than the true flaw area. Results in the classical statistical literature indicate that such errors-in-variables (EV) will bias the estimated regression coefficients. Thus, the presence of measurement errors in flaw sizing will also affect the linear regression in POD computations.

Fuller (1987) introduced the classical measurement error model, investigated the effects of measurement error on the ordinary least squares estimators, and provided proper methods to do estimation. Carroll, Ruppert and Stefanski (1995) extended these ideas to cover the nonlinear regression measurement error model and provided more general approaches to solve the measurement error problem. The effect of EV issues has not been studied carefully in NDE applications. Our objective is to adapt measurement error model methods to NDE applications and to extend the standard $\hat{a}$ versus a method to adjust for measurement error.

3.1.3 Overview

The remainder of this paper is organized as follows. Section 2 describes the classical measurement error model. Section 3 describes the flaw measurement process used in NDE applications that require flaw sizing. Section 4 explains the Burkel measurement error model, which is the first model we used to study and correct measurement error in our NDE applications. This method allows for truncation and censoring in the response. Section 5 presents an alternative measurement error model, which we call the geometrical measurement error model. Section 6 provides conclusion and discussion about future research.

3.2 The Classical Measurement Error Model

3.2.1 General Errors-in-Variables (GEV) Model

The $\hat{a}$ versus a method is the standard method to estimate POD for NDE inspections when quantitative signal strengths can be recorded. This method is based on simple linear regression. The model can be written as

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{3.1}$$

Here Y is the (possibly transformed) observed response and ϵ is unobservable residual, assumed to be *i.i.d.* with mean 0 and standard deviation σ_ϵ. X is the observable regressor (typically log flaw area in NDE applications).

If there is measurement error in the explanatory variable X, X is not observable. However, W, the measurement of X, can be observed. This motivates the classical measurement error model, which can be written as an extension of Equation (3.1)

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon \\ W &= X + U \end{aligned}$$

where,

- Y is the response.
- ϵ is the error in the response, which has normal distribution with mean 0 and standard deviation σ_ϵ.
- X is the true value of the explanatory variable.
- W is the measurement of X.
- U is the measurement error, which has normal distribution and is independent of ϵ.

If we assume that X is random and independent of measurement error U, it can be shown that the estimated regression slope is biased toward zero (page 5 of Fuller, 1987). In NDE

applications, the measurement error U has a skewed distribution, instead of the normal distribution assumed by Fuller. We used general EV method to handle the skewed distribution. Detailed discussion will be presented in Section 3.4 and Section 3.5.

3.2.2 General Maximum Likelihood Approach for Estimation

In POD estimation using the $\hat{a}$ versus a method, the regression coefficients (β_0, β_1) and standard deviation σ_ϵ are estimated using the method of maximum likelihood (ML) because the ML method can be used in more complicated (but important) situations involving censoring and truncation and because ML estimators have desirable statistical properties such as

- Asymptotic optimality properties: In large samples, the distribution of ML estimators is approximately a multivariate normal distribution and ML estimators converge to true parameter values.
- Invariance property: Under standard regularity conditions, ML estimation of a parameter function $g(\theta)$ is simply $g(\hat{\theta})$.

The likelihood function is central to making inferences when censoring and truncation mechanisms are active because it describes the probability for given observed data as a function of the unknown model parameters and because it produces a theoretical basis for structural inference.

In addition to the assumption of a linear relationship between the response to the exploratory variable X, the standard $\hat{a}$ versus a method assumes that the true value of exploratory variable is observed and fixed. In the presence of measurement error, however, the true value of the exploratory variable X is not observable. Instead, the response and the measured exploratory variable (both of which are random) are observed. To get the proper likelihood function, a straightforward approach is to use the joint probability distribution of the observables: Y and W. In the measurement error model, we only assume that:

- The true value X has certain distribution.
- The measurement error U has an arbitrary probability distribution.
- The measured value W has certain distribution implied by the distributions of X and U.

- The response Y has a distribution with the mean that is linearly related to X.

Based on this regression measurement error model, the conditional distributions of Y given X, W given X and the marginal distribution of X, one can obtain the joint distribution

$$\begin{aligned} f_{(Y,W)}(y,w) &= \int_x f_{(Y,W)|X}(y,w) f_X(x) dx \\ &= \int_x f_{Y|(W,X)}(y,w) f_{W|X}(w) f_X(x) dx \\ &= \int_x f_{Y|X}(y|x,\beta_0,\beta_1,\sigma) f_{W|X}(w|x,\sigma_U) f_X(x|\mu_X,\sigma_X) dx. \end{aligned} \tag{3.2}$$

In addition to the measurement error, we also need to deal with censoring and truncation in the response when we write the likelihood function.

3.3 The Flaw Area Measurement Process

In data collection efforts, such as those coordinated by JETEQC, the approximate flaw area is obtained by performing metallographic analysis on flaws. Flaws are located using ultrasonic testing. Hard alpha flaws tend to be cigar-shaped or ellipsoid-shaped with the long axis aligned with the axis of the billets. The length of flaws can be obtained, to a good approximation, from the ultrasonic C-Scan image. The billet is cut several times in the region where the flaw was detected to look for the largest diameter of a flaw. Assuming that a cut happens to be made at the largest cross-sectional area of a flaw, the measured flaw area will be equal to the true flaw area. Otherwise, measured flaw area will be less than true flaw area.

3.4 The Burkel Measurement Error Model

3.4.1 The Burkel Measurement Error Model

The model described in this section was suggested to us in private communication by Dr. Richard Burkel, from General Electric Transportation. Let D_{max} denote the maximum diameter of a flaw and let D denote the measured diameter of a flaw at cutting position. Let L be the

length of a flaw, then

$$\begin{aligned} \text{MeasuredFlawArea} &= \pi \times D \times L \\ \text{TrueFlawArea} &= \pi \times D_{max} \times L \end{aligned}$$

Because $D \leq D_{max}$,

$$\begin{aligned} \log(\text{MeasuredFlawArea}) &= \log(\pi \times D \times L) \\ &= \log\left(\pi \times D_{max} \times L \times \frac{D}{D_{max}}\right) \\ &= \log(\text{TrueFlawArea}) + \log\left(\frac{\text{MeasuredFlawArea}}{\text{TrueFlawArea}}\right), \end{aligned}$$

which we write as $W = X + U$. We assume that U is bounded between δ_1 and δ_2 (logarithms of the minimum and maximum ratio of measured flaw area to true flaw area, respectively) with a truncated normal distribution having a cumulative distribution function (CDF)

$$F_U(u, \delta_1, \delta_2) = \begin{cases} \frac{\Phi\left(\frac{u}{\sigma_U}\right) - \Phi\left(\frac{\delta_1}{\sigma_U}\right)}{\Phi\left(\frac{\delta_2}{\sigma_U}\right) - \Phi\left(\frac{\delta_1}{\sigma_U}\right)} & u > \delta_1 \quad \text{and} \quad u < \delta_2 \\ 0 & \text{otherwise.} \end{cases}$$

Here Φ is the CDF of the standard normal distribution.

Figure 3.1 shows simulated data from the Burkel measurement error model using the following parameters: $\beta_0 = 6.3$, $\beta_1 = 0.5$, $\sigma_\epsilon = 0.44$, $\mu_X = 7.5$, $\sigma_X = 0.85$, MinFlawRatio = 0.1, MaxFlawRatio = 1 ($\delta_1 = \log(\text{MinFlawRatio})$ and $\delta_2 = \log(\text{MaxFlawRatio})$), $\sigma_U = 2.0$. The selected values for the parameters β_0, β_1, σ_ϵ, μ_X and σ_X were estimated from one of the proprietary NDE data sets that we have been asked to analyze. These parameter values will also be used in other parts of this paper. The simulated response is "Effective Flat Bottom Hole (EFBH) Area" (in square mils). EFBH is defined as

$$\text{EFBH} = \frac{\%FSH}{80} \times \left(\frac{\text{Cal}}{64}\right)^2 \times \frac{\pi}{4} \times 10^6$$

. EFBH gives the size of the flat bottom hole that would result in the same % full screen height (%FSH), assuming calibration was done to a #Cal flat bottom hole (FBH). EFBH is

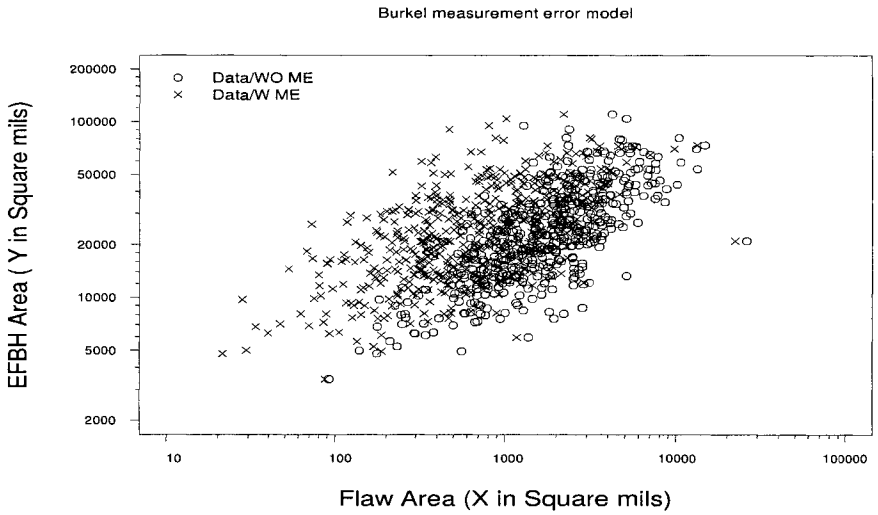

Figure 3.1 Simulated data with/without measurement error.

used because some data sets have %FSH as a response but a mix of calibration values were used (e.g., #2 and #3 FBHs were used to calibrate conventional inspections in the NDE applications that motivated this research) and EFBH serves as a common measure of signal strength.

In Figure 3.1, circles represent simulated data with true flaw area while the crosses represent simulated data with measured flaw area (i.e., flaw area with measurement error). With $\sigma_U = 2.0$, the measurement errors in flaw sizing cause data to shift importantly to the left.

To show the repeatability of the deviation shown in Figure 3.1, Figure 3.2 summarizes 50 data-generation/estimation simulations with a sample size of 500 showing the Ordinary Least Square (OLS) mean line for each trial (we can use OLS in this simulation because there is no censoring). The longer darker line is the mean line for the true distribution. This simulation illustrates the potential strength of the bias caused by using a standard regression model when there are substantial measurement errors in flaw sizing. The bias in regression coefficients will be propagated to the POD computation, leading to biased POD estimates.

As in the classical EV model, data without repeated measures on the explanatory variable are not sufficient to estimate all of the parameters in the measurement error model. Instead we assume that the measurement error distribution is available from other sources such as

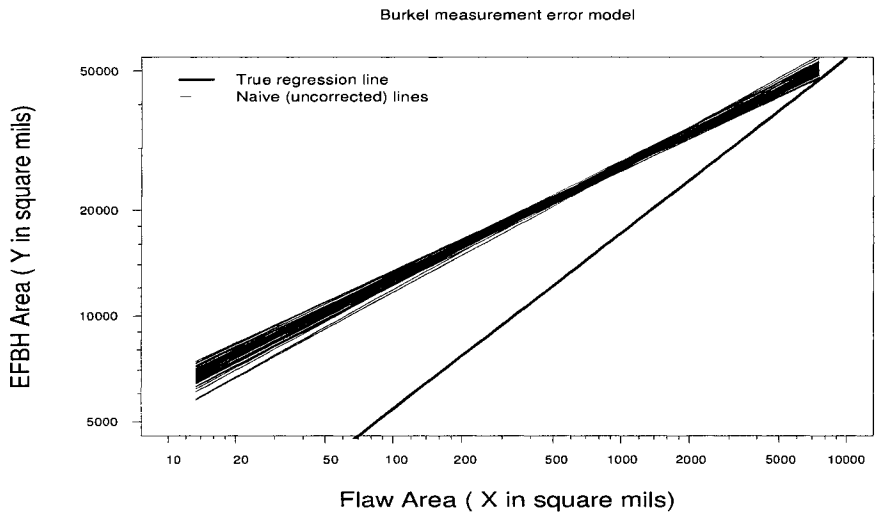

Figure 3.2 Effect of measurement error on the regression.

engineering judgement or analytical analysis.

3.4.2 Maximum Likelihood Estimation for the Burkel Model

3.4.2.1 Joint distribution

Following the general approach in Section 3.2.2 and using the Burkel measurement error model in Section 3.4.1, the joint probability density function of response Y and measured flaw area W can be derived as:

$$
\begin{aligned}
f_{(Y,W)}(y,w) = \int_{W-\delta_2}^{W-\delta_1} & \times \frac{1.0}{\sigma_\epsilon\sqrt{2\pi}} \exp\left(\frac{(y-\beta_0-\beta_1 x)^2}{-2\sigma_\epsilon^2}\right) \\
& \times \frac{1.0}{\sigma_U\sqrt{2\pi}} \frac{1}{c} \exp\left(\frac{(W-x)^2}{-2\sigma_U^2}\right) \\
& \times \frac{1.0}{\sigma_X\sqrt{2\pi}} \exp\left(\frac{(x-\mu_X)^2}{-2\sigma_X^2}\right) dx, \qquad (3.3)
\end{aligned}
$$

where, β_0, β_1 and σ_ϵ are the parameters used in the standard $\hat{a}$ versus a model. The constant c is

$$c = \Phi\left(\frac{\delta_2}{\sigma_U}\right) - \Phi\left(\frac{\delta_1}{\sigma_U}\right).$$

The other parameters were defined above.

3.4.2.2 Likelihood function

The likelihood function is proportional to the probability of the data. Taking logarithms simplifies numerical computations. The log likelihood function, for fixed values of σ_U, δ_1 and δ_2 and the data, can be written as the sum of the contributions for each of the n independent observations in the data set:

$$\mathcal{L}(\beta_0, \beta_1, \sigma_\epsilon, \mu_X, \sigma_X; \sigma_U, \delta_1, \delta_2, Y, W) = \sum_{i=1}^{n} \mathcal{L}_i. \tag{3.4}$$

Let $H_Y(y)$ be the marginal cumulative probability distribution of response Y, and let $H_{Y|W}$ be the conditional probability cumulative distribution of response Y, given measured exploratory variable W. Let g_W be the marginal probability density function of W.

In some applications, the experimental ultrasonic testing (UT) data will be either truncated or censored or both. Left truncation is used to account for the possibility of field flaw misses and right censoring arises from saturated signals (i.e., signals that are so large that only a lower bound on signal strength is recorded). The likelihood contributions for the different types of observations, assuming left truncation at a level y^{TL} and right censoring at a level y^{CR}, are given in Equations (3.5) and (3.6), respectively.

- The likelihood contribution for an exact observation (no censoring) is

$$\mathcal{L}_i = \log\left(\frac{f_{(Y,W)}(y,w)}{1 - \int_{-\infty}^{y^{\mathrm{TL}}}\int_{-\infty}^{\infty} f_{(Y,W)}(y,w)dydw}\right). \tag{3.5}$$

- The log likelihood contribution for a right censored (response saturation) observation is

$$\begin{aligned}\mathcal{L}_i &= \log\left(\frac{\int_{y^{\mathrm{CR}}}^{\infty} f_{(Y,W)}(y,w)dy}{1 - \int_{-\infty}^{y^{\mathrm{TL}}}\int_{-\infty}^{\infty} f_{(Y,W)}(y,w)dydw}\right) \\ &= \log\left(\frac{g_W(W=w) \times (1 - H_{Y|W}(y^{\mathrm{CR}}|W=w))}{1 - H_Y(y^{\mathrm{TL}})}\right),\end{aligned} \tag{3.6}$$

where $f_{(Y,W)}(y, w)$ is the probability density function of Y and W in Equation (3.3). The terms in the denominator of Equations (3.5) and (3.6) account for the possibility of unknown misses (left truncation) in field inspections. ML estimates of the unknown distribution parameters are obtained by finding those values of the parameters that maximize Equation (3.4).

