Statistical methods for extreme values and degradation analysis

Shiyao Liu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Statistics and Probability Commons

Recommended Citation
Liu, Shiyao, "Statistical methods for extreme values and degradation analysis" (2013). Graduate Theses and Dissertations. 13351.
https://lib.dr.iastate.edu/etd/13351

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Statistical methods for extreme values and degradation analysis

by

Shiyao Liu

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, Major Professor
Arka P. Ghosh
Mervyn G. Marasinghe
Alyson G. Wilson
Huaiqing Wu

Iowa State University
Ames, Iowa
2013

Copyright © Shiyao Liu, 2013. All rights reserved.
DEDICATION

To my parents
# TABLE OF CONTENTS

**LIST OF TABLES** ................................. vii

**LIST OF FIGURES** ............................... viii

**ACKNOWLEDGEMENTS** ........................... xiv

**ABSTRACT** .................................... xvi

**CHAPTER 1. GENERAL INTRODUCTION** ........... 1

1.1 Background .................................... 1

1.2 Motivation .................................... 1

1.2.1 Estimating the Minimum Thickness Along a Pipeline ......... 1

1.2.2 Degradation Model to Assess Pipeline Life .................... 2

1.2.3 Unbounded Likelihood Problem in Maximum Likelihood Estimation .. 3

1.3 Dissertation Organization .......................... 3

**CHAPTER 2. STATISTICAL METHODS FOR ESTIMATING THE MINIMUM THICKNESS ALONG A PIPELINE** .............. 5

2.1 Introduction ................................... 6

2.1.1 Motivation and Purpose .......................... 6

2.1.2 Pipeline Data ................................ 7

2.1.3 Some Previous Work on Extreme Value Analysis .................. 9

2.1.4 Overview .................................. 9

2.2 The Traditional Statistical Method to Estimate a Minimum ........... 10

2.2.1 Methods for Estimating the Distribution of a Minimum ........... 10

2.2.2 Application of the Distribution of a Minimum ................. 10

2.3 Methods for Estimating a Minimum Based on Extreme Value Theory .... 13
2.3.1 Extreme Value Distributions ........................................... 13
2.3.2 The Block Minima Method ............................................. 14
2.3.3 Estimating the Minimum over a Population .......................... 14
2.4 Design of the Simulation Experiment and Simulation Details .......... 15
  2.4.1 Objective of the Simulation .......................................... 15
  2.4.2 Experimental Factors and Their Levels .............................. 16
  2.4.3 ML Estimation and Comparison Criteria ............................. 17
2.5 Simulation Results: Effect of Block Size on MSE ........................ 18
  2.5.1 Graphical Summary of MSE Results for the Normal Parent Distribution 18
  2.5.2 Explanation of MSE Results for the Normal Parent Distribution .. 20
  2.5.3 The Effect of More Extreme Extrapolation ........................ 22
  2.5.4 MSE Results for the LEV and SEV Parent Distributions ............ 23
  2.5.5 Impact of the Shape Parameter on the GEV Quantile Estimates .... 25
2.6 Simulation Results: Effect of Block Size and Parent Distribution on Confidence Intervals ..................................................... 26
  2.6.1 Simulation Results on Profile Likelihoods for the Normal Parent Distribution ....................................................... 27
  2.6.2 Simulation Results on Profile Likelihoods for the SEV Parent Distribution ....................................................... 29
  2.6.3 General Conclusion from the Profile Likelihood Simulations ........ 29
2.7 Estimation of the Minimum Thickness in the Pipeline .................... 30
2.8 Conclusions, Recommendations, and Areas for Further Research .......... 31
2.9 Acknowledgments ............................................................. 35

CHAPTER 3. USING DEGRADATION MODELS TO ASSESS PIPELINE LIFE ................................................................. 41
3.1 Introduction ................................................................. 42
  3.1.1 Motivation and Purpose ............................................... 42
  3.1.2 Pipeline Data .......................................................... 42
  3.1.3 Related Work .......................................................... 43
  3.1.4 Overview ............................................................ 44
3.2 Modeling Pipeline Data from Circuit G in Facility 3 .......................... 45
  3.2.1 Degradation Model for Pipeline Data from Circuit G in Facility 3 .... 45
  3.2.2 Bayesian Estimation of the Parameters in the Degradation Model .... 46
  3.2.3 Statistical Model for Different Quadrants ............................... 46
3.3 Models Relating Degradation and Pipeline Failure in Circuit G of Facility 3 48
  3.3.1 Bayesian Evaluation of the Failure Time Distribution ................. 48
  3.3.2 Prediction of the Remaining Life of the Current Circuit ............. 51
3.4 Modeling Pipeline Data from Circuit Q in Facility 1 ........................ 53
  3.4.1 Degradation Model for Corrosion Initiation and Growth ................ 53
  3.4.2 Bayesian Estimation of the Parameters in the Degradation Model .... 54
3.5 Models Relating Degradation and Failure of Pipeline Data from Circuit Q in Facility 1 .......................................................... 57
  3.5.1 Bayesian Evaluation of the Failure Time Distribution ................ 57
  3.5.2 Prediction of the Remaining Life of the Current Circuit ............. 59
3.6 Effect of Additional Inspections on Identifiability ........................ 61
3.7 Hierarchical Model with Corrosion Initiation and Growth Along a Pipeline 63
3.8 Concluding Remarks and Areas for Future Research ........................ 64
3.9 Appendix: OpenBUGS Code for the Hierarchical Model ..................... 66

CHAPTER 4. UNDERSTANDING AND ADDRESSING THE UNBOUNDED “LIKELIHOOD” PROBLEM .................................................. 78
4.1 Introduction ................................................................. 79
  4.1.1 Background ......................................................... 79
  4.1.2 An Illustrative Example ........................................... 79
  4.1.3 A Simple Remedy ................................................... 81
  4.1.4 R. A. Fisher’s Definition of Likelihood ........................... 81
  4.1.5 Related Literature .................................................. 82
  4.1.6 Overview ........................................................... 83
4.2 A Classification of Unbounded “Likelihoods” ............................. 83
4.2.1 Continuous Univariate Distributions with Three or Four Parameters, Including a Threshold Parameter .............................................. 84
4.2.2 Discrete Mixture Models Where at Least One Component Has Both a Location and a Scale Parameter ........................................... 87
4.2.3 Minimum-Type (and Maximum-Type) Models Where at Least One of the Marginal Distributions Has Both a Location and a Scale Parameter 90
4.3 The Effect of the Round-off Error on Estimation .......................... 92
  4.3.1 Background ........................................................................ 92
  4.3.2 Numerical Examples ............................................................. 93
4.4 A Sufficient Condition for Using the Density-Approximation Likelihood .... 95
4.5 Concluding Remarks ................................................................. 97
4.6 Appendix .................................................................................. 97
  4.6.1 Proof of Theorem 1 ............................................................... 97
  4.6.2 An Example Showing the Equicontinuity Condition Is Not Necessary . 98

CHAPTER 5. GENERAL CONCLUSIONS ............................................ 104
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>ML estimates and likelihood-based approximate 95% confidence intervals for $\hat{x}_{0.05}$.</td>
</tr>
<tr>
<td>3.1</td>
<td>Marginal posterior distribution summaries of the degradation model parameter estimates for pipeline data from Circuit G in Facility 3 using the degradation model (3.1).</td>
</tr>
<tr>
<td>3.2</td>
<td>Marginal posterior distribution summaries of the degradation model parameter estimates for pipeline data from Circuit G in Facility 3 using the degradation model (3.2).</td>
</tr>
<tr>
<td>3.3</td>
<td>Marginal posterior distribution summaries of the parameters in the degradation model with lognormal corrosion rate for pipeline data from Circuit Q in Facility 1.</td>
</tr>
<tr>
<td>3.4</td>
<td>Marginal posterior distribution summaries of the parameters in the degradation model with Weibull corrosion rate for pipeline data from Circuit Q in Facility 1.</td>
</tr>
<tr>
<td>3.5</td>
<td>Marginal posterior distribution summaries of the degradation model parameter estimates of the simulated pipeline data with 12 circuits and 33 TMLs within each circuit in Facility 1.</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