3.4.3 Simulation of the Burkel Model GEV Method

In Section 3.2, we presented the general measurement error model and demonstrated that measurement error can cause the estimated regression line to deviate from the true regression line if the magnitude of the errors is large. To accurately estimate POD using the data with measurement error, we developed a GEV method. For this method, we assume a particular measurement error model for the flaw area and use ML method to estimate the regression coefficients required by POD computation. We used the same parameter values (β_0, β_1, σ_ϵ, μ_X and σ_X) as used in the simulation for Figure 3.1. Once the simulated data were generated, we applied the GEV method to these data to show how the GEV method works to correct the bias.

We use simulation to show how the changes in the standard deviation and lower truncation level, respectively, affect the magnitude of estimation bias. The effect of standard deviation was studied first.

Similar to Figure 3.1, Figure 3.3 shows that a large standard deviation will cause a large shift of the simulated data with measurement error away from the data without error. Figure 3.4 shows three different types of lines: the true regression line, the naive model estimate lines, and GEV model estimate lines. The simulation was based on the true regression line. The naive model estimate lines are fitted by OLS using the simulated data with measurement errors. The GEV model estimate lines are fitted by using the same data sets that were used in the naive model fitting. When the measurement error is very small, both the naive lines and the GEV lines are aligned very close to the true regression line. When the measurement error becomes large, the naive lines deviate from the true regression line and the bias is large. The GEV lines are, however, still centered with the true regression line even though the measurement error is large. This indicates the GEV ML estimations have little bias. The increased sampling

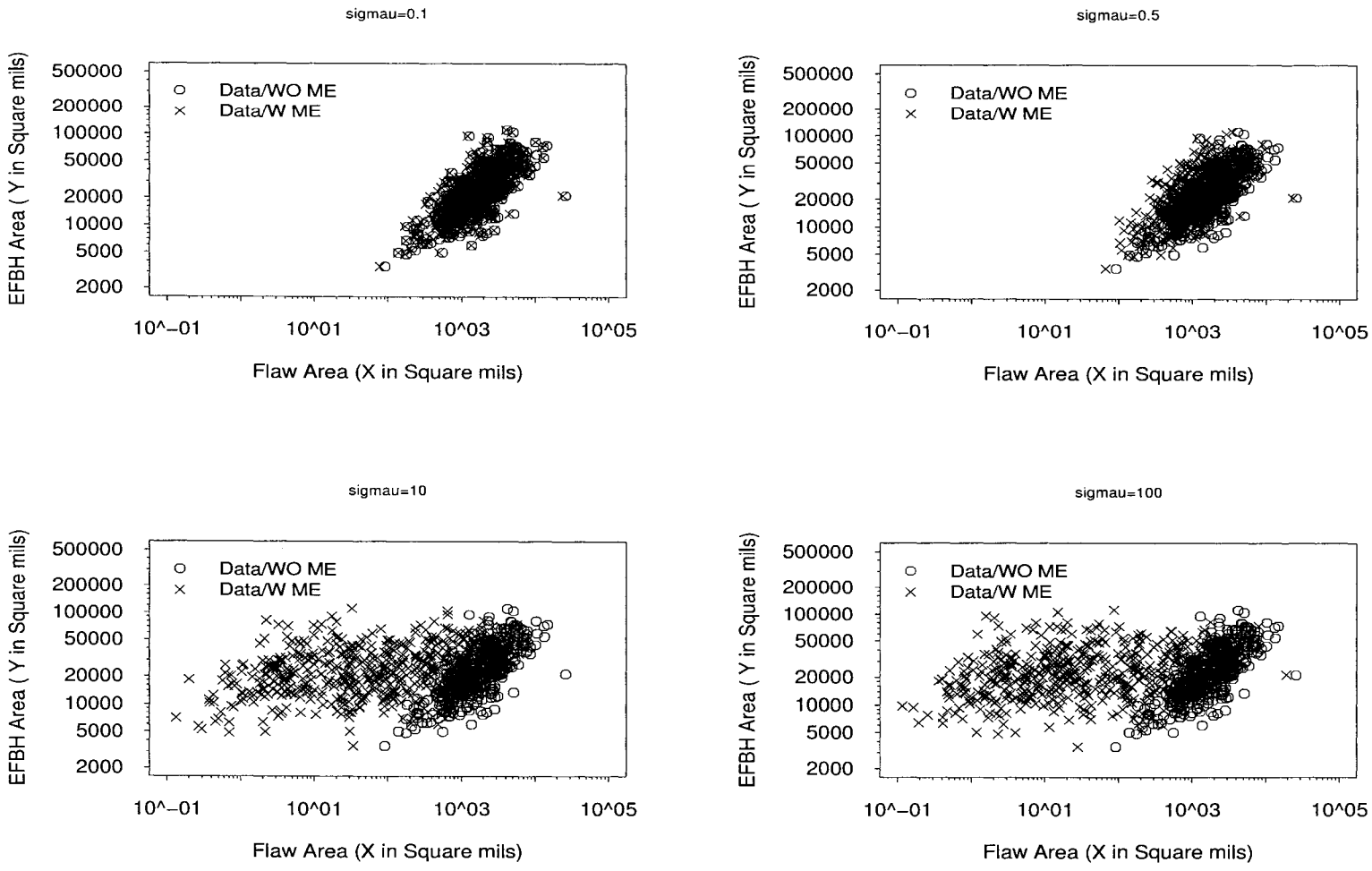

Figure 3.3 Effect of the standard deviation on measurement error using $\beta_0 = 6.3$, $\beta_1 = 0.5$, $\sigma_\epsilon = 0.44$, MinFlawRatio $= 0.0005$.

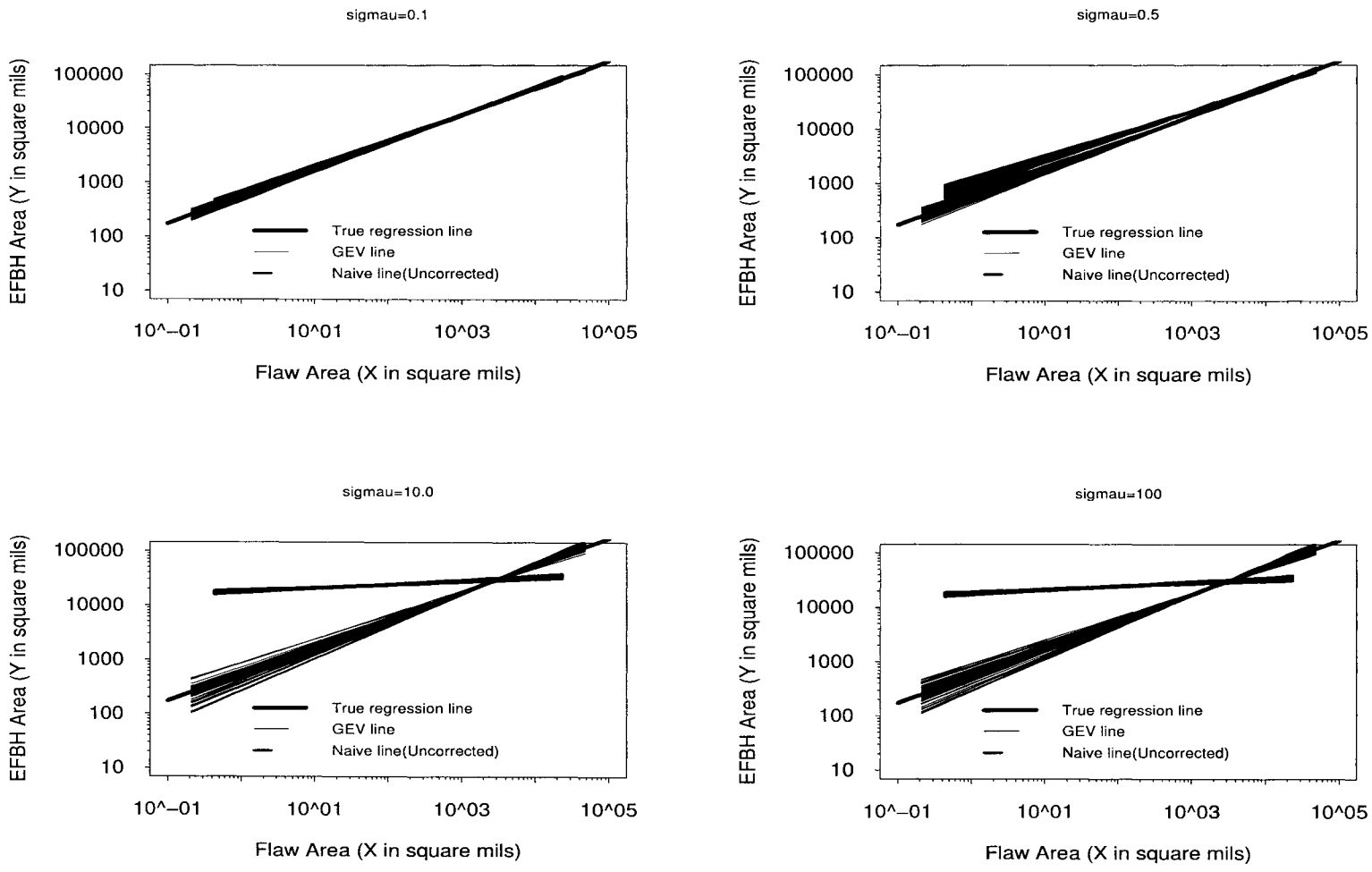

Figure 3.4 Comparison of the naive and GEV methods for different values of the standard deviation of measurement error.

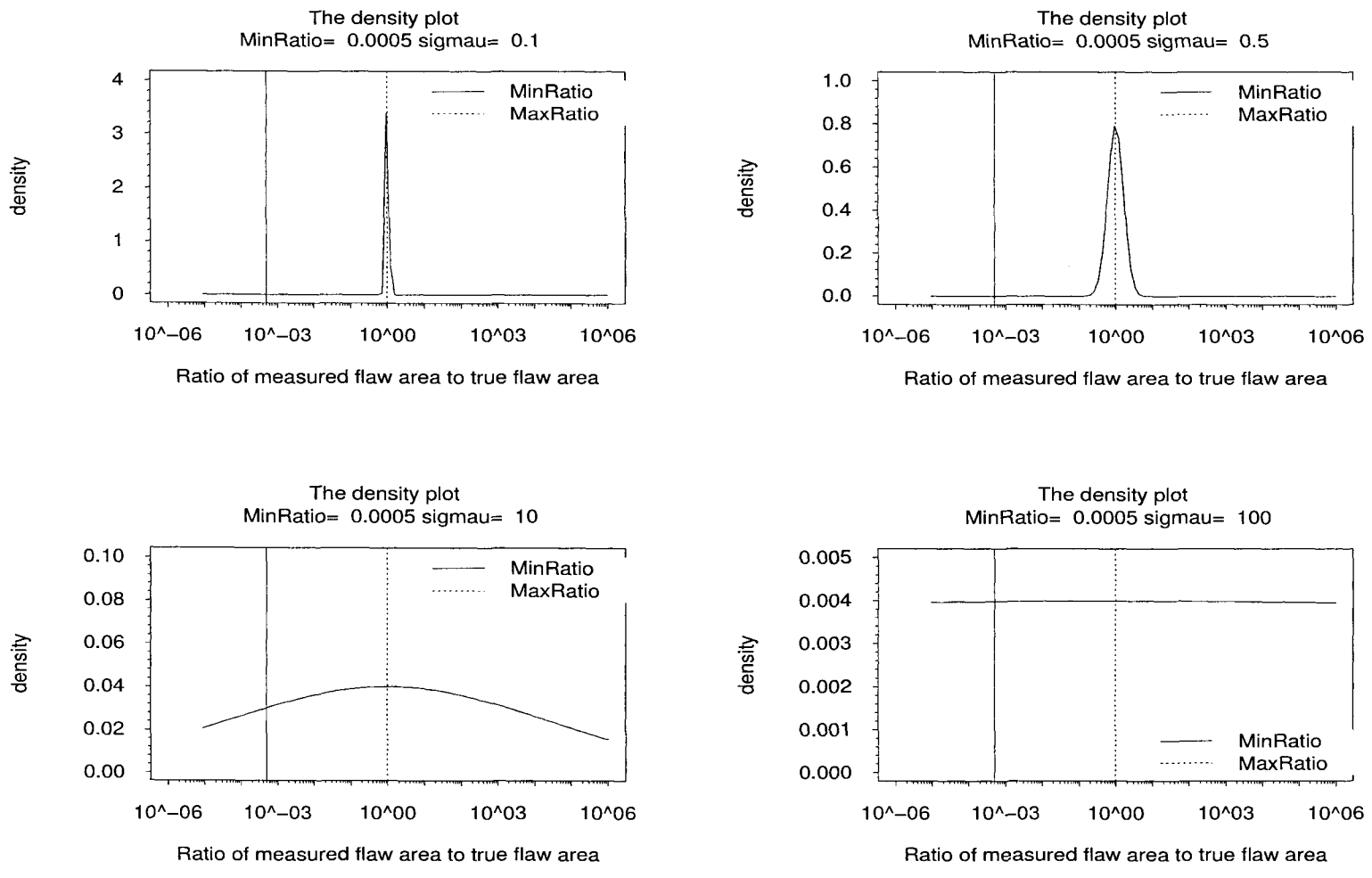

Figure 3.5 The density plot of the flaw Area ratio for different standard deviation of measurement error.

variability in the GEV ML estimate lines is caused by additional variability in the flaw area measurement.

To understand how the data shift increases with the increasing σ_U, as shown in Figure 3.3, we did further study. Figure 3.5 shows the distribution of flaw area ratio (i.e, the ratio of measured flaw area to true flaw area) as a function of σ_U. The increase in σ_U causes the truncated lognormal density function to spread out so that the distribution of measurement errors approaches to a uniform-like distribution over the truncated range. This makes the flaw area ratio have a higher probability of being further away from 1 (i.e., more measurement error).

The effect of the minimum flaw area ratio (i.e., $\exp(\delta_1)$) on the measurement error distribution was illustrated in Figure 3.6. The smaller values of $\exp(\delta_1)$ lead to larger measurement error bias. Similar to Figure 3.5, Figure 3.7 shows the reason behind the relationship between the measurement error distribution and the minimum flaw area ratio. When the minimum flaw area ratio becomes smaller, the lognormal density distribution function for flaw area ratio (i.e, ratio of measured flaw area to true flaw area) spreads out so that the measurement error distribution becomes wider. The wider spread forces the flaw area ratio to be further away from 1 (i.e., more measurement error) for more of the data.

The GEV method is capable of correcting the measurement error over a wide range of different minimum flaw ratios. Examples of simulations to illustrate this are shown in Figure 3.8.