| Figure 2.1 | Number of metal-loss features recorded among the 5,649 one-foot segments that had one or more detected features. | 8 |
| Figure 2.2 | Probability plots for the $n = 200$ pipeline wall thickness inspection data with 95% simultaneous confidence bands. | 11 |
| Figure 2.3 | A comparison of the MSEs of the ML quantile estimators of the minimum for the normal parent distribution when the sample size is $n = 200$ and ratio is $M/n = 10$. | 19 |
| Figure 2.4 | A comparison of the MSEs of the ML quantile estimators of the minimum for the normal parent distribution when the sample size is $n = 1,000$ and ratio is $M/n = 10$. | 20 |
| Figure 2.5 | A comparison of the sampling distributions of $\hat{x}_{0.05}$ using the BL$_{\text{min}}$-Gumbel and BL$_{\text{min}}$-GEV methods for the normal parent for $n = 200$ and $M/n = 10$ and different combinations of $B$ and $m$. The horizontal lines indicate the position of the true 0.05 quantile of the distribution of the minimum in the population. Note that some BL$_{\text{min}}$-GEV estimates are off scale for $m = 10$ and 20. | 21 |
| Figure 2.6 | A comparison of the sampling distributions of $\hat{x}_{0.05}$ using the BL$_{\text{min}}$-Gumbel and BL$_{\text{min}}$-GEV methods for the normal parent when the sample size is $n = 200$ with different combinations of $M/n$, $B$ and $m$. The horizontal lines indicate the position of the true 0.05 quantile of the distribution of the minimum in the population. Note that some BL$_{\text{min}}$-GEV estimates are off scale for $m = 10$. | 22 |
Figure 2.7  A comparison of the MSEs of the ML quantile estimators of the minimum for the LEV parent distribution when the sample size is \( n = 200 \) and ratio is \( M/n = 10 \). ........................................ 24

Figure 2.8  A comparison of the MSEs of the ML quantile estimators of the minimum for the SEV parent distribution when the sample size is \( n = 200 \) and ratio is \( M/n = 10 \). ........................................ 25

Figure 2.9  Scatter plot of the shape parameter estimates \( \hat{\xi} \) versus the \( x_{0.05} \) quantile estimates for the BL_{min}-GEV method for normal, SEV and LEV parent distributions when sample size is \( n = 200 \) and ratio is \( M/n = 10 \). In part (a), the block size is \( B = 4 \) and the number of blocks is \( m = 50 \); In part (b), the block size is \( B = 10 \) and number of blocks is \( m = 20 \). . 26

Figure 2.10  Comparison of quantile relative likelihood profile plots using (a) BL_{min}-GEV with the block size \( B = 4 \) and (b) BL_{min}-Gumbel with the block size \( B = 10 \) when the sample size is \( n = 200 \), the population size is \( M = 5,649 \), quantile is \( x_{0.05} \) and the parent distribution is normal. The vertical line indicates the position of the true 0.05 quantile of the distribution of the minimum in the population. The horizontal lines allow visualization of corresponding approximate 50% and 95% likelihood-based confidence intervals. ........................................ 28

Figure 2.11  Comparison of quantile relative likelihood profile plots using (a) BL_{min}-GEV with the block size \( B = 4 \) and (b) BL_{min}-Gumbel with the block size \( B = 10 \) when the sample size is \( n = 200 \), the population size is \( M = 5,649 \), quantile is \( x_{0.05} \) and parent distribution is the SEV distribution. The vertical line indicates the position of the true 0.05 quantile of the distribution of the minimum in the population. The horizontal lines allow visualization of corresponding approximate 50% and 95% likelihood-based confidence intervals. ........................................ 29
Figure 2.12  Relative likelihood profile plots for the quantile $x_{0.05}$ under the BL$_{\text{min}}$-Gumbel and the BL$_{\text{min}}$-GEV methods for $n = 200$ and $n = 1,000$. The horizontal lines indicate corresponding approximate 50% and 95% confidence intervals.

Figure 2.13  Weibull probability plots of estimates of the parent distribution and distribution of minimum (DoM) for the pipeline wall thickness inspection data under the three different estimation methods for $n = 200$ (left) and $n = 1,000$ (right). The shorter vertical tick marks on the $p = 0.05$ line indicate likelihood-based 95% confidence interval for the $x_{0.05}$ quantile of the minimum distributions. The longer vertical tick mark on the $p = 0.05$ line indicates “true quantile” $x_{0.05}$ based on the 5,649 population thickness values.

Figure 3.1  Time plot for pipeline data from Circuit G in Facility 3.

Figure 3.2  Time plot for pipeline data from Circuit Q in Facility 1.

Figure 3.3  Time plot showing the fitted thickness values for the pipeline data from Circuit G in Facility 3 using the degradation model (3.1).

Figure 3.4  Time plot showing the fitted thickness values for different quadrants of pipeline data from Circuit G in Facility 3 using the degradation model (3.2).

Figure 3.5  Degradation model estimates of (a) failure time distribution (years since pipeline installation) and (b) remaining lifetime distribution (years since the last inspection $t_c$) with two-sided 95% and 80% credible intervals on the lognormal paper for pipeline data from quadrant 1 of Circuit G in Facility 3.

Figure 3.6  Posterior density of the 0.1, 0.2, 0.3 and 0.4 quantiles of the minimum remaining lifetime distribution (years since the last inspection time $t_c$: January 2003) with the population size $M = 100$ of pipeline data from Circuit G in Facility 3 using the degradation models (3.1) and (3.2).
| Figure 3.7 | Trellis plot for pipeline data from Circuit Q in Facility 1. | 53 |
| Figure 3.8 | Trellis plot of the fitted thickness values for Circuit Q in Facility 1 using the lognormal corrosion rate distribution. The dotted lines indicate extrapolation. | 58 |
| Figure 3.9 | Trellis plot of the fitted thickness values for Circuit Q in Facility 1 using the Weibull corrosion rate distribution. The dotted lines indicate extrapolation. | 59 |
| Figure 3.10 | Samples from the marginal posterior distribution of the lognormal corrosion rates for each TML in Circuit Q in Facility 1. | 60 |
| Figure 3.11 | Samples from the marginal posterior distribution of the corrosion initiation times for each TML in Circuit Q in Facility 1 using a lognormal distribution to describe corrosion rates. | 61 |
| Figure 3.12 | Posterior densities of the initiation times for each TML in Circuit Q in Facility 1. | 62 |
| Figure 3.13 | Degradation model estimates of failure time distributions for pipeline components from Circuit Q in Facility 1 comparing normal, lognormal and Weibull corrosion rate distributions in the degradation model (3.7). | 68 |
| Figure 3.14 | Degradation model estimates (the center lines) of failure time distributions for pipeline components from Circuit Q in Facility 1 with the lognormal corrosion rate distribution in the degradation model (3.7) and two-sided 95% and 80% credible intervals. | 69 |
| Figure 3.15 | Degradation model estimates of remaining lifetime distributions for pipeline components from Circuit Q in Facility 1 comparing normal, lognormal and Weibull corrosion rate distributions in the degradation model (3.7). | 70 |
| Figure 3.16 | Degradation model estimates (the center lines) of remaining lifetime distributions for pipeline components from Circuit Q in Facility 1 with the lognormal corrosion rate distribution in the degradation model (3.7) and two-sided 95% and 80% credible intervals. | 71 |
Figure 3.17 Posterior density of the 0.1, 0.2, 0.3, and 0.4 quantiles of the minimum remaining lifetime distribution (years after the last inspection time $t_c$: January 2004) with the population size $M = 100$ using the lognormal corrosion rate distribution of pipeline data from Circuit Q in Facility 1.

Figure 3.18 Time plot for the simulated pipeline data from a single circuit with 33 TMLs.

Figure 3.19 Trellis plot of the fitted thickness values for the simulated pipeline data in a single circuit with 33 TMLs using the diffuse prior distributions.

Figure 3.20 Degradation model estimates of failure time distributions for the simulated pipeline data in a single circuit with 33 TMLs using the lognormal corrosion rate distribution and diffuse priors in the degradation model (3.7) and two-sided 95% and 80% credible intervals.

Figure 3.21 Time plot for simulated pipeline data from Circuits 1 to 4 in the same Facility.

Figure 3.22 Box plots of the corrosion rates within each of the 12 circuits.

Figure 4.1 Three-parameter lognormal profile log-likelihood plots of the threshold parameter $\gamma$ for the diesel generator fan data using (a) the unbounded density-approximation likelihood $L$ and (b) the correct likelihood $\mathcal{L}$ (with the round-off error $\Delta=5$).

Figure 4.2 The minimum GEV log-likelihood plots of the location parameter $\mu$ for data with sample size $n = 20$ using (a) the density-approximation likelihood $L$ and (b) the correct likelihood $\mathcal{L}$.

Figure 4.3 Mixture distribution of two univariate normal profile log-likelihood plots of $\delta$ using (a) the unbounded density-approximation likelihood and (b) the correct likelihood (with $\Delta=0.05$).
Figure 4.4 An MTD corresponding to a minimum of two independent random variables having the Weibull and exponential distributions. The profile log-likelihood plots of $\sigma_A$ using (a) the density-approximation likelihood and (b) the correct likelihood (with $\Delta=0.05$). .......................... 91

Figure 4.5 Plots of the minimum GEV location parameter estimates $\hat{\mu}$ versus the number of digits after the decimal point $k$ with $\sigma$ equal to 0.3, 0.5, 1, and 2. The $\Delta$ values on the top of each plot correspond to the round-off errors. ................................................................. 94

Figure 4.6 Mixture distribution of two univariate normal profile log-likelihood plots of $\delta$ using (a) the unbounded density-approximation likelihood and (b) the correct likelihood (with $\Delta = 0.005$). ................................. 95
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who helped me through the PhD journey.

First, I would like to thank Department of Statistics at Iowa State University for giving me such an invaluable opportunity to explore the hidden treasures of data and helping me prepare to become a statistician.

My deepest gratitude goes to my advisor, Professor William Q. Meeker, for his continuous guidance, numerous help, enthusiastic encouragement and great patience. Dr. Meeker inspired me with the research ideas, the ways of doing research, and how to think with a statistical mind. “Every time, when we observe something unexpected, either we made mistakes, or we could learn something.” What I learned from Dr. Meeker is not only the spirit of discovering the unknown, but also the courage to confront mistake. With the help of Dr. Meeker, I realized the special power of mathematics and statistics in solving practical problems. Dr. Meeker has set a perfect example for me to follow in my career and life. I feel really fortunate to have him as my advisor.

I am also very grateful to my committee members Dr. Arka Ghosh, Dr. Mervyn G. Marasinghe, Dr. Alyson G. Wilson, and Dr. Huaiqing Wu for their precious help on my dissertation research. I learnt a lot from them. Dr. Ghosh showed me the elegance and rigor of measure theory in STAT 642, the course I enjoyed most at Iowa State. Dr. Marasinghe provided precious comments and insightful ideas for improving our papers. Dr. Wilson led me into the amazing world of Bayesian statistics and I enjoyed reading her Bayesian reliability book. In particular, I am indebted to Dr. Wu for his nice collaboration and help. The discussions with Dr. Wu greatly broadened my horizon in Statistics.

I want to take this opportunity to thank many faculty members at Iowa State University. They extended my knowledge and kindly supported me in many different ways. Special thanks
goes to Dr. Sarah Nusser, Dr. Mack Shelley, Dr. Karin Dorman and Dr. Ranjan Maitra from Department of Statistics and Dr. Annette O’Connor from Department of Veterinary Diagnostics and Production Animal Medicine.

Last but not least, I would like to thank my parents for their countless love, constant support and utmost encouragement, which always give me strength and faith no matter how far away I am from them. Nothing could make me more proud than being their daughter.
ABSTRACT

Energy and chemical companies use pipelines to transfer oil, gas and other materials from one place to another, within and between their plants. Pipeline integrity is an important concern because pipeline leakage could result in serious economic or environmental losses. In this dissertation, statistical models and methods motivated by real applications were developed for pipeline reliability using extreme value theory and degradation modeling.

In Chapter 2, interval-censored measurements from a given set of thickness measurement locations (TMLs) along a three-phase pipeline are used to estimate the distribution of the minimum thickness. The block-minima method based on extreme value theory provides a robust approach to estimate the minimum thickness in a pipeline. In the block-minima method using the Gumbel and the generalized extreme value distributions, the choice of the number of blocks involves the trade-off between bias and variance. We conduct a simulation study to compare the properties of different models for estimating minimum pipeline thickness and investigate the effect of block size choice on MSE in the block-minima method.

The pipeline thickness estimation in Chapter 2 is based on data from a single time point at each TML. In the other pipeline applications, longitudinal inspections of the pipeline thickness at particular locations along the pipeline are available. Depending on different mechanisms of corrosion processes, we have observed various types of general degradation paths. In Chapter 3 of this thesis, we propose a degradation model describing corrosion initiation and growth behavior. The parameters in the degradation model are estimated using a Bayesian approach. We derive failure-time and remaining lifetime distributions from the degradation model and compute Bayesian estimates and credible intervals of the failure-time and remaining lifetime distribution. We also develop a hierarchical model to quantify the pipeline corrosion rate for similar circuits within a single facility.

The extreme value theorem suggests that no matter what the underlying parent distribu-
tion is, the limiting distribution of minima is the minimum generalized extreme value (GEV) distribution. The likelihood function, as it is usually written as a product of density functions, however, is unbounded in the parameter space. Due to rounding, all data are discrete and the use of densities for “exact” observations is only an approximation. In Chapter 4 of the thesis, we use the “correct likelihood” based on interval censoring to eliminate the problem of an unbounded density-approximation likelihood. We categorize the models that have an unbounded density-approximation likelihood into three groups, which are (1) continuous univariate distributions with both a location and a scale parameter, plus a threshold parameter, (2) discrete mixture models of continuous distributions for which at least one component has both a location and a scale parameter, (3) minimum-type (and maximum-type) models for which at least one of the marginal distributions has both a location and a scale parameter. For each category, we illustrate the density breakdown with specific examples. We also study the effect of the round-off error on estimation using the correct likelihood, and provide a sufficient condition for the joint density to provide the same maximum likelihood estimate as the correct likelihood, as the round-off error goes to zero.
CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

Energy and chemical companies use pipelines to transport material (e.g., in a three-phase pipeline that is a mixture of oil, gas and water) from one location to another location. Pipeline integrity is an important concern for these companies. If leakage occurs within a pipeline circuit, it could result in serious economic loss, personal injury, or damage to the environment. The main objective of this research is to develop statistical methods to estimate the minimum thickness of the pipeline circuit, develop a degradation model to describe the pipeline corrosion behavior for the longitudinal pipeline inspection data to predict the life time of the pipeline, and to investigate an unbounded likelihood problem that arises, for example, in the distribution of extremes.

1.2 Motivation

The research is motivated by applications from both energy and chemical companies regarding establishing statistical models to estimate pipeline integrity and reliability. The detailed motivation for each study is given below.

1.2.1 Estimating the Minimum Thickness Along a Pipeline

Due to the severe consequences of any leakage, the integrity of the pipeline has always been one of the most important concerns for energy and chemical companies. If the thickness of a pipeline at particular location is below a certain threshold there is a high risk of pipeline leakage. In real pipeline applications, there is usually limited information regarding the distribution of the measured pipeline thickness. Using a misspecified probability distribution to estimate the
distribution of the minimum thickness could cause substantial bias. The block-minima method based on the extreme value theory (Fisher and Tippett 1928) provides an alternative approach to estimate the minimum thickness, especially when the underlying corrosion distribution is unknown and a sufficient amount of data are available.

In the block-minima method, the choice of appropriate block sizes to construct the block-minima data set for fitting either the minimum generalized extreme value distribution (GEV) or one of three limiting distributions of minima involves the trade-off between the bias and variance of the small quantile estimators. There are many factors that need to be considered; for example, the underlying parent distribution of the pipeline thickness, the ratio of the population size to the sample size, and the different extreme value distributions. We conducted a simulation study to compare the properties of different models for estimating minimum pipeline thickness and to investigate the effect of using different size blocks. We illustrate the methods using pipeline inspection data from a three-phase pipeline.

1.2.2 Degradation Model to Assess Pipeline Life

In the first application of estimating the minimum thickness along a pipeline, there is single thickness measurement for each location. Longitudinal inspections of the pipeline thickness at particular locations along the pipeline provide useful information to assess the lifetime of the pipeline when compares to lifetime data. The degradation process data usually provides more information about the corrosion processes and the lifetime of the pipeline. We have observed various types of general degradation paths from pipeline data. In one application, we use a degradation model describing the corrosion initiation and growth behavior in the pipeline assuming that for each location, before the corrosion has been initiated, there is no thickness loss. Our model assumes a constant corrosion rate at each location and linear degradation path after the corrosion initiation. Under the assumption that both the corrosion initiation time and the corrosion rate after initiation are positive, we estimate the parameters in the proposed degradation model using a Bayesian approach. We derive the time to failure and remaining lifetime distributions from the degradation model to predict the lifetime of the pipeline circuit and also estimate the small quantiles of the remaining lifetime distribution within the Bayesian frame-
work. For another data set, we use a degradation model for the longitudinal pipeline thickness measurements with no initiation time but different corrosion rates at different locations.