3.4.4 Application to Simulated Inspection Data

The real data sets that motivated this work and that we have analyzed with the developed method are proprietary. Following closely the data structure of these real data sets and using the Burkel measurement error model, we simulated a UT inspection data set. We use these simulated data to illustrate the results that we saw in the analysis of the actual data. The simulated data include flaw response amplitude in units of percent of FSH and measured flaw area in units of square mils (1/1000 of an inch). We simulated flaw amplitudes (i.e., %FSH) instead of EFBH areas because the real data sets are reported in these units and all used the same calibration specification. We assume that flaw area has a lognormal distribution with $\mu_X = 8.63$ and $\sigma_X = 1.65$. The values of the parameters used to simulate response

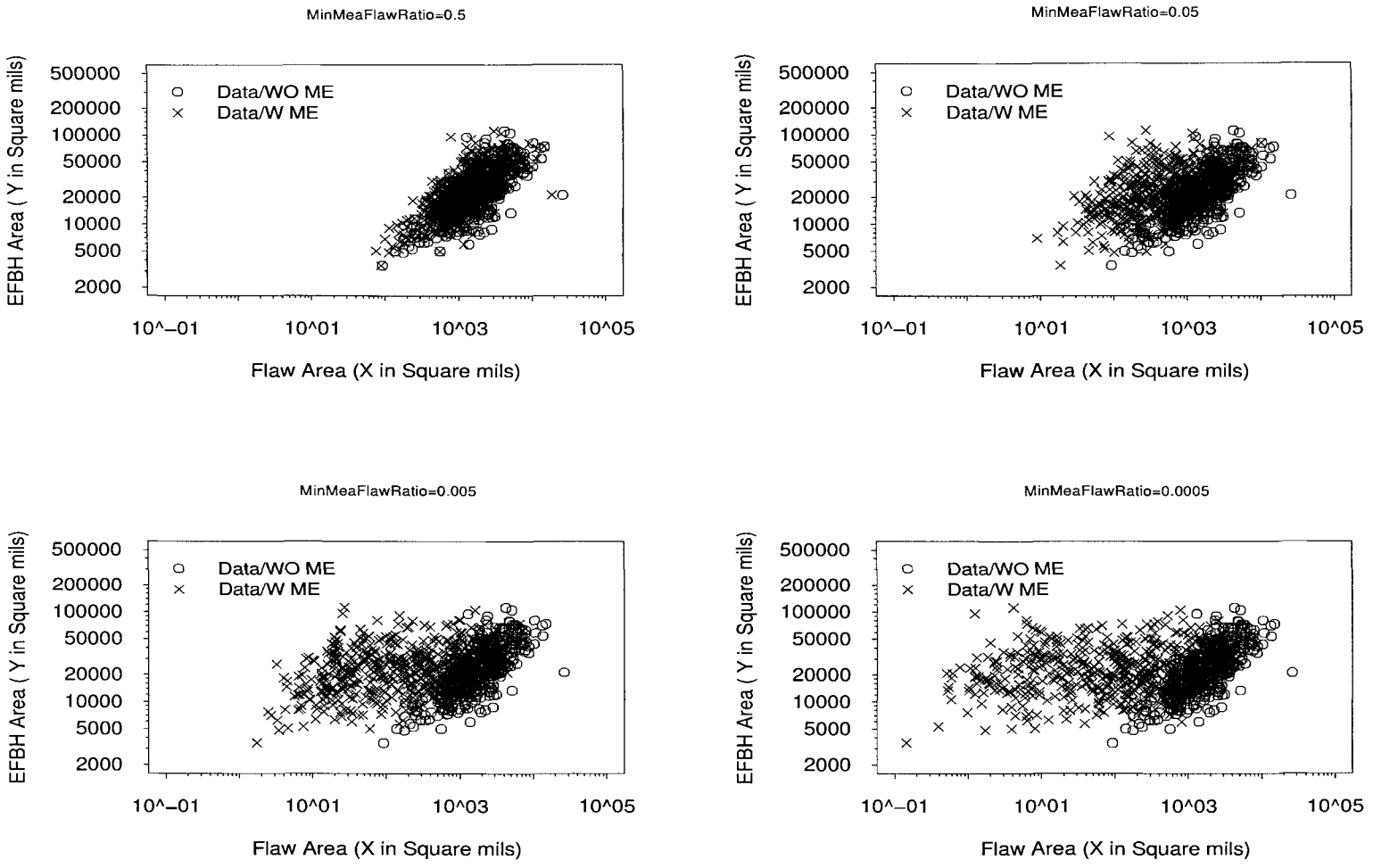

Figure 3.6 Effect of the minimum flaw area ratio on measurement error.

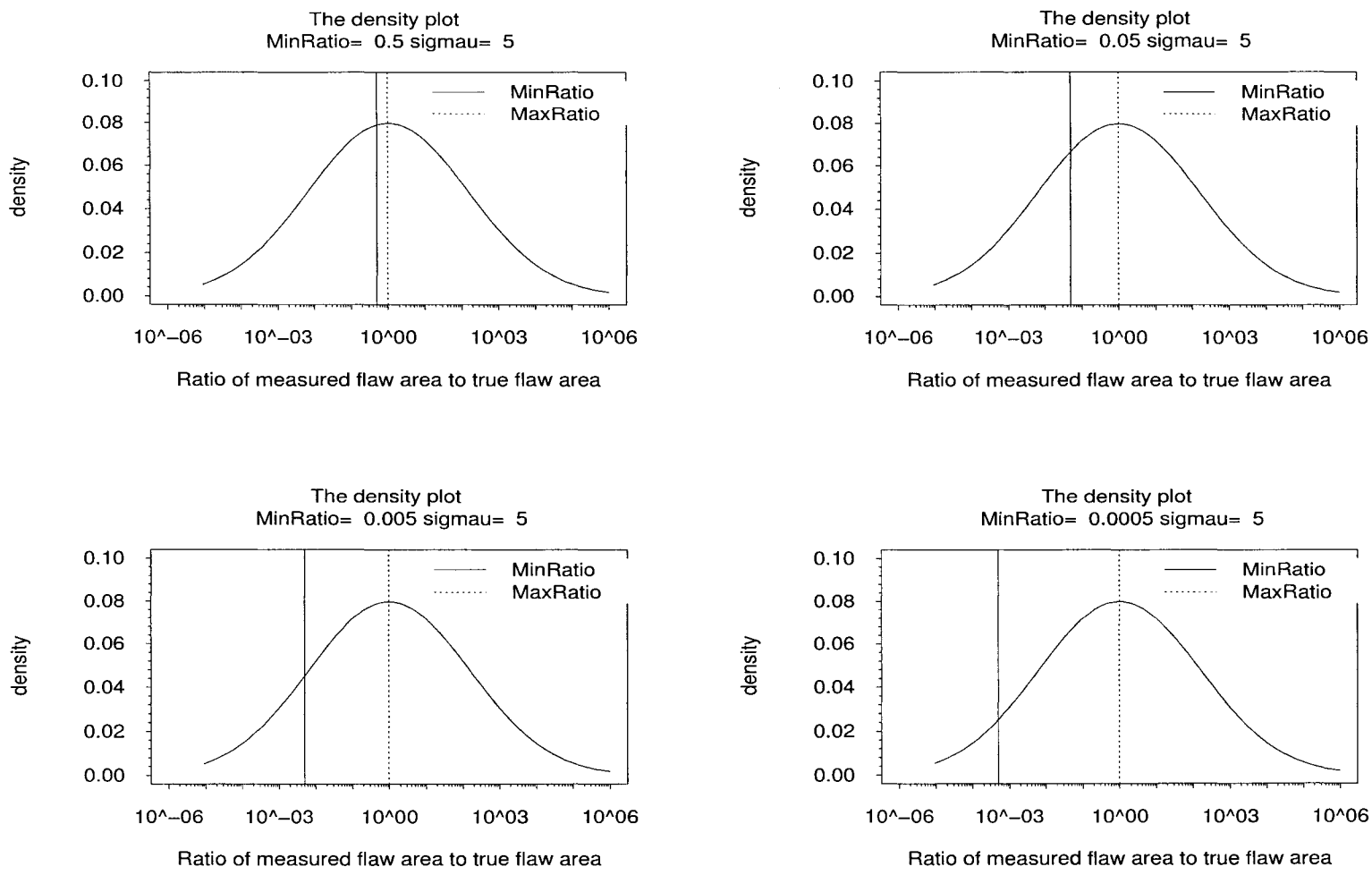

Figure 3.7 The density plot of the flaw area ratio at different minimum flaw area ratio.

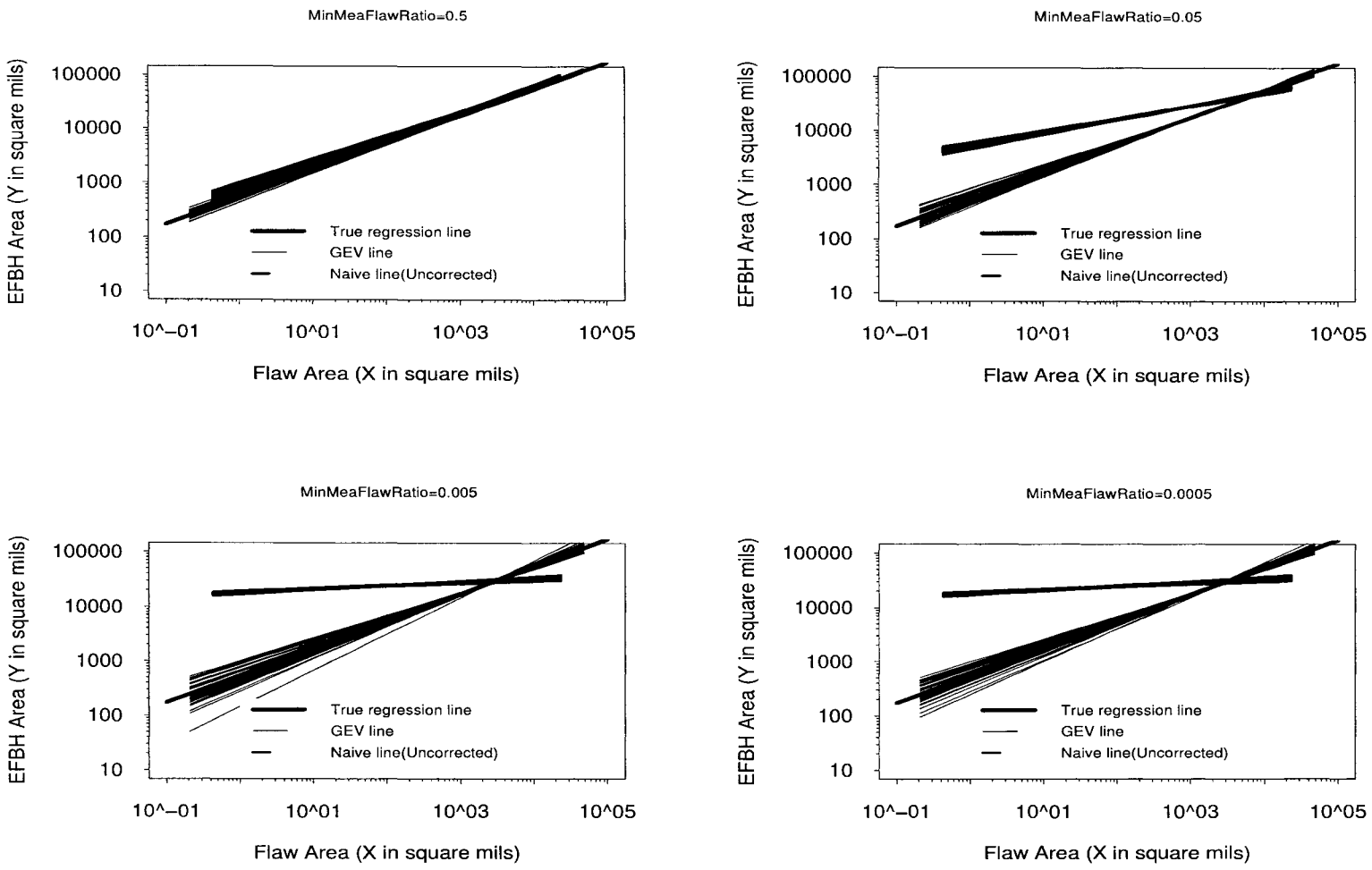

Figure 3.8 Comparison of the naive and the GEV methods for different values of minimum flaw ratio $\exp(\delta_1)$.

under the $\hat{a}$ versus a model are: $\beta_0 = 1.45$, $\beta_1 = 0.3$, and $\sigma_\epsilon = 0.4$. These above parameter values are estimated from the real proprietary data. Richard Burkel (private communication) suggested that the minimum flaw area ratio should be 0.5. In addition, he suggested that one standard deviation of the untruncated normal distribution for measurement error should correspond approximately to a 6dB change in the responses. Using this information, we did a transformation under assumption of $\beta_1 = 0.3$ (a typical value for UT inspection), and derived $\sigma_U = 2.31$. Also following the approach used by Burkle, Sturges, Tucker and Gilmore (1996), the truncation level in the simulation is obtained by adding 10% screen height to the observed noise level.

Figure 3.9 shows the simulated UT inspection data with measurement error using these parameters, with saturation for signals greater than 100% FSH. The estimated regression lines based on the naive (uncorrected) regression and the GEV method are also plotted with the simulated data. The naive line deviates importantly from the GEV line. This result is consistent with the simulation results presented in Section 3.4.3.

Figure 3.10 shows two POD curves for the simulated inspection study. The thin curve is estimated using the naive (uncorrected) method while the thicker curve is estimated using GEV method assuming the Burkel measurement error model. The POD curve estimated by the GEV method shifts to right and is less steep, when compared to the POD estimated by naive method. The comparison suggests that the naive method can give seriously inaccurate results when there is substantial measurement error.

3.5 The Geometrical Measurement Error Model

3.5.1 The Geometrical Measurement Error Model

As mentioned in Section 3.4.1, due to cost constraints, the "true" flaw area required in the standard regression model employed by the $\hat{a}$ versus a method is usually not available for flaws found in actual inspections. In the real inspection data that we studied, the flaws had cigar-shape or ellipsoid-shape with the long axis aligned with the axis of the billets. Flaw area is estimated by cutting the billet (and thus the flaw) perpendicular to its long axis. Because

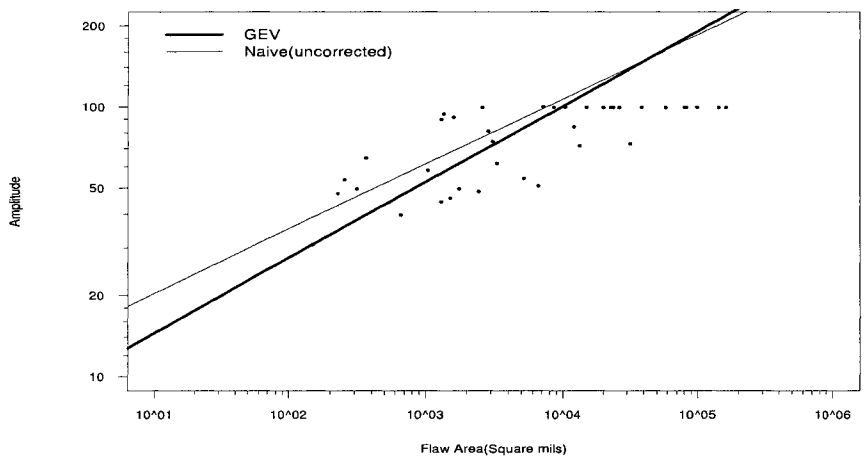

Figure 3.9 Plot showing simulated UT inspection data with measurement error and both naive and GEV regression lines.