1.2.3 Unbounded Likelihood Problem in Maximum Likelihood Estimation

The performance of ML estimators for quantiles based on fitting a general extreme value distribution can be poor unless there is a large number of observations (or blocks if the block-minimum method is being used)—say, greater than 50. The density approximation likelihood function for the GEV distribution is unbounded in the parameter space. The unbounded behavior of the likelihood function can result in a breakdown of the maximum likelihood (ML) estimation. The unbounded likelihood problem, which causes both numerical and statistical problems in ML estimation, arises in a number of other statistical models. We look at a variety of unbounded likelihood problems and classify them into three categories. Instead of using the density-approximation likelihood, following the suggestion of Kempthorne and Folks (1971), one can eliminate the unbounded likelihood problem by using the correct likelihood based on small intervals that result from round off (e.g., implied by the data’s precision). In the application of the estimation of the minimum thickness of a pipeline, because the pipeline measurements are subject to round-off error, we treat the thickness measurements as interval-censored observations. We also explore the effect that the round-off error has on estimation with the correct likelihood and provide a sufficient condition for the density approximation and the correct likelihoods to give the same maximum likelihood estimates.

1.3 Dissertation Organization

This dissertation consists of three main chapters, preceded by this general introduction and followed by a general conclusion. Each of these main chapters corresponds to paper that is to be submitted to a journal. Chapter 2 develops methods to estimate the thinnest wall thickness along a pipeline circuit. Chapter 3 uses the degradation model to estimate the lifetime of a pipeline circuit. Chapter 4 studies the models leading to unbounded likelihoods and investigates the effect of the round-off error on estimation by using the “correct likelihood”.
Bibliography


CHAPTER 2. STATISTICAL METHODS FOR ESTIMATING THE MINIMUM THICKNESS ALONG A PIPELINE

A paper to be submitted
Shiyao Liu and William Q. Meeker
Department of Statistics
Iowa State University

Abstract

Pipeline integrity is important because leaks can result in serious economic or environmental losses. Inspection information from a sample of locations along the pipeline can be used to estimate corrosion levels. The traditional parametric model method for this problem is to estimate parameters of a specified corrosion distribution and then to use these parameters to estimate the minimum thickness in a pipeline. Inferences using this method are, however, highly sensitive to the distributional assumption. Extreme value modeling provides a more robust method of estimation if a sufficient amount of data is available. For example, the block-minima method produces a more robust method to estimate the minimum thickness in a pipeline. To use the block-minima method, however, one must carefully choose the size of the blocks to be used in the analysis. In this paper we use simulation to compare the properties of different models for estimating minimum pipeline thickness, investigate the effect of using different size blocks, and illustrate the methods using pipeline inspection data.

Key Words: Block minima; extreme value; maximum likelihood; simulation.
2.1 Introduction

2.1.1 Motivation and Purpose

Energy companies use pipelines to transfer oil, gas and other materials from one place to another. Manufacturers of chemical products use pipelines within and between their plants. When the thickness at a location falls below a fixed threshold, there is risk of leakage that could result in serious economic loss, personal injury, or damage to the environment. It is possible to use statistical methods to estimate the minimum thickness of a pipeline. The traditional parametric statistical method of modeling the minimum is to estimate the distribution of thickness from the measured pipeline thickness data and then calculate the corresponding probability distribution of the minimum thickness. However, we usually have uncertain knowledge about the particular parent distribution that appropriately describes the data generating process. Small discrepancies in the specified parent distribution can lead to substantial bias in estimating the minimum distribution.

Extreme value theory, originating with Fisher and Tippett (1928), serves as an alternative approach to model extrema. Instead of estimating the parent distribution from observations, we accept the fact that the parent distribution is unknown. An immediate consequence of the Extreme Value Theorem is that under mild conditions, the limiting distribution of properly standardized minima (or maxima) extreme values has a generalized extreme value distribution. This distribution includes three classes of extreme value distributions as special cases, and these are called the Gumbel, Fréchet, and Weibull distributions respectively. The choice among these three distributions depends on the domain of attraction of the relevant tail of the parent distribution. Here we explore the use of extreme value distributions to model minimum pipeline thickness. We use simulation to investigate the alternative procedures for estimating a minimum and apply the methods to inspection data from a three-phase pipeline (i.e., a pipeline carrying a mixture of oil, gas, and water). Our results show that whether one fits a generalized extreme value distribution or one of the special extreme value distributions under an assumed domain of attraction has a large effect on the choice of block size.
2.1.2 Pipeline Data

In some pipeline integrity applications it is possible to do in-line pipeline inspection (ILI) by using a “smart-pig” utilizing magnetic flux leakage or ultrasonic testing technology to detect and measure corrosion and other metal-loss features in a pipeline. The smart-pig is pulled through the pipeline acquiring information with high spatial resolution (e.g., 3mm).

In most pipeline applications, however, such in-line inspections are impossible and the pipeline operators must rely on external inspections that are done at a set of sample locations, known as Thickness Measurement Locations (TMLs). The resulting sample data are then used to make inferences about the integrity of the entire pipeline. Ultrasonic and radiographic (X-ray) testing are the most commonly-used external inspection methods to measure pipeline thickness.

To illustrate the application of the different methods that one can use to estimate the distribution of a minimum in applications like pipeline integrity, we use data from a three-phase pipeline that had an original thickness of 0.375 inches. To protect sensitive information, the name of company that provided the data and the location of the pipeline cannot be disclosed.

The raw pipeline data that we received were ILI pipeline inspection data giving the location (in feet, measured from one end of the pipeline), size, and depth of observed metal-loss features. We partitioned the data into features (a feature is an indication of metal loss) observed in the 32,272 one-foot segments along the pipeline. The smart-pig identified features in 5,649 of these one-foot segments. In the other 26,623 segments, there was no detectable metal loss. Within the 5,649 locations with detectable metal loss, the number of features that were recorded ranges between 1 and 27, as shown in Figure 2.1. For each one-foot segment with observed metal loss, we take the minimum of all of the thickness measurements as the wall thickness response within that one-foot segment.

Analysis of the data suggests, in agreement with knowledge that we have gained from experts in pipeline industry, that metal loss tends to concentrate in certain areas of the pipeline. These points of concentration tend to be in the area of certain physical characteristics of the pipeline such as near supports, places where the slope of the pipeline changes, and near welds that join
Figure 2.1  Number of metal-loss features recorded among the 5,649 one-foot segments that had one or more detected features.

two sections of pipe. Thus we will take as the population the 5,649 locations with detectable metal loss. To simulate the common kind of external inspection, we draw simple random samples of size $n = 200$ (approximately the smallest sample size required to use the more robust statistical methods based on extreme-value theory) and $n = 1,000$ (approximately the largest sample size we have seen in external inspection pipeline applications) of the one-foot segments with metal loss. Such selective sampling would correspond, roughly, to the industry practice of over-sampling at locations in a pipeline where one would expect to see higher rates of metal loss.

Because of measurement resolution limitations of the smart-pig system, (resulting from an analog-to-digital conversion with a limited number of bits per reading to allow storage of large
amounts of data) the pipeline wall measurements are not known exactly due to round-off error. As suggested by Vardeman and Lee (2005), we therefore treated the data as interval-censored observations.

2.1.3 Some Previous Work on Extreme Value Analysis

Extreme value analysis has been used widely in many areas of application ranging from insurance and finance to meteorology and hydrology. There is a large number of books and articles regarding both the mathematical theory and applications of extreme value analysis. Gumbel (1958) is one of the earliest books and is still an important reference in extreme value analysis. Coles (2001) describes the common approaches of extreme value analysis including the block maxima (or minima) method and the threshold excess models. Castillo, Hadi, Balakrishnan and Sarabia (2005) focus particularly on applications of extreme value analysis in the engineering areas. Engeland, Hisdal and Frigessi (2004) use extreme value methods to model hydrological floods and droughts. Kowaka et al. (1994) use extreme value statistical methods to investigate corrosion phenomena. Laycock, Cottis and Scarf (1990), Laycock and Scarf (1993) and Scarf and Laycock (1994) apply the extreme value analysis to corrosion and propose a four parameter time-dependent model to extrapolate of extreme pit depths into future exposure time and larger area of metal. Shibata (1994) reviews the application of the extreme value statistics to corrosion using several examples. Scarf and Laycock (1996) use extreme value theory to model the maximum penetration caused by pitting corrosion on metal surfaces. Fougères, Holm and Rootzén (2006) design and analyze experiments to compare treatments with extreme responses, using corrosion experiments to illustrate their approach.

2.1.4 Overview

The remainder of this paper is organized as follows. Section 2.2 briefly describes the traditional statistical method of using a parametric distribution to model the minimum directly and illustrates this method with a pipeline thickness example. Section 2.3 introduces the extreme value distributions, presents the block minima method, and shows how to apply the block minima method with the Gumbel and the generalized extreme value distributions to estimate
quantiles of the distribution of a minimum over the population. Section 2.4 gives the details of the design of a simulation experiment for comparing the different methods of estimating a minimum. Section 2.5 presents the simulation results and investigates the effect of block size choice on MSE in the block minima method. Section 2.6 explores the effect of block size choice on the interval estimates by comparing the relative likelihood profile plots for quantiles of the distribution of a minimum. Section 2.7 compares different methods of estimating quantiles of the distribution of a minimum and the corresponding confidence intervals with the pipeline thickness inspection data. Section 2.8 provides some summary conclusions and recommendations and suggests some areas for future research.

2.2 The Traditional Statistical Method to Estimate a Minimum

2.2.1 Methods for Estimating the Distribution of a Minimum

For independent and identically distributed (iid) random variables $X_1, X_2, \ldots, X_M$ with a cumulative distribution function (cdf) $F(x; \theta)$, the distribution of the minimum $Y_M = \min\{X_1, X_2, \ldots, X_M\}$ can be expressed as:

$$
\Pr[Y_M \leq x_c] = 1 - [1 - F(x_c; \theta)]^M
$$

(2.1)

where $M$ is the population size and $\theta$ is parameter vector. From (2.1), in order to estimate the minimum distribution, one needs first to specify the parent distribution $F(x; \theta)$. Then substituting an estimate of $\theta$ provides an estimate of the distribution of the minimum.

2.2.2 Application of the Distribution of a Minimum

In this section, we use the pipeline wall thickness inspection data to illustrate the application of the traditional statistical method to model a minimum. In order to estimate the probability that the minimum wall thickness of the population is less than a critical limit (say, 0.10 inches in this application), one needs first to choose an appropriate parent distribution to fit the wall thickness data from the $n = 200$ locations. Because inferences on the minimum thickness generally require extrapolation into the lower tail of the distribution, the inferences can be
Figure 2.2  Probability plots for the $n = 200$ pipeline wall thickness inspection data with 95% simultaneous confidence bands.

... highly sensitive to the assumed distribution. When the data are consistent with more than one distribution, it is important to do sensitivity analysis to assess the effect that different distributional assumptions will have on the final answers.

Probability plots (described, for example, in Chapter 6 of Meeker and Escobar 1998) provide a useful graphical method for assessing the adequacy of an underlying parent distribution. Figure 2.2 provides the normal, lognormal, smallest extreme value (SEV) and Weibull probability plots for the $n = 200$ pipeline thickness observation. These probability plots indicate a good fit for the Weibull distribution, although the lognormal distribution is also consistent with...
the data and provides a reasonable description, at least in the lower tail of the distribution. Therefore, we choose Weibull and lognormal distributions as candidate parent distributions to make comparative statements about the probability that the minimum thickness is less than a specified value. The Weibull and lognormal distribution cdfs are

\[ F(x; \mu, \sigma) = \Phi_{\text{sev}} \left( \frac{\log(x) - \mu}{\sigma} \right) \quad \text{and} \quad F(x; \mu, \sigma) = \Phi_{\text{nor}} \left( \frac{\log(x) - \mu}{\sigma} \right). \]

Here \( \Phi_{\text{sev}}(z) = 1 - \exp[-\exp(z)] \) is the standardized \((\mu = 0, \sigma = 1)\) smallest extreme value (or SEV) cdf and \( \Phi_{\text{nor}}(\cdot) \) is the standard \((\mu = 0, \sigma = 1)\) normal cdf. Although one might question the assumption of independence in this application to estimate the distribution of a minimum, if the dependence is positive, as might be expected, the value given by the method is conservative (e.g., Chapter 2 of Barlow and Proschan 1975).

Here, we illustrate the traditional method to estimate a minimum (i.e., the minimum method). For the Weibull distribution, the ML estimate of the probability that the minimum pipeline wall thickness out of the \( M = 5,649 \) one-foot segments (i.e., the population) is less than 0.10 would be:

\[
\Pr[\min(X_1, X_2, \ldots, X_{5649}) \leq 0.10] = 1 - [1 - \Pr(X_1 \leq 0.10)]^{5649} \\
= 1 - \left[ 1 - \Phi_{\text{sev}} \left( \frac{\log(0.10) + 1.157}{0.089} \right) \right]^{5649} \\
= 0.0144. \quad (2.2)
\]

Here \( \hat{\mu} = -1.157 \) and \( \hat{\sigma} = 0.089 \) are respectively the maximum likelihood estimates of the SEV parameters based on the logarithm of the wall-thickness measurements from a simple random sample of 200 randomly chosen one-foot segments out of the 5,649 one-foot segments in the pipeline. Similarly, for the lognormal distribution,

\[
\Pr[\min(X_1, X_2, \ldots, X_{5649}) \leq 0.10] = 1 - \left[ 1 - \Phi_{\text{nor}} \left( \frac{\log(0.10) + 1.2089}{0.1148} \right) \right]^{5649} \approx 0, \quad (2.3)
\]
where \( \hat{\mu} = -1.2089 \) and \( \hat{\sigma} = 0.1148 \) are respectively the maximum likelihood estimates of the mean and standard deviation of log thickness based on the same sample of \( n = 200 \) out of the \( M = 5,649 \) pipeline wall thickness measurements.

As expected, the Weibull distribution is more conservative than the lognormal distribution in terms of estimating the probability that the minimum thickness is less than a critical limit.

### 2.3 Methods for Estimating a Minimum Based on Extreme Value Theory

#### 2.3.1 Extreme Value Distributions

As mentioned in Section 2.2, because the parent distribution function \( F \) is not always known and inferences on the minimum imply extrapolation into the lower tail of the distribution, the use of (2.1) to estimate distribution of minima carries risk of serious bias. Extreme value theory provides an alternative method of modeling a minimum. Classic extreme value theory gives the asymptotic distribution for a minimum (e.g., Section 9.1.1 of Castillo et al. 2005 or Section 3.2 of Coles 2001). The limiting distribution of the minima belongs to one of the three forms known as the Gumbel, Fréchet, and Weibull families (and there are corresponding distributions for maxima that we will not explicitly consider here).

The three limiting distributions are embedded in the minimum generalized extreme value (GEV) family with a cdf

\[
G(x) = 1 - \exp \left\{ - \left[ 1 - \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \tag{2.4}
\]

where \( \xi \neq 0 \) and \( 1 - \xi (x - \mu)/\sigma \geq 0 \). The GEV family has three parameters: a location parameter, \( -\infty < \mu < \infty \), a scale parameter, \( \sigma > 0 \), and a shape parameter, \( -\infty < \xi < \infty \). The limit of (2.4) as \( \xi \to 0 \), leads to the minimum Gumbel family with cdf

\[
G(x) = 1 - \exp \left[ - \exp \left( \frac{x - \mu}{\sigma} \right) \right], \quad -\infty < x < \infty. \tag{2.5}
\]

The quantiles of the GEV distribution are obtained by solving \( G(x_p) = p \) for \( x_p \) giving:
If the parent distribution $F$ has a minimum limiting distribution $G$, then $F$ is said to be in the minima domain of attraction of $G$. In many practical applications, physical considerations will indicate the particular form of $G$. Table 9.5 of Castillo, Hadi, Balakrishnan and Sarabia (2005) also summarizes the maxima and minima domain of attraction of these three types of parametric limiting distributions. From that table, the normal, the SEV distribution (minimum Gumbel distribution) and the LEV distribution (maximum Gumbel distribution) all belong to the Gumbel minima domain of attraction.

Although the minimum Gumbel and the SEV distributions are equivalent, for clarity of purpose, we use the term minimum Gumbel to refer to a limiting distribution of minima and SEV to refer to a parent distribution.