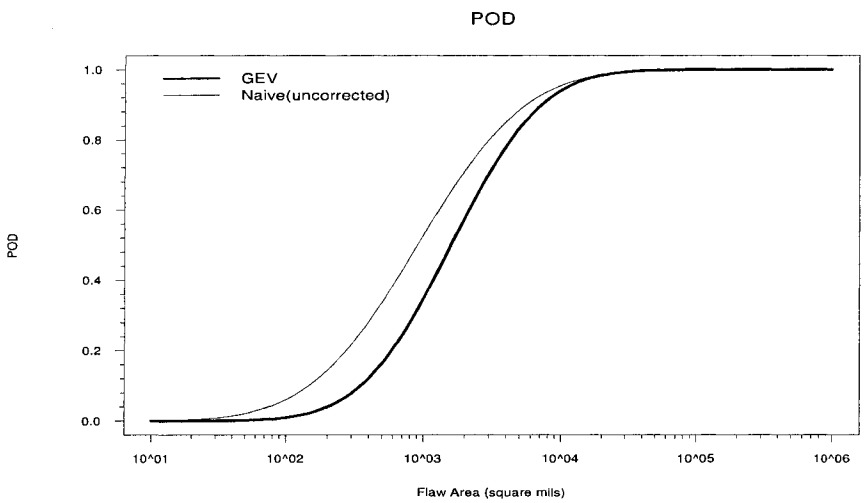

Figure 3.10 POD plots for the conventional inspection using simulated data.

we know that the major source of flaw area error is the cutting location, here we develop an alternative measurement model from a geometrical view point. The assumptions used in this model are:

- The shape of a flaw can be described by a three dimensional ellipsoid whose two short axes have the same length, denoted by M. The length of the long axis is denoted by K.
- The cutting plane is normal to the long axis of the ellipsoid.
- The dimension of the long axis of the flaw is estimated from an ultrasonic C-scan image and the measurement error in this dimension is negligible.
- C, the distance between the cutting plane and the flaw center, has a truncated normal distribution. $C \sim TN(0, {\sigma_C}^2, -K, K)$.
- P, the ratio of K to M also follows a truncated normal distribution. That is, $P = K/M \sim TN(\mu_P, {\sigma_P}^2, 1, \infty)$.
- C and P are independent.

The surface of an ellipsoid is given by

$$\frac{x^2}{M^2} + \frac{y^2}{M^2} + \frac{z^2}{(P \times M)^2} = 1.$$

In terms of C, P and the surface equation of a flaw, the measured flaw area can be written as:

$$\text{MeasuredFlawSize} = \pi P M^2 \sqrt{\left(1 - \frac{C^2}{K^2}\right)},$$

where

$$\begin{aligned} W &= \log(\text{MeasuredFlawSize}) \\ &= \log\left(\pi P M^2\right) + \frac{1}{2}\log\left(1 - \frac{C^2}{K^2}\right) \\ &= \log\left(\text{TrueFlawSize}\right) + \frac{1}{2}\log\left(1 - \frac{C^2 \times \pi}{\text{TrueFlawSize} \times P}\right). \end{aligned}$$

We can write $W = X + U$. Thus,

$$\begin{aligned} U &= \frac{1}{2}\log\left(1 - \frac{C^2}{K^2}\right) \\ &= \frac{1}{2}\log\left(1 - \frac{C^2 \times \pi}{\text{TrueFlawArea} \times P}\right) \\ &= \frac{1}{2}\log\left(1 - \frac{C^2 \times \pi}{\exp(X) \times P}\right). \end{aligned} \tag{3.7}$$

Because both C and P have a truncated normal distribution and they are independent of each other, their joint PDF can be written as Equation (3.8) for a given $X = x$. Again, the length measured from ultrasonic C-scan image is assumed to be accurate.

$$\begin{aligned} f_{(C,P)}(c,p) = {} & \frac{1}{\sigma_X \times \sqrt{2\pi}} \exp\left[\frac{-c^2}{2\sigma_C^2}\right] \\ & \times \frac{1}{\sigma_P \times p\sqrt{2\pi}} \exp\left[\frac{-(\log(p) - \mu_P)^2}{2\sigma_P^2}\right] \\ & \times \frac{1}{2\Phi\left(\frac{\sqrt{\frac{\exp(x)\times p}{\pi}}}{\sigma_C}\right) - 1} \times \frac{1}{\Phi(\frac{\mu_P}{\sigma_P})}. \end{aligned} \tag{3.8}$$

Using the probability integral transformation technique and the derived joint distribution of C and P, the PDF of measured error U is

$$\begin{aligned} f_U(u) &= \frac{d}{du}\text{Prob}(U <= u) \\ &= \int_1^{\infty} \frac{d}{du}\left\{ \frac{2\left[\Phi\left(\frac{\sqrt{\frac{\exp(x)p}{\pi}}}{\sigma_C}\right) - \Phi\left(\frac{\sqrt{\frac{\exp(x)p}{\pi}}}{\sigma_C} \times \sqrt{1 - \exp(2u)}\right)\right]}{2\Phi\left(\frac{\sqrt{\frac{\exp(x)p}{\pi}}}{\sigma_C}\right) - 1} \right. \\ &\times \frac{1}{\sigma_P \times p\sqrt{2\pi}} \times \frac{1}{\Phi\left(\frac{\mu_P}{\sigma_P}\right)} \\ &\times \left. \exp\left[-\frac{(\log(p) - \mu_P)^2}{2{\sigma_P}^2}\right] \right\} dp \, . \end{aligned} \tag{3.9}$$

3.5.2 Maximum Likelihood Estimation for the Geometrical Model

3.5.2.1 Joint distribution

Based on the general approach in Section 3.2.2 and the geometrical measurement error model in Section 3.5, the joint probability density function of Y and W, can be derived as:

$$f_{Y,W}(y,w) = \int_w^{\infty}\int_1^{\infty}\left\{\frac{1}{\sigma_\epsilon\sqrt{2\pi}}\exp\left[-\frac{(y-\beta_0-\beta_1 x)^2}{2{\sigma_\epsilon}^2}\right]\right.$$
$$\times\frac{\frac{2\exp(2w-1.5x)}{\sqrt{1-\exp(2w-2x)}}}{\sqrt{\pi}\mathrm{Derf}\left[\frac{\sqrt{\exp(x)p}}{\sigma_C\sqrt{2\pi}}\right]\left[0.5+0.5\mathrm{Derf}\left(\frac{\mu_P}{\sigma_P\sqrt{2}}\right)\right]}$$
$$\times\frac{1}{\sigma_C\sqrt{2\pi}}\exp\left[-\frac{(\exp(x)-\exp(2w-x))\,p}{2\pi{\sigma_C}^2}\right] \tag{3.10}$$
$$\times\frac{1}{\sigma_P\sqrt{2\pi}\sqrt{p}}\exp\left[\frac{-(\log(p)-\mu_P)^2}{2\sigma_P^2}\right]$$
$$\left.\times\frac{1}{\sigma_X\sqrt{2\pi}}\exp\left[\frac{-(x-\mu_X)^2}{2\sigma_X^2}\right]\right\}dp\;\;dx\;,$$

where,

$$\mathrm{Derf}(x) = 2\Phi(\sqrt{2}\times x) - 2.$$

3.5.2.2 Likelihood function

The loglikelihood function, for fixed values of μ_P, σ_P and σ_C, can be written as the sum of the contributions for each of the n observations in the data set under the geometrical measurement error model:

$$\mathcal{L}(\beta_0,\beta_1,\sigma_\epsilon,\mu_X,\sigma_X;\mu_P,\sigma_P,\sigma_C,Y,W) \;=\; \sum_{i=1}^{n}\mathcal{L}_i. \tag{3.11}$$

As discussed in Section 3.4.2.2, the experimental UT data can be truncated or censored. For data with left truncation at a level y^{TL} and right censoring at a level y^{CR}, the likelihood contributions from different types of observations are again given in Equations (3.5) and (3.6),

respectively. In the geometrical measurement error model, $f_{(Y,W)}(y,w)$ is the probability density function of Y and W and it is defined in Equation (3.10). ML estimates are obtained by maximizing Equation (3.11).

The ML method is capable of handling different types of observations, exact, censored and truncated data. An algorithm to evaluate Equations (3.5) and (3.6) using the probability density function in Equation (3.10) would, however, be computationally intensive. Thus in our examples, we do not use censoring or truncation.

3.5.3 Simulation of the Geometrical Model GEV Method

Following the same procedure that was used with the Burkel model, we generated simulated data using the same parameters values (β_0, β_1, σ_ϵ, μ_X and σ_X) used in the simulation for Figure 3.1. The other parameter values used in simulation are: $\mu_P = 1.0$, $\sigma_P = 0.36$, and $\sigma_C = 50.0$. Here the simulated response is EFBH area. One simulated data set is shown in Figure 3.11. In the figure, crosses represent simulated data with true flaw area while the circles represent simulated data with measured flaw area, (i.e., flaw area with measurement error). The measurement errors in flaw sizes cause data to shift to the right, leading to bias in the regression coefficient estimators.

We illustrate this bias further by using 50 data-generation/estimation simulations, each with a sample size of 500. The results are plotted in Figure 3.12. The single longer line is the true regression line. The short lines are naive regression lines for 50 trials. This simulation illustrates the potential strength of the bias caused by using naive method when there are substantial measurement errors in flaw sizing. The bias in regression coefficients will be propagated to the POD, leading to biased POD estimates. Please note that we did not plot the GEV lines for the 50 trails as we did in Figure 3.4 because of the computation time constraints. But GEV estimates for one of the trails are summarized in Table 3.5.3 along with the parameter values used in the simulation. The ML estimates agree very well with the true parameter values.

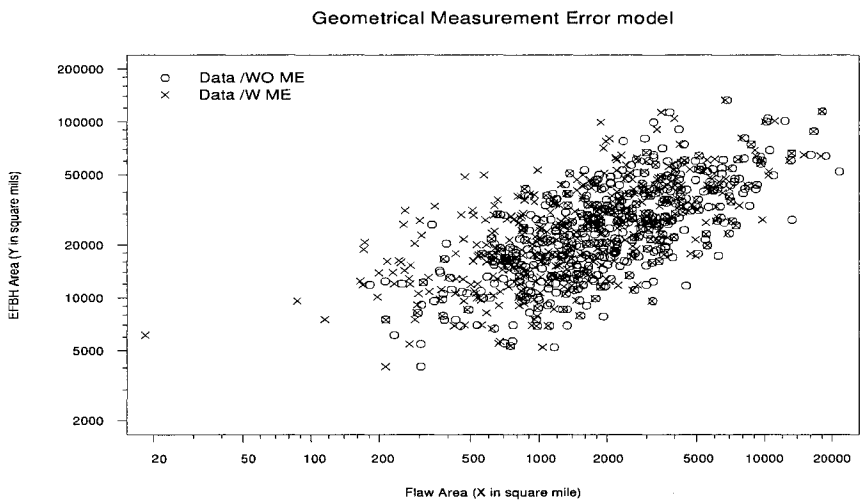

Figure 3.11 Simulated data with/without measurement error for geometrical model.

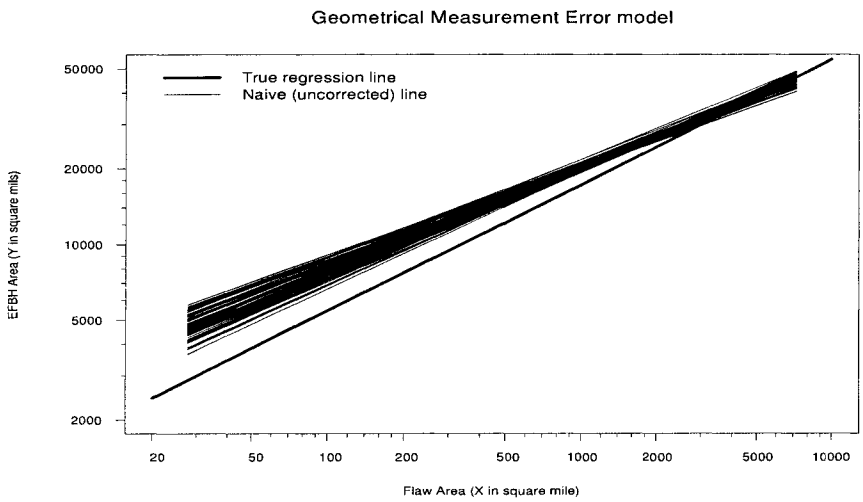

Figure 3.12 Simulated data with/without measurement error for geometrical model.

Table 3.1 Example Parameter Estimates for the Geometrical Model

Parameters	*Parameter Values*	*ML Estimates*	*Std.Err*	*95% Lower*	*95% Upper*
β_0	6.3	6.43	0.35	5.74	7.11
β_1	0.5	0.48	0.05	0.39	0.57
σ_ϵ	0.44	0.48	0.03	0.42	0.53
μ_X	7.5	7.53	0.07	7.40	7.66
σ_X	0.85	0.88	0.05	0.78	0.99

3.5.4 Application to the Simulated Inspection Data

The simulation in this section is similar to that in Section 3.4.4, except that we simulated the UT inspection data using the geometrical measurement error model. Then we analyzed the simulated data set using the corresponding GEV approach for the simulation. We assume that flaw area has a lognormal distribution with $\mu_X = 8.63$ and $\sigma_X = 1.65$. The values of the parameters used to simulate response under the $\hat{a}$ versus a model are: $\beta_0 = 1.45$, $\beta_1 = 0.3$, $\sigma_\epsilon = 0.4$. The values of the parameters used to simulate measurement error are: $\mu_P = 1$, $\sigma_P = 0.36$ and $\sigma_C = 50$ under the geometrical measurement error model. The values of μ_P and σ_P were estimated from the experimental data from a metallographic study. Here the simulated response is amplitude (i.e., %FSH). Figure 3.13 shows the simulated UT inspection data with measurement error using these parameters. Also shown are the estimated regression lines based on the naive (uncorrected) regression and the GEV method.

The two POD curves corresponding to the two regression lines (i.e., the naive estimate line and GEV estimate line) are plotted in Figure 3.14. The thin curve was estimated using the naive (uncorrected) method while the thicker curve was estimated using GEV method and the geometrical measurement error model. The comparison suggests that the naive method can lead to inaccurate POD estimate when there is substantial measurement error and that the GEV method using the geometrical model will correct for such bias.

3.6 Concluding Remarks and Future Research Work

This paper extends the classic measurement error model to NDE application. Two measurement error models, the Burkel measurement error model and the geometrical measurement

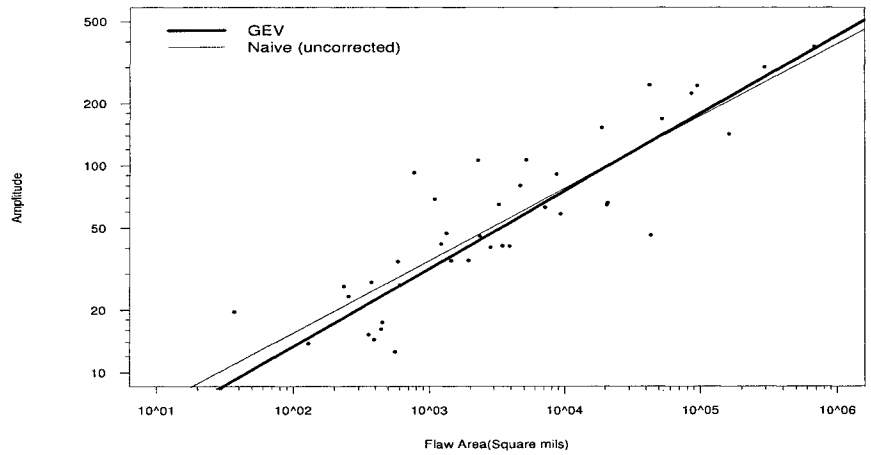

Figure 3.13 Plot showing simulated UT inspection data with measurement error under geometrical measurement error model and both naive and GEV regression lines.

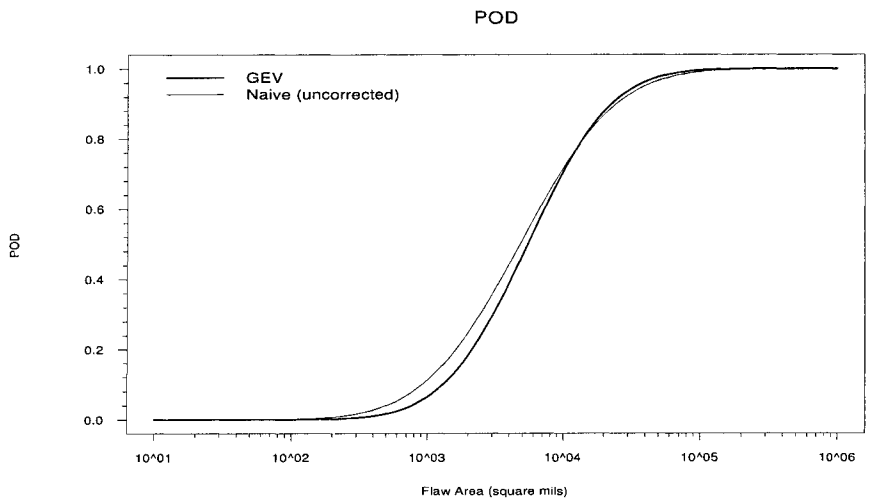

Figure 3.14 POD plots for the conventional inspection using simulated data under geometrical measurement error model.

error model, are developed. These two models can be used to adjust bias in POD computation due to flaw sizing error. The systematic simulations in this paper provide insights of how measurement error affects regression coefficients. The GEV methods that we have developed can make important corrections in regression coefficients giving more accurate estimates of POD. We demonstrate the GEV methods with simulated inspection data based on the experimental NDE inspection data.

Presently, we are not able to include data truncation and censoring in the geometrical measurement error model because of the computation restraints. A method capable of dealing with data truncation and censoring will be valuable for NDE applications.

3.7 Acknowledgements

This material is based upon work supported by the Federal Aviation Administration under Contract #DTFA03-98-D-00008, Delivery Order # 0034 and performed at Iowa State University's Center for NDE as part of the Engine Titanium Consortium program, through the Airworthiness Assurance Center of Excellence. We would like to express special thanks to R. Bruce Thompson and Floyd Spencer for their helpful suggestions relating to this research. We also acknowledge help comments on the CBS data and on our modelling methods, as they evaluated, from Jon Bartos, Richard Burkel, Waled Hassan, and Tim Mouzakis.

References

Berens, A. P. (1989), "NDE Reliability Data Analysis," *Metals Handbook* (9th ed., Vol. 17, *Nondestructive Evaluation and Quality Control*), Metals Park, OH: American Society for Metals 689-701.

Burkel, R. H., Sturges, D. J., Tucker, W. T. and Gilmore, R. S., (1996) "Probability of Detection for Applied Ultrasonic Inspection," Review of Progress in Quantitative NDE, Vol. 15, edited by D. O. Thompson and D. E. Chimenti, Plenum Press, New York, NY, 1996, 1991-1998

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

Carroll, R. J., Ruppert D., and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*, London; New York, Chapman Hall.

Brasche, L., Chiou, C. P., Thompson, R. B., Smith, K., Meeker, W. Q., Margetan, F., Panetta, P., Chenail, R., Galli, F., Umbach, J., Raulerson, D., Degtyar, A., Bartos, J., Copley, D., McElligott, R., Howard, P., Bashyam, M., Contaminated Billet Study, DOT/FAA/AR-xx/xx to be published in 2005.

MIL-HDBK-1823 (1999), Non-Destructive Evaluation System Reliability Assessment, Standardization Order Desk, Building 4D, 700 Roberts Avenue, Philadephia, PA 19111-5094.

CHAPTER 4. APPLICATION OF STATISTICAL METHODS FOR ASSESSMENT OF COMPONENTS OF VARIANCE IN PROBABILITY OF DETECTION MODELS

A paper to be submitted to Technometrics

Yurong Wang, William Meeker[1] and Waled Hassan[2]

[1] Department of Statistics

Iowa State University

Ames, IA 50011

[2] Honeywell Engines, Systems, and Services

Pheonix, AZ 85034

Abstract

Nondestructive evaluation (NDE) methods are used widely in industries to assure the integrity of critical system components. Examples include rotating components in jet engines and heat-transfer tubes in nuclear power plants. There is an important need to quantify and improve the probability of detection (POD) for NDE inspection used in both production quality control and in-service reliability. Improvement of POD, especially, requires the identification and quantification of sources of variability. A standard NDE assessment method uses a manufactured "block" of material containing seeded defects of known size and character. This block is then inspected according to an experimental design that will capture the important sources

of variability. The commonly used NDE data analysis/modelling method, known as $\hat{a}$ versus a, uses a linear regression to relate the NDE signal response to the flaw or defect size. The model behind this method contains only one component of variance for the response. There are, however, many random factors causing variability in NDE inspection. In this paper, we develop a Bayesian hierarchial model to identify and quantify the inspection variance components in the presence of data censoring. Using Markov Chain Monte Carlo (MCMC) simulation implemented in WinBUGS (the MS Windows operating system version of BUGS: Bayesian analysis Using Gibbs Sampling) software, we demonstrate the effectiveness of the Bayesian approach with simulated data and the experimental data.

Key words: Bayesian, Censoring, Hierarchical Model, Markov Chain Monte Carlo, Mixed Model, Nondestructive Evaluation, Probability of Detection

4.1 Introduction

4.1.1 Background and Motivation

Nondestructive Evaluation (NDE) is widely used in various industries for ensuring quality and reliability because of its advantages of lower cost and repeatability, when compared to destructive methods. Important applications include rotating components of jet engines and heat-transfer tubes in nuclear power plants. There are, however, many factors that cause inspection variability, necessitating a probabilistic characterization of inspection capability. Factors that can affect the performance of an inspection system include material properties, flaw geometry, flaw orientation, operator differences and so on. These factors can be partitioned into three groups: factors relating to the inspection system ($\underline{x}_{\text{SYS}}$), factors relating to the material ($\underline{x}_{\text{PART}}$), and factors relating to the flaw itself ($\underline{x}_{\text{FLAW}}$).

1. The factors $\underline{x}_{\text{SYS}}$ relating to a NDE inspection system include the transducer parameters, scan resolution (mechanical increment of the scanning system in X and Y dimensions), system alignment/angulations, as well as operators in the experiment.

2. The factors $\underline{x}_{\text{PART}}$ relating to the part to be inspected include part geometry (particularly the degree of curvature at the inspection location), material microstructure, anisotropy, surface roughness, etc.

3. The factors $\underline{x}_{\text{FLAW}}$ characterizing a flaw include flaw size, shape, orientation, depth and density (e.g., percent nitrogen and degree of cracking and voiding for a hard alpha inclusion), etc.

Figure 4.1 is the ultrasonic test C-scan image of 32 synthetic hard alpha flaws with nominal sizes #5, #4, #3 and #2 (corresponding to cylindrical flaw diameters of 5/64 inches, 4/64 inches, 3/64 inches and 2/64 inches, respectively). The image gives the strength of the ultrasonic image at each of 1100 by 400 pixels. In each of the 4 rows, 8 flaws with the same nominal size have different ultrasonic responses. The variability within a flaw size is caused by the random factor, $\underline{x}_{\text{FLAW}}$.

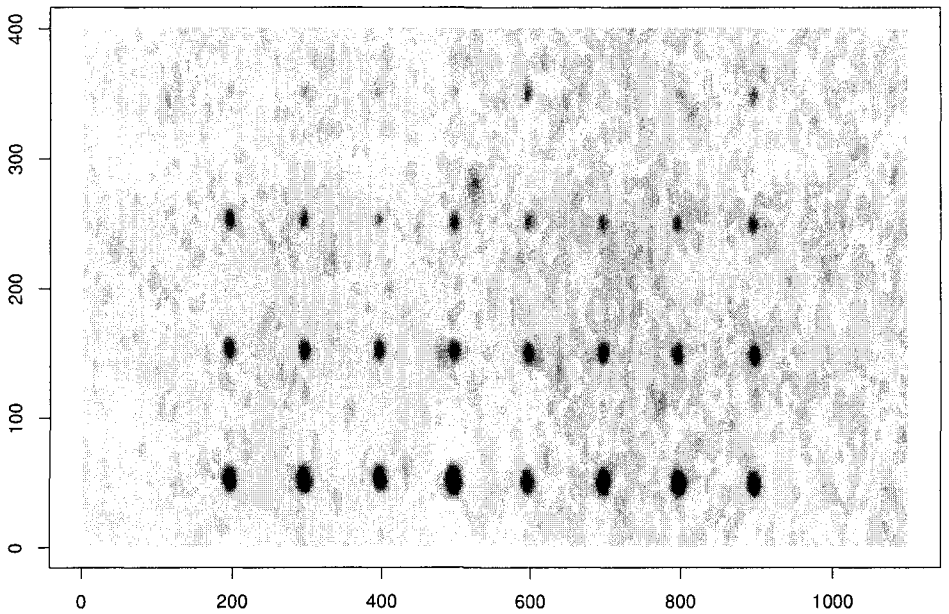

Figure 4.1 Measured ultrasonic responses from synthetic hard alpha flaws.

Excessive variability from various sources can degrade NDE inspection quality. There are important needs to identify and quantify sources of variability in NDE applications. Such qualitative and quantitative information about variability can provide basis for later experimental designs to improve NDE inspection. In NDE applications involving designed experiments, it is common that data are left censored due to known flaws being missed or right censoring due to signal saturation. In this paper, we use a Bayesian approach to estimate components of variance in NDE inspection processes, allowing for censored data.

4.1.2 Related Work

Variance component analysis has been widely used in industry. An comprehensive review can be found in Searl (1992). Searl (1992) also discusses analysis of variance for unbalanced data, predictions of random variables, and hierarchical models. These methods are commonly

used to identify critical factors affecting process variability (and thus product quality) and to improve product quality and reliability by either controlling these key factors or making the system less sensitive to the noises. There is much classical literature on variability component analysis, most of which uses mixed effects models. Methods for analyzing censored data are well established for fixed effects models. For example, Meeker and Escobar (1998) discussed such methods. But only a few works in the literature deal with data censoring in the presence of random effects. For example, Feiveson and Kulkarni (2000) utilized Bayesian methods to analyze censored data using the mixed effects model.

Statistical software packages have developed functions or procedures for variance component analysis. These include S-plus (function LME) and SAS (PROC MIXED and the GLIMMIX macro). But none of these functions or procedures has the capability of dealing with censored data. When data are unbalanced and either censored or truncated, variance component analysis is difficult using "classical" likelihood-based methods or REML. This is because the likelihood requires evaluation of high dimensional integrals. Bayesian methods provide an important and useful alternative statistical method. Congdon (2003) illustrates the Bayesian approach to data analysis and modelling in various applications using the WinBUGS software. The Bayesian approach can handle complicated problems of variance component analysis, even allowing for data censoring.

Besides the aforementioned work on variance component analysis in the statistical literature, Hassan (2002) conducted a study to assess multizone inspection capability based on the average of the experimental NDE data. We use his experiment and data as the basis for our work in this paper.

4.1.3 Overview

The paper is organized in the following way: Section 4.2 reviews the classical mixed effects model. Section 4.3 outlines the experimental study and data. Section 4.4 describes the mixed effects model for our application. Section 4.5 provides a general Bayesian approach for variance component analysis for NDE experiments. Section 4.6 describes the MCMC simulation technique used in our application. Sections 4.7 and 4.8 analyze the simulated data and the

experimental data respectively, to quantify the variability of various sources as example of validation and application. Section 4.9 contains concluding remarks and describes areas for future research.

4.2 Classical Mixed Effects Model

Mixed effects models are useful for estimating components of variance. The mixed effects linear model can be written as:

$$Y = X\beta + Zb + \epsilon, \tag{4.1}$$

where,

- Y is the response vector of length n.
- X is the $n \times p$ model matrix for fixed factors and β represents the p fixed effects parameters.
- Z is the $n \times r$ model matrix for the random factors and b is the random vector of length r containing the random effects parameters.
- The means of both b and ϵ are 0.
- b and ϵ are statistically independent.

4.3 Experimental Study

4.3.1 Experimental Design

As mentioned in Section 4.1, there are many factors that can cause variability in NDE inspections. A factorial experiment was designed to study NDE multizone ultrasonic inspection system in order to obtain information on variance components and further improve billet inspection. The multizone system does simultaneous data acquisition in four to six different depth zones of a cylindrical billet.

The experimental sample or block used in this example is an 8 inch-diameter titanium billet 28 inches in length. In this titanium forging block, synthetic flaws were seeded. The nominal diameters of these flaws are #3, #4, and #5 respectively, corresponding to flat bottom hole

(FBH) sizes. Here the flaw size measure was adapted from the FBH standard where #3 is 3/64 inches, #4 is 4/64 inches in diameter, etc. The flaws are located at specified depths from 0.2 inches to 4.0 inches below the surface of the billet.

In the variance assessment experiment, the 8 inch-diameter titanium billet containing synthetic flaws was inspected in three sites, corresponding to different billet manufactures. In the multizone inspection systems used in this study, there were 5 zones covering the range of depths needed to be inspected. Each zone within a site has a separate transducer. Zone 1 starts at a depth of 0.2 inches and ends at a depth of 0.9 inches. Zones 2, 3, 4, and 5 cover the following depths, 0.9 inches to 1.8 inches, 1.8 inches to 2.7 inches, 2.7 inches to 3.6 inches, and 3.6 inches to 4.5 inches, respectively. In zone i, k_i flaws were seeded in the block at different depths within each zone. The nominal flaw sizes were #3, #4, and #5. The exact values of k_i, $i = 1, \ldots, 5$ are proprietary.

At each site, the titanium billet was inspected four different times (four runs) with at least three different operators, during four different shifts (in some cases, an inspection was started in one shift and finished in another). At the beginning of each run at an inspection site, the inspection system was set up separately for each zone. The setup includes a step to calibrate the inspection system so that the amplitude response from #2 FBH is 80% full screen height (FSH) on an oscilloscope. In addition to this calibration, the setup also involves transducer alignment, water path, and other instrument settings.

4.3.2 Inspection Data

C-Scan image data were collected from each of 4 runs at 3 different sites. These data were taken on indication amplitude in terms of % FSH (% full screen height on an oscilloscope) and signal-to-noise ratio (SNR). In addition, the depth of the indications was measured by using the longitudinal wave speed and the time of arrival of the signal from the seeded flaws.

In NDE experiments, data are often censored. Indication amplitude or SNR may be left censored due to flaw misses or right censored due to signal saturation. In addition to data censoring, some flaws were "seen" in more than one zone in the same run. In our analysis, we only use data from one zone for a given flaw. For our analysis we allow a particular flaw to

appear only once within each run. The selection criterion is to keep the zone in which the flaw has the largest amplitude, when compared to other zones, for the same run. Figure 4.2 shows the real inspection data used in this paper. For each (#3, #4 and #5) FBH size, there are k_i different flaws at different depths in zone i. Each flaw appears 12 times in the data set because each flaw was inspected in 4 runs within each of 3 sites. Figure 4.2 is a plot of percent screen height versus size for all of the data from the multizone experiment. Some horizontal "jitter" was introduced into the plot in order to make it easier to see the data at the single flaw size. This plot also shows much variability which may be caused by the factors other than flaw size: Flaw-to-flaw, Run, Setup or measurement error. The source of flaw-to-flaw variability is caused by both flaw morphology and local microstructural differences from flaw to flaw. Flaw depth has little or no effect because the multizone inspection system uses separate calibration in each zone, resulting in increased gain in deeper zones.