2.3.2 The Block Minima Method

To estimate the distribution of minima from the thickness measurements using the extreme value distributions, we need to obtain data from a minimum process. The block minima method (described, for example, in Section 3.3.1 of Coles 2001) provides an alternative method of estimating the distribution of minima from the thickness measurements by grouping the data into blocks of equal or approximately equal size and taking as data the minimum in each block. For $n$ iid observations $X_1, \ldots, X_n$, let $m$ denote the number of blocks, so there are $B = n/m$ observations in each block. Let $X_{\text{min}_i} = \min\{X_{B(i-1)+1}, \ldots, X_{Bi}\}, i = 1, \ldots, m$ be the minimum value in block $i$. Then the block minima $X_{\text{min}_1}, \ldots, X_{\text{min}_m}$ are independent observations that will follow, approximately, a minimum extreme value distribution (either one of the minimum-type distributions or the generalized extreme value distribution).

2.3.3 Estimating the Minimum over a Population

In order to estimate the minimum of a population, when using the block minima method, one will generally need to extrapolate further into the tail of the extreme value distribution.
that is estimated by using the block minima method (e.g., Section 1.1.3 of Glegola 2007). In particular, this additional extrapolation is needed because when using the block minima method with blocks of size $B$, we obtain an estimate of the parent distribution of minima for a population of size $M/B$. If the constructed block minima data set $\{X_{\text{min}}^{1}, \ldots, X_{\text{min}}^{m}\}$ follows a minimum distribution $G_{M}(x)$ (either the Gumbel or the GEV distribution) corresponding to blocks of size $B$, then the minimum thickness $Y_{M} = \min\{X_{1}, \ldots, X_{M}\}$ of the population with size $M$ can be treated as the minimum of a sample of $M/B$ independent block minima with blocks of size $B$ and the minimum distribution of $Y_{M}$ is

$$G_{M}(x_{c}) = \Pr(Y_{M} \leq x_{c}) = 1 - \Pr\left(Y_{\text{min}}^{1} > x_{c}, \ldots, Y_{\text{min}}^{M/B} > x_{c}\right) = 1 - [1 - G_{B}(x_{c})]^{M/B},$$

(2.7)

where $Y_{\text{min}}^{j} = \min\{X_{B(j-1)+1}, \ldots, X_{Bj}\}, j = 1, \ldots, M/B$ is the minimum value in block $j$ (having size $B$) and $Y_{M} = \min\{Y_{\text{min}}^{1}, \ldots, Y_{\text{min}}^{M/B}\}$.

If one wants to control $G_{M}(x)$, such that $G_{M}(x) = \Pr(Y_{M} \leq x) = p$, then one would choose the threshold to be $x_{p} = G_{M}^{-1}(p)$, the $p$ quantile $x_{p}$ of the distribution of the population minimum $Y_{M}$. The translation to the adjusted quantile in terms of the block minima distribution $G_{B}(x)$ is as follows:

$$x_{p} = G_{M}^{-1}(p) = G_{B}^{-1}\left(1 - (1 - p)^{B/M}\right) = G_{B}^{-1}(p^{*}),$$

(2.8)

where $p^{*} = 1 - (1 - p)^{B/M}$. It is this quantile that will be the focus of our simulation to study estimation performance in Sections 2.4, 2.5 and 2.6.

2.4 Design of the Simulation Experiment and Simulation Details

2.4.1 Objective of the Simulation

Cox, Isham and Northrop (2002) investigate the asymptotic variance of the quantile estimates of a distribution of maxima under different estimation methods. Asymptotic variances
do not take into account the bias. Bias, however, can play a central role in evaluation of the accuracy of estimators in finite samples. Extreme value theory is based on large-sample asymptotic results. It is important to understand how methods based on this large-sample theory will perform with finite samples. The objective of this section is to describe the design of a simulation experiment and to suggest criteria with which we can compare various statistical methods for modeling and making inferences about the minimum thickness of a pipeline and other applications where the block minima method might be used. As we will see, choice of block size plays an important role in the performance of the block minima estimation method. We then use this simulation to explore the impact of block size in the block-minima extreme value method for estimating small quantiles of a distribution of a minimum.

2.4.2 Experimental Factors and Their Levels

In the simulation, we used three parent distributions: the normal, the SEV (minimum Gumbel) and the LEV (maximum Gumbel) distributions. All of these distributions belong to the Gumbel minima domain of attraction. In the simulation we used sample sizes \( n = 200 \) (approximately the smallest sample size that would be suitable when using the block minima method) and \( n = 1,000 \) (approximately the largest sample size we have seen in pipeline applications). In the block minima method, with the sample size \( n = 200 \), the observations were divided into equal-size blocks with sizes in the set \( S_{200} = \{2, 4, 10, 20, 40\} \). With sample size \( n = 1,000 \), the block sizes in our evaluations were chosen from \( S_{1000} = \{5, 10, 20, 50, 100\} \).

The ratio of the population size to the sample size \( M/n \) affects the effective amount of extrapolation. The three levels for this factor are taken from the set \( M/n = \{10, 100, 1000\} \).

The methods used to model the distribution of a minimum in this simulation are:

- Method 1 corresponds to the traditional minimum method described in Section 2.2.1 where the form of the parent distribution is specified.

- Method 2 is based on the block minima method (abbreviated as BL\(_{\text{min}}\)) described in Section 2.3.2 using the minimum Gumbel distribution to describe the minima of the blocks. This would be the appropriate estimation method if the parent distribution is
known to be one of the distributions that has the minimum Gumbel distribution as its limiting distribution (e.g., the normal, SEV, or LEV that are used as parent distributions in the simulation).

- Method 3 is also based on the block minima method, using the minimum GEV distribution to fit the block minima data. This would be an appropriate method to use if there were no information about the underlying parent distribution.

The probability plots for the pipeline wall thickness data (Figure 2.2 in Section 2.2.2) and physical knowledge about the thickness data (they must be positive and were obtained by taking the minimum value in each one-foot segment) suggest that the parent distribution of the pipeline wall thickness can be adequately described by the distribution in the Weibull minima domain of attraction (e.g., the Weibull or the lognormal distribution). Thus, the distribution of the logarithm of the pipeline wall thickness is in the Gumbel minima domain of attraction and the simulation results can be used as the guide to model the logarithm of the pipeline wall thickness.

In practice, one seldom knows the most appropriate parametric form of the parent distribution. Without knowledge of the parent distribution, one might use an inadequate parametric distribution to estimate the minimum distribution. A slight discrepancy in the parent distribution can cause a substantial bias error in the estimation of the minimum distribution. To illustrate the sensitivity to an incorrect choice of a parent distribution, in simulation, we also use different assumed parent distributions.

2.4.3 ML Estimation and Comparison Criteria

The maximum likelihood (ML) method is used for estimating the parameters in the extreme value distribution. For details on the ML estimation of the parameters in the two-parameter extreme value distribution, see, for example, Meeker and Escobar (1998) and Lawless (2002). Coles (2001) provides more details on the ML method for estimating the parameters in the generalized extreme value distribution.
We use the usual definitions of mean square error (MSE), variance, and bias for comparison of estimators. For an unknown quantity $\theta$ with $\hat{\theta}$ as an estimator, the MSE of $\hat{\theta}$ is:

$$\text{MSE} (\hat{\theta}) = E \left[ (\hat{\theta} - \theta)^2 \right] = \text{Var} (\hat{\theta}) + \left[ \text{Bias} (\hat{\theta}) \right]^2 \quad (2.9)$$

where $\text{Bias} (\hat{\theta}) = E (\hat{\theta} - \theta)$.

In the simulation, we evaluate these properties of the ML estimators of the lower quantiles of the distribution of the minimum. The ML estimators of the $p$ quantile $x_p$ are obtained by substituting the ML estimators of the parent distribution parameters into the quantile expressions in Section 2.3.1.

### 2.5 Simulation Results: Effect of Block Size on MSE

We investigated plots of the MSE, variance, and bias for the three different parent distributions. The ordering of the estimation method MSE curves for sample size $n = 1,000$ are generally similar to those for sample size $n = 200$. Thus, we will primarily display MSE results for $n = 200$. Subsequently, we will provide other plots that help understand the variance-bias trade-off.

#### 2.5.1 Graphical Summary of MSE Results for the Normal Parent Distribution

Figure 2.3 compares the MSEs of ML estimators of quantiles ranging from $x_{0.0015}$ to $x_{0.15}$ of the distribution of the minimum for the different combinations of number of blocks for samples of size $n = 200$ using a normal parent distribution and ratio $M/n = 10$. Figure 2.4 displays similar simulation results for the normal parent distribution when the sample size $n = 1,000$.