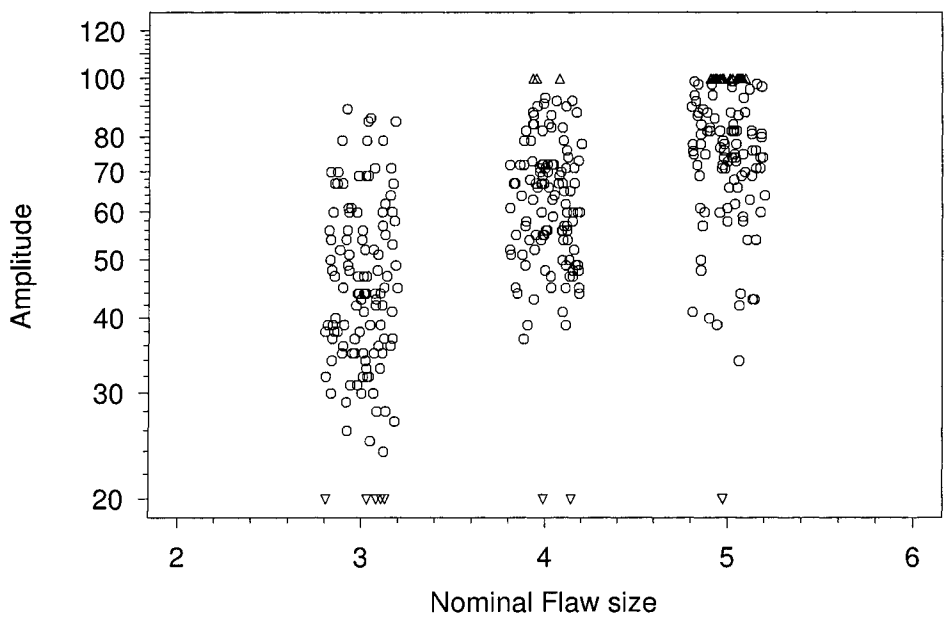

Figure 4.2 Plot of the multizone amplitude inspection data. Note the saturated observations at 100% FSH.

In the following sections, we will provide the variance component analysis using the Bayesian approach. Section 4.5 describes the Bayesian method for variance component analysis including the Bayesian method and MCMC solution. Sections 4.7 and 4.8 provide the components of variance analysis for the simulated data and the actual experimental data, respectively.

4.4 Mixed Effects Model

There are both fixed factors and random factors in the multizone NDE experiment described in Section 4.3. Fixed factors include nominal flaw size and site. Flaw-to-flaw is a random factor due to variability in flaw characteristics like flaw morphology and local microstructural differences from flaw to flaw. When the synthetic flaws with the same nominal size are seeded in the billet, their shapes, orientations and local microstructure around flaws are hard to control exactly. The nominal size is the flaw size that was in the specification for the block when it was manufactured. For example, when to produce a #3 synthetic flaw, ideally, the flaw would be 3/64 inches in diameter. But usually the true flaw size is not exactly 3/64 inches in diameter, due to fabrication and microstructure variability. Another random factor is inspection run within each site because of setup tasks such as alignments of the transducers, that are carried out before each inspection run. In the multizone inspection, operators do adjustment and alignment for transducers in each zone within an inspection run. Thus each combination of transducer and run within a site corresponds to one random setup. Obviously, the data set in this paper is multilevel (hierarchical) which includes the information at three levels of analysis: setup, run and site. The corresponding model is a hierarchical mixed effects model. An important advantage of a hierarchical mixed effects model is its ability to directly incorporate data at multiple levels of analysis, at the cost of increased complexity. The hierarchical mixed effects model can be written as

$$\begin{aligned} Y = \quad & \beta_0 + \beta_1(\log(\text{Nominal Size}) + \text{Flaw-to-flaw}) + \text{Site} + \\ & \text{Run (Nested in Site)} + \text{Setup(Nested in Run)} + \epsilon, \end{aligned} \tag{4.2}$$

where,

- Y is the response, the logarithm of indication amplitude. The units of indication amplitude is percent of FSH, which is proportional to voltage produced by the transducer.

- Nominal Size is the nominal flaw size, and β_0 and β_1 represent the regression coefficients. In our computations, however, we used a reparameterization of $\beta_0^\star = \beta_0 - \beta_1 \times \log(\overline{\text{Nominal Size}})$, corresponding to the use of the centered log size values with this reparameterization. $\hat{\beta}^\star$ and $\hat{\beta}_1$ have little correlation, improving the performance of our numerical methods.

- Site is the fixed effect parameter for the three different sites. Site_3 is considered as a baseline (i.e., $\text{Site}_3 = 0$).

- Flaw-to-flaw is the random effect parameter representing differences among the flaws with the same nominal size.

- Run and Setup are random effect parameters for experimental runs and setups, respectively.

- ϵ denotes random residuals, representing measurement error in responses.

In this paper, we assume that the random factors have normal distributions. For example, the distribution of ϵ is assumed to have a normal distribution with mean 0 and standard deviation, σ_ϵ. The distribution of random factors, Flaw-to-flaw, Run and Setup are also assumed to have normal distribution with mean 0 and standard deviations $\sigma_{\text{Flaw-to-flaw}}$, σ_{Run}, and σ_{Setup}, respectively. Our model assumes the random factors and ϵ are all independent. The eight unknown model parameters are

$\theta = (\beta_0^\star, \beta_1, \text{Site}_1, \text{Site}_2, \sigma_{\text{Flaw-to-flaw}}, \sigma_{\text{Run}}, \sigma_{\text{Setup}}, \sigma_\epsilon)'$.

4.5 Bayesian Model

4.5.1 Bayesian Hierarchical Model

In contrast to non-Bayesian methods, Bayesian methods use probability distributions to quantify uncertainty in unknown model parameters (e.g., the eight unknown model parameters

mentioned in Section 4.4). More specifically, Bayesian methods assign prior distributions to characterize prior knowledge about parameter values that is available before data collection, and uses the joint posterior distribution of parameters given the data as the basis of inference. In the mixed effects model described by (Equation 4.2), the random effects parameters are random due to the physical reasons and are used to describe the inspection process variability. For example, the flaw-to-flaw effect is random. The flaw seeding process can introduce variability to flaw size because it is impossible to exactly control flaw characteristics and the local microstructure around a flaw. The mixed effects model captures these characteristics using a random flaw-to-flaw term. For Bayesian inference, however, a prior distribution is used to describe the uncertainty about parameters of interest. The Bayesian hierarchical model includes "random" elements for both uncertainty about parameters and random factors in the classical mixed effects model. The Bayesian hierarchical model in this paper uses the mixed effects model described in Section 4.4 with prior distributions on the 8 unknown parameters. The parameters of the prior distributions are called hyperparameters.

4.5.2 Prior Distributions

Bayesian analysis has two inputs: the prior distribution (including the hyperparameters) and the likelihood. The prior distribution is a probability distribution on the parameter space which reflects prior knowledge on a set of unknown parameters and this distribution is denoted by $\pi(\theta)$. The likelihood function reflects the information on parameters from data Y and is denoted by $f(y|\theta)$. Bayes theorem combines $\pi(\theta)$ and $f(y|\theta)$ to give the posterior distribution, a probability measure on the parameter space that combines the information of the parameter in the prior and likelihood. In particular, the posterior distribution for θ, given prior $\pi(\theta)$ and data y is:

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}, \tag{4.3}$$

where, in our example, $\theta = (\beta_0^\star, \beta_1, \text{Site}_1, \text{Site}_2, \sigma_{\text{Flaw-to-flaw}}, \sigma_{\text{Run}}, \sigma_{\text{Setup}}, \sigma_\epsilon)'$. The likelihood $f(y|\theta)$ depends on the mixed effects model defined in Equation (4.2) and the available data.

Prior distributions may be chosen to be informative if prior information is available. When little or no prior information is available on a parameter, it is common practice to specify

a diffuse (approximate flat) prior distribution. A diffuse prior distribution should have little effect on the posterior distribution. On the other hand, informative priors can have a strong influence on the posterior distribution, especially when there is not much information in the data. Informative priors are typically obtained from past data, experience, or expert opinion.

Because external prior information is not available for the inspection process that generated the data described in Section 4.3.2, we will use independent diffuse prior distributions for hyper-parameters needed in the Bayesian hierarchal model.

We first consider the hyperparameters for the unknown fixed effects parameters. We assume that $\beta_0^\star$, β_1, Site_1, and Site_2 have independent normal distributions with a very large standard deviation:

$$\begin{aligned} \beta_0^\star &\sim N(0.0, 1000000.0) \\ \beta_1 &\sim N(0.0, 1000000.0) \\ \text{Site}_i &\sim N(0.0, 1000000.0), \text{ i=1, 2.} \end{aligned}$$

In the WinBUGS software, however, the spread of the prior distribution is specified in terms of "precision" which is defined as the reciprocal of the variance.

As mentioned in Section 4.4, the random effects flaw-to-flaw, Run, Setup and the random measurement error all have independent normal distributions with mean 0 and that their corresponding standard deviations are $\sigma_{\text{Flaw-to-flaw}}$, σ_{Run}, σ_{Setup}, and σ_ϵ, respectively. The assumed diffuse prior distributions for all three of these unknown standard deviations are:

$$\frac{1}{\sigma^2} \sim \text{Gamma}(0.001, 10000000), \tag{4.4}$$

where, Gamma(α, β) has the following density function and parameterizations:

$$\begin{aligned} f(z;\alpha,\beta) &= \frac{1}{\beta^\alpha \Gamma(\alpha)} z^{\alpha-1} \exp^{-\frac{z}{\beta}}, \quad z >= 0 \\ \alpha &= \text{shape} \quad \text{parameter} > 0, \\ \beta &= \text{scale} \quad \text{parameter} > 0. \end{aligned}$$

4.5.3 Posterior Distributions

Let $\mu = \beta_0^{\star}+\beta_1(\log(\text{Nominal Size})-\overline{\log(\text{Nominal Size})})+\text{Site}$ and let Σ denote the covariance matrix of the response vector $Y = (y_1, y_2, \ldots, y_n)'$. The likelihood can be written as:

$$f(\theta|Y) = \int_{y_1^l}^{y_1^u} \cdots \int_{y_i^l}^{y_i^u} \cdots \int_{y_n^l}^{y_n^u} \phi_{Norm}(y_1, \ldots, y_i, \ldots, y_n|(\mu, \Sigma))dy_1 \ldots dy_i \ldots dy_n, \tag{4.5}$$

where,

- When y_i is left censored, $y_i^l = -\infty$ and $y_i^u = y_i^{CL}$, y_i^{CL} is the left censoring level for y_i.
- When y_i is right censored, $y_i^l = y_i^{CR}$ and $y_i^u = \infty$, y_i^{CR} is the right censoring level for y_i,
- When y_i is not censored, $y_i^l = y_i - \delta$ and $y_i^u = y_i + \delta$.

Here,

$$\phi_{Norm}(Y|(\mu, \Sigma)) = \left(\frac{1}{2\pi}\right)^{(\frac{n}{2})} (|\Sigma|)^{(-\frac{1}{2})} \exp((Y-\mu)'\Sigma^{-1}(Y-\mu))$$

is the joint probability density function of a multivariate normal distribution of dimensions n, where n is the length of the response vector Y. Σ is the covariance matrix of $n \times n$ depending on unknown parameters $\beta_1, \sigma_{\text{Flaw-to-flaw}}, \sigma_{\text{Run}}, \sigma_{\text{Setup}}$, and σ_ϵ. The variance or covariance elements in Σ can be described as follows,

- The variance of an observation is $\beta_1^2\sigma^2_{\text{Flaw-to-flaw}} + \sigma^2_{\text{Run}} + \sigma^2_{\text{Setup}} + \sigma^2_\epsilon$
- The covariance between two observations from the same flaw but different runs is $\beta_1^2\sigma^2_{\text{Flaw-to-flaw}}$
- The covariance between two observations from the different flaws but same run and same setup is $\sigma^2_{\text{Run}} + \sigma^2_{\text{Setup}}$
- The covariance between two observations from the different flaw but same run and different setups is σ^2_{Run}
- For other cases, the covariance between two observations is 0.

The likelihood in Equation (4.5) is very complicated. Thus the posterior distribution based on the likelihood is complicated and intractable for direct computation

4.6 Evaluation of Posterior Distributions via Simulation

4.6.1 Simulation

For Bayesian inference, one must evaluate the marginal posterior distribution for parameters and functions of parameters of interest. Direct evaluation of marginal posterior distributions requires the computation of the high-dimensional integrals with respect to posterior distributions. It is impossible to evaluate the high-dimensional integrals using conventional numerical methods. Special simulation methods have, however, provided useful approximations of these integrals. The basic idea is to obtain information about the posterior distribution by drawing a large sample of parameter vectors from the posterior distribution. Let $\underline{\theta} = (\theta_1, \ldots, \theta_p)$ be the p unknown model parameters. Practical applications require estimation of functions of $\underline{\theta}$, such as the POD at a given flaw size. Let $h(\underline{\theta})$ denote such a function. Direct evaluation of the marginal posterior distribution of $h(\underline{\theta})$ is usually impossible due to the required high dimensional integration. Let $\underline{\theta}_1, \ldots, \underline{\theta}_B$ denote an $i.i.d$ sample from the posterior distribution of $\underline{\theta}$. Typically, B would be on the order of 100. Then $h(\underline{\theta}_1), \ldots, h(\underline{\theta}_B)$ is a sample from the marginal posterior. If draws from the posterior distribution are available at acceptable computational cost, we could simply use the mean of a large number of $h(\underline{\theta}_k)$ values to estimate the mean of the marginal posterior and the quantiles of the empirical distribution given by $h(\underline{\theta}_1), \ldots, h(\underline{\theta}_B)$ to define credible intervals for $h(\underline{\theta})$.

4.6.2 MCMC and WinBUGS

The general problem of drawing an $i.i.d$ sample from an arbitrary multivariate distribution is difficult. This is particularly true for the complicated hierarchal model in this paper where the posterior distribution is a nonstandard multidimensional distribution. However, instead of drawing $i.i.d$ samples, it is generally possible to define an algorithm that samples from a Markov chain that has the desired posterior as its stationary distribution. The statistic of interest can be estimated using this sequence of samples. This method is called MCMC simulation. WinBUGS is a versatile package that has been designed to carry out MCMC computations for a wide variety of Bayesian models. Having specified the model as a full joint distribution on all

quantities including both parameters and other unobservables in the likelihood, one can sample values from their conditional (posterior) distribution given the stochastic nodes that have been specified. In WinBUGS system, stochastic nodes may be observed (i.e., data), or may be unobserved, like parameters and observations that are unobservable due to censoring. Within WinBUGS, an expert system will choose the sampling method used to produce a sequence of samples (Congdon, 2003). The sampling methods are used in the following hierarchies (in each case a method is only used if no previous method in the hierarchy is appropriate):

- Direct sampling using the Gibbs sampler, if conjugacy is identified.
- Derivative-free adaptive rejection sampling (Gilks, 1992) for non-conjugate problems with log-concave sampling densities.
- Slice sampling (Neal, 1997) for non-conjugate problems without log-concavity sampling density but on a restricted range of the parameters .
- Current point Metropolis for non-conjugate problems without log-concavity with an unrestricted range of the parameters.

The basic idea behind the Gibbs sampling algorithm is to successively sample from the conditional distribution of each node given all of the other nodes (these are known as full conditional distributions). A slice-sampling algorithm is used for non log-concave densities on a restricted range. This has an adaptive phase of 500 iterations which will be discarded from all summary statistics. The current Metropolis MCMC algorithm is based on a symmetric normal proposal distribution, whose standard deviation is tuned over the first 4000 iterations in order to get an acceptance rate of between 20% and 40%. See Spiegelhalter, Thomas, Best, and Lunn (2003) for introductions to WinBUGS software and MCMC methodology.

4.6.3 Full Conditional Distribution and Gibbs Sampling in WinBUGS

The Bayesian analysis in this paper (implemented in WinBUGS) used the Gibbs sampling algorithm. The Gibbs sampling algorithm requires specification of the full conditional distributions for the unknown parameters. Gelfand and Smith (1990) developed a general framework

for calculating marginal densities needed for Gibbs sampling. Gelfand, Smith and Lee (1992) extended the general framework for full conditional distributions for truncated distributions, constrained parameter models, censored data models and group data models.

To handle censored data, the basic idea is to treat each censored observation as an unknown parameter. To avoid the difficulty of specifying full conditional distributions and sampling in case of data censoring, they treated the response Y as an unobservable and included it in the Gibbs sampler. Following Gelfand, Smith and Lee (1992), we use the following notation.

- [ ] denotes the density function.
- Y denotes the response vector of length n. V and W denote vectors giving the censoring levels for the response vector (In general, different responses may have different censoring levels).
- θ denotes the unknown parameters.
- $[\lambda]$ denotes the independent prior distribution defined in Section 4.5.2 for the unknown parameters θ. $[\eta]$ denotes the prior distribution for V and W.
- Z is defined as

$$Z_j = \begin{cases} V_j & Y_j \leq V_j \\ Y_j & V_j < Y_j < W_j \\ W_j & Y_j \geq W_j \end{cases} \tag{4.6}$$

According to the general framework developed by Gelfand, Smith and Lee (1992), the full conditional distribution for $\theta = (\beta_0^\star, \beta_1, \text{Site}_1, \text{Site}_2, \sigma_{\text{Flaw-to-flaw}}, \sigma_{\text{Run}}, \sigma_{\text{Setup}}, \sigma_\epsilon)'$ can be expressed as

$$[\theta_i | Z, Y, V, W, \theta_j, j \neq i, \lambda] \propto [Y|V, W, \theta][\theta|\lambda] \tag{4.7}$$

The remaining full conditionals required by the Gibbs sampler are given by $[\eta|Z, Y, V, W, \theta, \lambda] \propto [V, W|\eta][\eta]$, and $[\lambda|Z, Y, V, W, \theta, \eta] \propto [\theta|\lambda][\lambda]$ and finally, $[Y|Z, V, W, \theta, \eta, \lambda]$ $\propto [Z|Y, V, W][Y|V, W, \theta]$. For illustration, one typical case is considered. Y_j are conditional independent and have full conditionally distribution as follows

- If $V_j < Z_j < W_j$, then Y_j is not censored and the full conditional distribution of Y_j is degenerate at Z_j.
- If $Z_j = V_j$, then Y_j is left censored and the full conditional distribution of Y_j is restricted to $[-\infty, V_j]$.
- If $Z_j = W_j$, then Y_j is right censored and the full conditional distribution of Y_j is restricted to $[W_j, \infty]$.

Given the conjugate normal priors and Gamma prior distributions defined in Section 4.5.2, the full conditionals for unknown parameters are the updated conjugate forms obtained by standard Bayesian analysis. The full conditional distribution for Y is also a normal distribution. Sampling Y_j is therefore routine according to full conditional distributions.

But in our application, $[\eta]$ and its conditional distributions are degenerate distributions due to the fixed values in V and W.

4.7 Bayesian Application for the Simulation Study

4.7.1 Model

To illustrate the ideas used in our analysis, and to assess our ability to accurately estimate components of variance, we first analyze simulated data that have variance components corresponding to the actual application that motivated this research. Using the simulated data, we can assess how well the Bayesian approach will estimate the variance components in the NDE application. The simulated data have the same data structure as the experimental multizone data. In particular, we simulated data from 3 sites, with 4 runs at each site (a total 12 runs). Flaws have 3 nominal sizes, #3, #4 and #5 FBH size. In the simulation, we assume,

- Actual flaw size is different from nominal flaw size, caused by variation in flaw-to-flaw. log(Actual flaw size) = log(nominal flaw size) + flaw-to-flaw
- Site has a fixed effect on flaw response because the different three sites have slightly different inspection procedures and different equipment.

- Both run and setup are random effects. Run is nested within site and setup is nested within run.

In addition, the simulated flaw response may be left censored (missed), right censored (saturated) or exact. The hierarchial model used in the simulation is described in Section 4.5.1. The true values of the parameters used in the simulation are given in the Table 4.1.

4.7.2 WinBUGS Inputs

A WinBUGS program listed in appendix was written to analyze the data. Even for a standard problem (e.g., normal distribution prior), the direct coding in terms of full conditional densities can be complicated. In WinBUGS, however, users do not need to specify the actual full conditional distributions and only need to provide the simpler inputs to WinBUGS.

The necessary inputs to WinBUGS are specifications of the prior distributions and the model. The hierarchal structure of models can be implemented by using indexing. The WinBUGS program includes two parts:

- Model specifications, corresponding to specification of the likelihood.
- Prior distribution for unknown model parameters and the corresponding hyperparameters

Based on the prior distributions and the likelihood in WinBUGS, each MCMC trial results in values of the unknown parameters and responses corresponding to the censored observations. The sampling method for censored observations can be found in subsection 4.6.3 based on the conditional distribution inferred by WinBUGS. Censoring is indicated in the WinBUGS system by using the notation I(lower, upper). For example,

$$y \sim \text{dnorm}(\mu, \tau)\text{I}(\text{lower}, \text{upper}).$$

indicates a response y from the normal distribution with parameters μ and τ (the precision, $\tau = 1/\sigma^2$), which had been observed to lie between lower and upper. Leaving either lower or upper blank corresponds to no limit. For example, I(lower,) corresponds to an observation known to lie above lower, i.e., a right censored observation. Similarly, I(,upper) corresponds to a left censored observation.

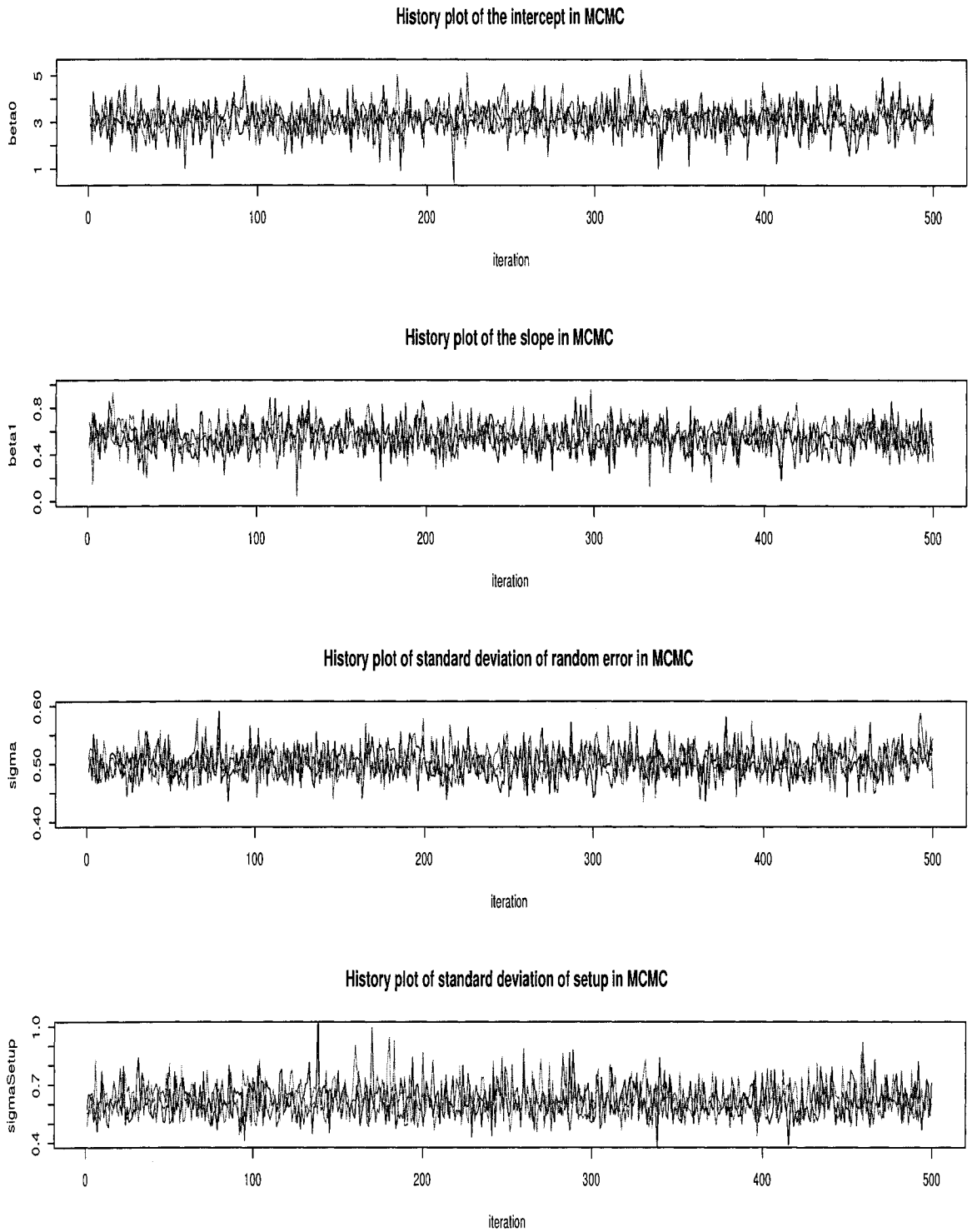

Figure 4.3 Plot of the three thinned MCMC samples for four of the unknown parameters.

4.7.3 Simulation Data Analysis

The specified model was run with three chains of 50,000 iterations each. The first 25,000 iterations were discarded as a burn-in, to assure that no startup transients are in the MC sample. The purpose of such a large number of iterations is to provide a high degree of assurance that the sequence has converged, especially for sequences that might have a very slow convergence rate. The purpose of using multiple chains is to ensure the convergence of the sampling process. If the samples from the three different chains are similar, it is evidence of successful convergence. Although using a single chain sometimes is acceptable for a relatively straightforward problem, using multiple parallel chains is preferable to reduce the chance that the sampling will be trapped in some small region, which would result in incorrect results. Another consideration in WinBUGS programming is to choose a parametrization that will make convergence faster. For our model, as explained in Section 4.4, we centered the explanatory variable in the linear regression relationship between the response and flaw size. If this had not been done, the values of β_0 and β_1 in the MCMC samples would be highly correlated, causing convergence to be relatively slow.

The time series plots of three chains are used to monitor the convergence of the samples. Figure 4.3 shows the history of the samples of the parameters in the simulation. We are confident that convergence has been achieved because all three of the chains appear to be overlapping one another for all the parameters.

The Gelman-Rubin scale reduction factor is another useful metric for assessing convergence of MCMC sampling procedures and can be used to decide if all parameters have converged. The scale reduction factor compares variation in the sampled parameter values within and between chains. If there is a wide divergence in the sample paths between different chains, variability of sampled parameter values between chains will significantly exceed the variability within chains (Congon, 2003). A value of the Gelman-Rubin Rhat statistic under 1.2 indicates approximate convergence (Congon, 2003). In our application, the values of Gelman-Rubin Rhat statistic are 1, indicating that the chains have converged to the same distribution.

For the fixed effect factor Site, we set the effect of the baseline to $\text{Site}_3 = 0$. The effects of

Table 4.1 Parameter Estimation for the Simulated MultiZone Inspection

Parameters	*True Value*	*Bayes Estimate*	*Std.Err*	*95% Lower*	*95% Upper*	*Rhat*
β_0	3.0	3.14	0.58	1.94	4.32	1
β_1	0.5	0.55	0.12	0.33	0.80	1
Site_1	0	-0.63	0.81	-2.24	1.15	1
Site_2	0	-0.20	0.81	-1.76	1.42	1
$\sigma_{\text{Flaw-to-flaw}}$	0.5	0.40	0.10	0.22	0.62	1
σ_{Run}	0.5	1.06	0.32	0.39	1.61	1
σ_{Setup}	0.5	0.62	0.08	0.48	0.80	1
σ_ϵ	0.5	0.50	0.02	0.46	0.55	1

Site_1, and Site_2, shown in Table 4.1, correspond to differences in the size of the effect between them and Site_3. Table 4.1 shows that the 95% Bayes credible intervals contain the true value for each of the parameters. All of the estimates are close to the true values of parameters except for σ_{Run} and the effect of sites. The reason that the estimate of σ_{Run} (1.06) deviates true value (0.5) is the small number of Runs (12) in the NDE experiment.

4.8 Bayesian Approach for the Experimental Study

In this section we fit the Bayesian hierarchial model to the experimental data described in Section 4.3. Again, we used MCMC techniques, as implemented in the WinBUGS software package. Similar to Section 4.7, the model was run with 3 chains with 50,000 iterations each using the experimental data.

Table 4.2 Parameter Estimation for the Multizone Inspection Experiment

Parameter estimates	*Mean*	*Std.Err*	*95% Lower*	*95% Upper*	*Rhat*
β_0	-0.47	0.6	-1.74	0.74	1
β_1	0.58	0.08	0.42	0.74	1
Site_1	-0.01	0.07	-0.14	0.13	1
Site_2	-0.2	0.07	-0.32	-0.06	1
$\sigma_{\text{Flaw-to-flaw}}$	0.30	0.07	0.20	0.48	1
σ_{Run}	0.04	0.03	0.01	0.11	1
σ_{Setup}	0.16	0.02	0.12	0.21	1
σ_ϵ	0.20	0.01	0.18	0.22	1

Table 4.2 summarizes the parameter estimates from the fitted model. We can see that the

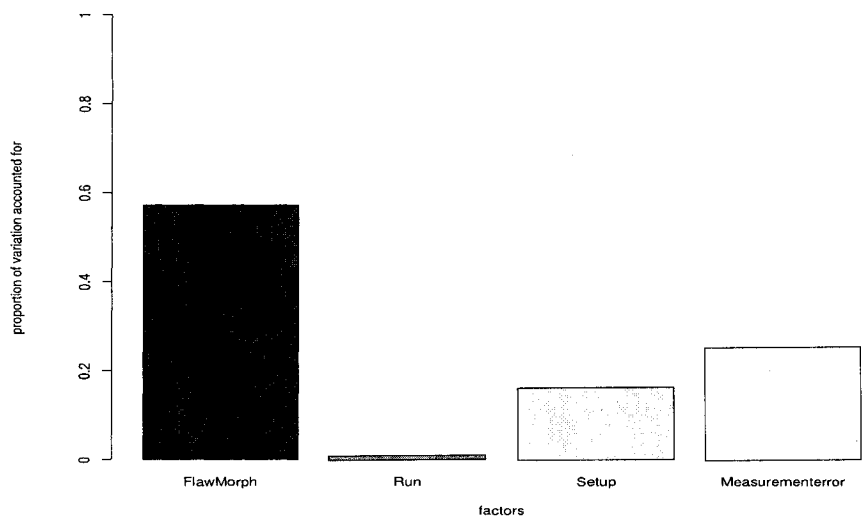

Figure 4.4 The proportion of variation accounted for by sources.

estimated coefficient of β_1 is statistically significantly positive. This is consistent with the fact that larger flaws tend to produce stronger NDE responses. The estimate of the sites effect suggested that site has an important effect on the NDE responses. Figure 4.4 shows the proportions of variability accounted for by the sources, flaw-to-flaw, random error, setup and run. The figure shows that flaw-to-flaw accounts for 60% of total inspection variability. This big proportion indicates that flaw-to-flaw is a key driver for the inspection variability. Run-To-Run variability only accounts for 5% of the total inspection variability. The proportions of variability are 25% and 20% for random error and Setup, respectively. The small contribution of run-to-run variability to the total variability is probably because most of the variability due to adjustment or alignment are caused by Setup in the experiment.