As seen in Figures 2.3 and 2.4, with data from a normal parent distribution, and small block sizes, using the BL_{min}-Gumbel method results in ML estimators with a large MSE caused by large bias because the asymptotic extreme value approximation is poor. With a sample size $n = 200$, and ratio $M/n = 10$, a block size of at least 20 is needed for the distribution of minima to be adequately described by the BL_{min}-Gumbel method. In further simulations (details not given here, but note Figure 2.6 in Section 2.5.3), as the ratio $M/n$ increases from 10 to 1,000,
resulting in a larger amount of effective extrapolation, bias is amplified, and the needed block size increases, say to 40.

As seen in Figure 2.4, with a sample size \( n = 1,000 \) and ratio \( M/n = 10 \), a block size of at least 20 is again needed for the BL\(_{\text{min}}\)-Gumbel method. As the ratio \( M/n \) increases from 10 to 1,000 (again, details not given here), the block size needed to compensate for the additional bias and to provide good performance increases to say 50. The BL\(_{\text{min}}\)-GEV method, however, results in ML estimators with relatively small bias, even with small block sizes. Generally, however, the variance is much larger with the BL\(_{\text{min}}\)-GEV method because an additional shape parameter must be estimated. In Figure 2.3, when the number of blocks is 5, the MSE curve of the quantile estimates using the BL\(_{\text{min}}\)-GEV method is so large that it is off-scale.

For an actual normal parent distribution, the MSEs of the quantile estimators using the distribution-of-minimum method based on the mis-specified SEV parent distribution are much greater than the MSEs of quantile estimators based on the other minimum distributions. This
Figure 2.4 A comparison of the MSEs of the ML quantile estimators of the minimum for the normal parent distribution when the sample size is $n = 1,000$ and ratio is $M/n = 10$.

is true for all values of the ratio $M/n$. The minimum-distribution estimators based on the mis-specified LEV parent distribution, however, have relatively small MSEs. This difference in behavior is because the lower tail of the LEV (SEV) distribution is similar to (different from) the lower tail of the normal distribution.

2.5.2 Explanation of MSE Results for the Normal Parent Distribution

Here we look in more detail at the sampling distributions of the ML estimators of quantiles of the distribution of a minimum in order to better understand the reasons for the behaviors seen in Section 2.5.1 and to provide insight into the choice of estimation method and block size.

The box plots in Figure 2.5 show the empirical sampling distribution of $\hat{x}_{0.05}$ using the $\text{BL}_{\text{min}}$-Gumbel and $\text{BL}_{\text{min}}$-GEV methods for different block sizes $B = \{4, 10, 20\}$ when using a normal parent distribution with sample size $n = 200$ and ratio $M/n = 10$. The box plots
provide some insight into the reasons for the differing behaviors (in bias and variance) between the BL_{min}-GEV and BL_{min}-Gumbel methods. In particular, when the number of blocks \( m \) is not large (say 20 or less), the BL_{min}-GEV method generates a substantial fraction of extremely small estimates, resulting in both large bias and variance. The median of the sampling distributions from the BL_{min}-GEV method, however, remains relatively close to the truth even when the block size is small. This is in contrast to the BL_{min}-Gumbel method where there is substantial bias when applied to the minimum of small blocks of normally-distributed variates.

![Graph showing comparison of BLmin methods](image)

**Figure 2.5** A comparison of the sampling distributions of \( \hat{x}_{0.05} \) using the BL_{min}-Gumbel and BL_{min}-GEV methods for the normal parent for \( n = 200 \) and \( M/n = 10 \) and different combinations of \( B \) and \( m \). The horizontal lines indicate the position of the true 0.05 quantile of the distribution of the minimum in the population. Note that some BL_{min}-GEV estimates are off scale for \( m = 10 \) and 20.
2.5.3 The Effect of More Extreme Extrapolation

Figures 2.3, 2.4 and 2.5 provide basic comparisons for a modest amount of extrapolation into the lower tail of the distribution (i.e., $M/n = 10$). In order to compare the BL$_{\text{min}}$-Gumbel and BL$_{\text{min}}$-GEV methods with larger amounts of extrapolation, Figure 2.6 displays a pair of box plots of the empirical sampling distributions of $\hat{x}_{0.05}$ using for a normal parent distribution with block sizes $B = 4$ and $B = 20$ when sample size $n = 200$ with different ratios $M/n = \{10, 100, 1000\}$.

![Figure 2.6](image)

Figure 2.6 A comparison of the sampling distributions of $\hat{x}_{0.05}$ using the BL$_{\text{min}}$-Gumbel and BL$_{\text{min}}$-GEV methods for the normal parent when the sample size is $n = 200$ with different combinations of $M/n$, $B$ and $m$. The horizontal lines indicate the position of the true 0.05 quantile of the distribution of the minimum in the population. Note that some BL$_{\text{min}}$-GEV estimates are off scale for $m = 10$.

Figure 2.6 (a) shows results from samples of size $n = 200$ from a normal parent and blocks size $B = 4$. We chose $B = 4$ for this example because the MSE of the quantile estimates using the BL$_{\text{min}}$-GEV method with block size $B = 4$ are the smallest among all choices of the blocks in Figure 2.5. The biases of the quantile estimators using the BL$_{\text{min}}$-GEV method are less than...
the biases of quantile estimators using the BL$_{\text{min}}$-Gumbel method for all ratios $M/n$. As the ratio increases, the biases of quantile estimators using both BL$_{\text{min}}$-GEV and BL$_{\text{min}}$-Gumbel methods increase. The biases of quantile estimators using the BL$_{\text{min}}$-Gumbel method increase more rapidly than the biases using the BL$_{\text{min}}$-GEV method. As the ratio $M/n$ becomes fairly large, say 1,000, the BL$_{\text{min}}$-GEV method, particularly when the number of the blocks (effective sample size) is not large, will generate a substantial fraction of small outliers, contributing to increased bias and variance. The behaviors of the BL$_{\text{min}}$-GEV and BL$_{\text{min}}$-Gumbel methods in the above box plots are consistent with what we observed before in the MSE plots in Figures 2.3 and 2.4. Figure 2.6 (b) shows box plots, similar to Figure 2.6 (a), but with the larger block size $B = 20$. We chose $B = 20$ for this example because the MSE of the quantile estimators using the BL$_{\text{min}}$-Gumbel method with block size $B = 20$ are the smallest among all choices of the blocks in Figure 2.5. We see that with the larger blocks (and thus a smaller number of blocks for estimation), the BL$_{\text{min}}$-GEV method has a large variance due to the existence of a substantial number of small outliers. Also, as the ratio $M/n$ increases, resulting in a large amount of effective extrapolation, the variance increases tremendously and the MSE of the BL$_{\text{min}}$-GEV quantile estimators grows explosively.

2.5.4 MSE Results for the LEV and SEV Parent Distributions

Here we look at behavior of the competing estimation methods under alternative parent distributions. Figures 2.7 and 2.8 are similar to Figure 2.3 and provide comparisons of the MSEs of the quantile estimators of the distribution of the minimum in the population under the LEV and the SEV parent distributions, respectively, for a sample size $n = 200$, and a ratio $M/n = 10$.

As can be seen in Figure 2.7, with data from the LEV parent distribution, and small block sizes, using the BL$_{\text{min}}$-Gumbel method again results in ML estimators with a relatively large MSE unless the block size is large (e.g., 20 or more). We know from box plots similar to Figures 2.5 and 2.6 (not shown here) that poor performance of the BL$_{\text{min}}$-Gumbel method is due mostly to negative bias in the estimates of quantiles $x_p$ with small $p$. On the other hand, Figure 2.7 also suggests that the BL$_{\text{min}}$-GEV method performs relatively well when there is a
Figure 2.7 A comparison of the MSEs of the ML quantile estimators of the minimum for the LEV parent distribution when the sample size is \( n = 200 \) and ratio is \( M/n = 10 \).

large number of blocks (say more than 50), even if those blocks are not large. In simulation results not displayed here, these conclusions remain the same as the ratio \( M/n \) varies from 10 to 1,000.

In Figure 2.7, for the LEV parent distribution, the MSEs of quantile estimators using the minimum distribution based on the mis-specified SEV parent distribution are very large. The MSE curves for the quantile estimators using the mis-specified minimum normal distribution, however, behaves much better when compared with the performance of the mis-specified minimum SEV distribution. This is because the lower tail behavior of the normal distribution is more similar to that of the LEV than it is to the SEV.