4.9 Concluding Remarks and Areas for Future Research

In this paper, we developed a Bayesian hierarchial model to identify and quantify the variance components of inspection in the presence of data censoring. The Bayesian approach is demonstrated with simulated data and experimental data using MCMC simulation in Win-

BUGS.

The estimates of variance components from the experimental data suggest that flaw-to-flaw has the most important influence on inspection variability. This suggests that determination and use of actual flaw size (sometimes called "ultrasonic flaw size") should be used in data analysis assess inspection variability.

This Bayesian approach could be extended to allowing for truncation. For example, in some NDE studies, when there are misses that are not recorded (as in field data), the response can be viewed as having come from a left-truncated distribution and are said to be "left truncated data" (Burkel, Sturges, Tucker, and Gilmore, 1996). In this paper, the developed Bayesian approach deals with continuous responses. This approach could also be applied to "hit/miss" data, where the data has binary response, as described in MIL-HDBK-1823 (1999).

4.10 Acknowledgements

This material is based upon work supported by the Federal Aviation Administration under Contract #DTFA03-98-D-00008, Delivery Order # 0034 and performed at Iowa State University's Center for NDE as part of the Engine Titanium Consortium program, through the Airworthiness Assurance Center of Excellence. We would like to express special thanks to R. Bruce Thompson and Floyd Spencer for their helpful suggestions relating to this research. We also acknowledge help comments on the CBS data and on our modelling methods, as they evaluated, from Jon Bartos, Richard Burkel, and Tim Mouzakis.

References

Berens, A. P. (1996), "NDE Reliability Data Analysis," *Metal Handbook*, Vol. 17, Edition 9th, 689-701.

Congdon, P. (2003), *Applied Bayesian Modelling*, New York: John Wiley & Sons.

Feiveson, A. H. and Kulkarni, P. M. (2000), "Reliability of Space-Shuttle Pressure Vessels With Random Batch Effects," *Technometrics*, Vol. 42, 332-344.

Hassan, W. (2002), "Evaluation of Multizone Inspection Variability at the Supply Base for 8 Inch-Diameter Ti-6A1-4V: A Round Robin Study," *Review of Progress in Quantitative NDE*, Vol. 22, 1870-1877.

Meeker, W. Q. and Escobar, L. A (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

MIL-HDBK-1823 (1999), *Non-Destructive Evaluation System Reliability Assessment*, Standardization Order Desk, Building 4D, 700 Roberts Avenue, Philadephia, PA 19111-5094.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992), *Variance Components*, New York: John Wiley & Sons.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D., (2003), *WinBugs Version 1.4 User Manual* (Cambridge MRC Biostatistics Unit, http://www.mrc-bsu. cam.ac.uk/bugs).

Gilks, W. (1992), Derivative-free adaptive rejection sampling for Gibbs sampling, *Bayesian Statistics 4*, (J M Bernardo, J O Berger, A P Dawid, and A F M Smith, eds), Oxford University Press, UK, pp. 641-665.

Neal, R. (1997), Markov chain Monte Carlo methods based on 'slicing' the density function, *Technical Report 9722*, Department of Statistics, University of Toronto, Canada: http://www.cs.utoronto.ca/ radford/publications.html

Gelfand, A. E., and Smith, A. F. M (1990), Sampling-based approach to calculating marginal densities, *Journal of the American Statistical Association*, Vol. 85, 398-409.

Gelfand, A. E., Smith, A. F. M, and Lee, T. M. (1992), Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling, *Journal of the American Statistical Association*, Vol. 87, 523-535.

CHAPTER 5. CONCLUSIONS

In this thesis, we studied three POD-related topics: POD assessment methods for bivariate responses allowing for data truncation and censoring; POD assessment methods for adjusting for bias due to flaw sizing errors; statistical models to identify and quantify the variance components in NDE operations. Our research developed a more complete understanding of these subjects and provided useful tools to estimate POD for advanced NDE operations.

Chapter 1 extended standard univariate $\hat{a}$ versus a model to bivariate responses in the presence of data censoring and truncation. We used the extended $\hat{a}$ versus a method to analyze multizone inspection data and compute the POD curve estimate using a dual detection criterion. Motivated by the need to analyze conventional data with a large number of misses, we developed a model with accommodation terms that will allow for atypical misses. POD was then computed and compared with the POD that could be achieved if the cause of the atypical misses could be eliminated.

In Chapter 2, two measurement error models, the Burkel measurement error model and the geometrical measurement error model, were developed. These two models are capable of correcting for potential flaw-sizing bias in POD computations. The systematic simulations in this chapter provided insights of how measurement error affects regression coefficients and demonstrated that the a generalized errors-in-variables (GEV) method can correct the regression coefficients that are used to estimate POD. We illustrated the GEV methods with simulated inspection data based on the actual experimental NDE inspection data.

In Chapter 3, we developed the Bayesian hierarchial model to identify and quantify the variance components of inspection in the presence of data censoring. The Bayesian approach was demonstrated with both simulated data and experimental data, using MCMC simulation in the Winbugs software program. The estimates of variance components from the experimental

data suggested that flaw-to-flaw has the most important influence on inspection variability.

APPENDIX A. SUMMARY OF CBS DATA

Table A.1 1995 CBS Multizone Inspection Data.

FlawID	Amplitude	SNR	Flaw area	FlawID	Amplitude	SNR	Flaw area
B1AW1A	190	12.24	4886	B1AW1B	201	4.36	12502
B1AW1C	83	5.47	56464	B1AW1D	73	4.65	1760
B1AW1E	127	4.47	7033	B1AW2A	134	13.59	5872
B1AW2B	225	21.7	7784	B1AW2C	213	16.29	16317
B1AW2D	134	5.46	17526	B1AW2E	160	4.29	137240
B1AW2F	268	12.6	41767	B1AW2X	94	5.05	17850
B1AW3A	142	5.02	16590	B1AW3B	69	2.05	2160
B1AW3C	253	11.79	34576	B1AW3D	101	4.49	5792
B1AW3E	47	2.52	4242	B1AW3Y	99	5.05	0
B1AW3Z	77	3.48	47059	B1BW1A	63	2.96	8530
B1BW1B	113	5.67	8530	B1BW1C	142	8.09	28669
B1BW1D	56	2.79	0	B1BW1E	69	3.05	3791
B1BW1F	450	26.6	60656	B1BW1G	213	8.28	46440
B1BW1H	127	6.56	46440	B1BW2BA	179	13.27	155064
B1BW2BB	160	15.74	28478	B1BW2BC	113	7.55	71400
B1BW2BD	160	11.38	79111	B1BW2BE	142	8.38	87218
B1BW3A	113	3.92	3873	B1BW3B	160	7.56	3873
B1BW3C	127	6.67	102335	B1BW3D	74	5.06	26142
B2W1A	142	11.35	729824	B2W1B	151	11.05	87100
B2W1C	127	6.67	29191	B2W1D	87	6.46	8684
B2W2A			0	B2W2B	107	6.55	388179
B2W2C	74	4.09	17808	B2W2D	113	4.46	139744
B2W2X	61	4.97	54595	B2W3A	107	6.81	3170
B2W3B	72	3.54	793	B2W3C	50	2.59	9706
B3W1BA	84	4.8	155563	B3W1BB	73	2.45	7145
B3W1BC	99	3.52	124022	B3W1BD	134	9.52	96051
B3W2A	357	12.25	914604	B3W2B	113	8.39	557583
B3W2C	201	8.32	638598	B3W2D	127	5.06	222417
B3W2E	213	12.53	10987	B3W2F	160	6.83	8407
B3W2G	253	5.88	20764	B3W2H	0	7.87	20764
B3W3A	113	7.65	103930	B3W3B	120	5.74	28405
B3W3C	127	8.99	268969	B3W3D	142	8.93	154042

Table A.2 1994 CBS Conventional Inspection Data.

Flaw ID	Amp Normal	Status Normal	Amp Angle	Status Angle	Flaw area	Flaw ID	Amp Normal	Status Normal	Amp Angle	Status Angle	Flaw area
B1AW1A	30	Left	30	Left	4886	B1AW1B	45	Exact	70	Exact	12502
B1AW1C	30	Left	30	Left	56464	B1AW1D	30	Left	30	Left	1760
B1AW1E	70	Exact	100	Right	7033	B1AW2A	30	Left	30	Left	5872
B1AW2B	30	Left	30	Left	7784	B1AW2C	80	Exact	80	Exact	16317
B1AW2D	30	Left	30	Left	17526	B1AW2E	90	Exact	90	Exact	137240
B1AW2F	80	Exact	30	Left	41767	B1AW2X	30	Left	80	Exact	17850
B1AW3A	100	Right	90	Exact	16590	B1AW3B	30	Left	30	Left	2160
B1AW3C	100	Right	80	Exact	34576	B1AW3D	70	Exact	30	Left	5792
B1AW3E	30	Left	30	Left	4242	B1AW3Y	30	Left	30	Left	0
B1AW3Z	30	Left	30	Left	47059	B1BW1A	30	Left	30	Left	8530
B1BW1B	70	Exact	30	Exact	8530	B1BW1C	70	Exact	100	Right	28669
B1BW1D	30	Left	30	Left	0	B1BW1E	30	Left	100	Right	3791
B1BW1F	100	Right	70	Exact	60656	B1BW1G	100	Right	90	Exact	46440
B1BW1H	30	Left	30	Left	46440	B1BW2BA	90	Exact	90	Exact	155064
B1BW2BB	30	Left	30	Left	28478	B1BW2BC	30	Left	30	Left	71400
B1BW2BD	90	Exact	90	Exact	79111	B1BW2BE	90	Exact	90	Exact	87218
B1BW3A	30	Left	30	Left	3873	B1BW3B	30	Left	30	Left	3873
B1BW3C	30	Left	30	Left	102335	B1BW3D	30	Left	30	Left	26142
B2W1A	80	Exact	80	Exact	729824	B2W1B	100	Exact	90	Exact	87100
B2W1C	80	Exact	80	Exact	29191	B2W1D	30	Left	30	Left	8684
B2W2A		Left		Left	0	B2W2B	30	Left	30	Left	388179
B2W2C	30	Left	30	Left	17808	B2W2D	65	Exact	80	Exact	139744
B2W2X	30	Left	30	Left	54595	B2W3A	30	Left	30	Left	3170
B2W3B	30	Left	30	Left	793	B2W3C	40	Exact	60	Exact	9706
B3W1BA	70	Exact	40	Exact	155563	B3W1BB	30	Left	30	Left	7145
B3W1BC	100	Right	30	Left	124022	B3W1BD	100	Right	30	Left	96051
B3W2A	30	Left	100	Right	914604	B3W2B	100	Right	30	Left	557583
B3W2C	30	Left	30	Left	638598	B3W2D	80	Exact	80	Exact	222417
B3W2E	80	Exact	80	Exact	10987	B3W2F	30	Left	30	Left	8407
B3W2G	80	Exact	80	Exact	20764	B3W2H	30	Left	30	Left	20764
B3W3A	50	Exact	90	Exact	103930	B3W3B	50	Exact	40	Exact	28405
B3W3C	100	Right	30	Left	268969	B3W3D	100	Right	60	Exact	154042

APPENDIX B. WINBUGS PROGRAM FOR NDE VARIANCE COMPONENT ANALYSIS AND PART OF INSPECTION DATA

WINBUGS PROGRAM

```
model VCAPROGRAM; {

#####################################################################
################ Generation of prior values#########################
################ Diffuse distributions are selected ################
#####################################################################
###### Prior distribution for the reparameterized intercept#######

beta0star ~ dnorm(0.0,1.0E-6)

###### Prior distribution for flaw size fixed effect        ######

beta1     ~ dnorm(0.0,1.0E-6)

###### Prior distribution for the residual precision        ######
###### (reciprocal variance)                                ######

taub~ dgamma(0.0001,0.0001)

######  The residual standard deviation                     ######

sigma<- sqrt(1.0/taub)

###### Prior distribution for the the random effect precisions ###
###### (reciprocal variances) ######

taubFlawMorph ~ dgamma(0.0001,0.0001)
```

```
taubRun ~ dgamma(0.0001,0.0001)

taubSetup ~ dgamma(0.0001,0.0001)

######  Standard deviations for the random effects          ######

sigmaFlawMorph<- sqrt(1.0/taubFlawMorph)

sigmaRun<- sqrt(1.0/taubRun)

sigmaSetup<- sqrt(1.0/taubSetup)

######  reparameterization due to centering the data        ######

beta0 <- beta0star + beta1 *(- x.bar)

#########################################################################
###############  factors: ###############################################
###############  unknown parameters #####################################
#########################################################################

######  Site fixed effect                                   ######
        ######  S is the number of sites (3)
for(i in 1: S - 1) {
    Site[i] ~ dnorm(0,1.0E-6)
}
        ###### Last site effect (Baseline)
```

```
    Site[S] <- 0


######   Run random effect and              ######################
####     specification that Runs are nested in Site         ######
######   R is the number of runs

for ( k in 1: R) {
    Run[k] ~ dnorm(0.0, taubRun)
}

######  Setup random effect and Setup are nested in Run     ######
        ######  SN is the number of setup

for(k in 1: SN) {
    Setup[k] ~ dnorm(0.0,taubSetup)
}



######  Flaw-to-flaw random effect                         ######
        ######  F is the number of flaws

for(i in 1:F) {
    FlawMorph[i] ~ dnorm(0.0,taubFlawMorph)
}

######  Mean for centering the data                        ######
```

```
    x.bar <- mean(flawsize[])




#####################################################################
################ Model#################################################
################ This loop reads the data and defines that##########
################ likelihood by model specification     ############
#####################################################################

######  Model Specification                                    ######

######  N is  the number of rows in the data                   ######

for(j in 1:N ) {
        # Specify censoring and model
   resp[j] ~ dnorm(mu.reg[j], taub)I(cenR[j],cenL[j])

# This is a function for mu linking to model.

mu.reg[j]<-  beta0star + beta1  * (flawsize[j]-x.bar +
                            FlawMorph[flawid[j]])+
             Site[Siteid[j]] + Run[Runid[j]] + Setup[Setupid[j]]

 }

######  Model end                                              ######
 }
```

Table B.1 Part of Variability Study Inspection Data.

FlawID	Amplitude	CenL	CenU	FlawSize	Site	Run	Setup
1	43			3	A	1	1
2	60			5	A	1	1
3	NA	100		5	A	1	2
4	NA		30	3	A	1	5
1	40			3	B	5	21
2	56			5	B	5	21
3	NA	100		5	B	5	22
4	55			3	B	5	25