In Figure 2.8, for the SEV parent distribution, the distribution of the minimum is exactly the minimum Gumbel distribution, and there is no model-specification bias when using the BL\(_{\text{min}}\)-Gumbel method. As the number of blocks increases, the MSE of the quantile estimates using the BL\(_{\text{min}}\)-Gumbel method decreases due primarily to reduction in variance.
2.5.5 Impact of the Shape Parameter on the GEV Quantile Estimates

To understand why some estimates of the quantiles can be extremely small when using the BL\textsubscript{min}-GEV method, especially with a small number of blocks, we looked at scatter plots of BL\textsubscript{min}-GEV method quantile estimates versus the corresponding shape parameter estimates. Figure 2.9 (a), for example, shows estimates of $x_{0.05}$ for sample size $n = 200$ and block size $B = 4$ (so there are $m = 50$ blocks in the sample). The plot shows that the small estimates of the quantile $x_{0.05}$ result when the shape parameter estimates are large. The range of shape parameter estimates under the SEV parent distribution is larger than those under the normal and the LEV parent distributions, resulting in the smaller quantile estimates with the SEV parent. For the LEV and the normal parent distributions, because the sampling distribution of the shape parameter estimates does not extend far into the positive range, the bias of the BL\textsubscript{min}-GEV quantile estimators is not as large as it is for the SEV parent.

Figure 2.9 (b) shows a similar scatter plot for block size $B = 10$ (number of blocks $m = 20$).
Compared with Figure 2.9 (a), the shape parameter estimates using the BL\textsubscript{min}-GEV method in Figure 2.9 (b) are, overall, larger and the smallest quantile estimates are much smaller than those in Figure 2.9 (a). The variances of the GEV parameter estimates are large when the number of blocks is small, leading to large variances for the estimators of the quantiles.

Figure 2.9 Scatter plot of the shape parameter estimates \( \hat{\xi} \) versus the \( \hat{x}_{0.05} \) quantile estimates for the BL\textsubscript{min}-GEV method for normal, SEV and LEV parent distributions when sample size is \( n = 200 \) and ratio is \( M/n = 10 \). In part (a), the block size is \( B = 4 \) and the number of blocks is \( m = 50 \); In part (b), the block size is \( B = 10 \) and number of blocks is \( m = 20 \).

### 2.6 Simulation Results: Effect of Block Size and Parent Distribution on Confidence Intervals

In the previous sections, our discussion focused on investigating the properties of point estimators of the quantiles of the minimum distribution. Interval estimates that quantify the uncertainty of the point estimator are usually needed in statistical analysis. This was certainly true in our pipeline example. Although one can generally expect that confidence interval
procedures based on point estimators with good (poor) properties will lead to well (poorly) behaved confidence interval procedures, in this section we compare quantile relative likelihood profile plots that give a sense of the effect that the block-size choice will have on the interval estimates. These profile plots also provide an alternative graphical tool for comparing the behaviors of the quantile estimates under different block size and distribution combinations. Because quantile relative likelihood profile curves are random, we plotted multiple realizations of these curves, corresponding to simulated data sets. In the pipeline wall thickness application, physical knowledge, the measurement process, and the data (Figure 2.2) suggest that the parent distribution in most cases can be adequately described by the Weibull or the lognormal distribution (i.e., the logarithm of pipeline wall thickness can be adequately described by the SEV or the normal distribution). Thus, in this section, we will mainly investigate quantile relative likelihood profile plots under the normal and the SEV parent distributions.

2.6.1 Simulation Results on Profile Likelihoods for the Normal Parent Distribution

Here we look at the confidence interval of the ML estimators of the quantiles of the distribution of the minimum under the normal parent distribution. Both plots in Figure 2.10 show 50 profile curves. The sampling distribution of the quantile point estimators can be visualized by looking at the maxima of the profiles. Additionally, the width of the likelihood-based approximate confidence intervals can be used to visually assess the precision of the corresponding point estimator.

The simulation results displayed in Figures 2.3 and 2.4 suggest that if the parent distribution is the normal distribution, a fairly large number of blocks is needed to use the BL_{min}-GEV method and a large block size is needed to use the BL_{min}-Gumbel method. As explained at the end of Section 2.5.4, with an SEV parent distribution, however, we know that there is no model specification bias when using the BL_{min}-GEV and BL_{min}-Gumbel methods to estimate small quantiles and the use of blocks would not be needed at all. For our likelihood profile evaluations for the BL_{min}-GEV method, we used a block size $B = 4$ (which with $n = 200$ gives 50 blocks). For the BL_{min}-Gumbel method, if we know that the parent is the normal distribution, Figure
suggests the block size \( B = 20 \). If we know parent is the SEV distribution, Figure 2.8 suggests the block size \( B = 1 \) (i.e., no blocking). Thus, we use a compromise block size \( B = 10 \) (giving 20 blocks).

![Comparison of quantile relative likelihood profile plots](image)

**Figure 2.10** Comparison of quantile relative likelihood profile plots using (a) \( \text{BL}_{\text{min}} \)-GEV with the block size \( B = 4 \) and (b) \( \text{BL}_{\text{min}} \)-Gumbel with the block size \( B = 10 \) when the sample size is \( n = 200 \), the population size is \( M = 5,649 \), quantile is \( x_{0.05} \) and the parent distribution is normal. The vertical line indicates the position of the true 0.05 quantile of the distribution of the minimum in the population. The horizontal lines allow visualization of corresponding approximate 50% and 95% likelihood-based confidence intervals.

The quantile relative likelihood profile plots in Figure 2.10 show that the \( \text{BL}_{\text{min}} \)-Gumbel method with a block size \( B = 10 \) results in less variability than the quantile relative likelihood profile plots using \( \text{BL}_{\text{min}} \)-GEV method with block size \( B = 4 \), especially in the lower endpoints of the likelihood-based approximate 50% and 95% confidence intervals (the horizontal lines in these plots are based on a simple chi-square distribution calibration). The MSE of the quantile estimates using the \( \text{BL}_{\text{min}} \)-GEV method, however, is smaller than that using the \( \text{BL}_{\text{min}} \)-Gumbel method. Similar simulations using \( n = 1,000 \) data sets and correspondingly larger number of blocks (details not shown here) gave similar results except that, as expected, precision was improved.
2.6.2 Simulation Results on Profile Likelihoods for the SEV Parent Distribution

Figure 2.11 provides comparisons of quantile relative likelihood profile plots using the SEV parent distribution with a sample size $n = 200$. The $\text{BL}_{\min}$-GEV method performs poorly, even with 50 blocks. The results in Figure 2.9 for the SEV parent help explain this behavior. As expected, due to the lack of model-specification bias, the MSE of the quantile estimators using the $\text{BL}_{\min}$-Gumbel method with block size $B = 10$ performs well.

![Figure 2.11](image.png)

Figure 2.11  Comparison of quantile relative likelihood profile plots using (a) $\text{BL}_{\min}$-GEV with the block size $B = 4$ and (b) $\text{BL}_{\min}$-Gumbel with the block size $B = 10$ when the sample size is $n = 200$, the population size is $M = 5,649$, quantile is $x_{0.05}$ and parent distribution is the SEV distribution. The vertical line indicates the position of the true 0.05 quantile of the distribution of the minimum in the population. The horizontal lines allow visualization of corresponding approximate 50% and 95% likelihood-based confidence intervals.

2.6.3 General Conclusion from the Profile Likelihood Simulations

If the parent distribution is close to the lognormal distribution (normal on the log scale), our results (e.g., in Figures 2.3, 2.4 and 2.10) indicate that the combination of a smaller block size (e.g., $B = 4$ when $n=200$; $B = 10$ when $n = 1,000$) and the $\text{BL}_{\min}$-GEV method is an appropriate choice according to the MSE criterion. If the parent distribution is close
to the Weibull distribution (SEV on the log scale), the results suggest that one should use the BL\textsubscript{min}-Gumbel method to estimate small quantiles of a distribution and, for the sake of robustness, choose a moderately large block size (e.g., $B = 10$ when $n=200$ or $B = 20$ when $n = 1,000$).

### 2.7 Estimation of the Minimum Thickness in the Pipeline

Here we return to the pipeline wall thickness inspection data and compare the estimates of $p = 0.05$ quantile and the corresponding likelihood-based approximate confidence intervals under the different methods for estimating the distribution of a minimum thickness. As in Section 2.2, the pipeline wall thickness data consist of two simple random samples of size $n = 200$ and $n = 1,000$ locations from the population of $M = 5,649$ measurements at locations that had metal-loss features along the three-phase pipeline.

<table>
<thead>
<tr>
<th>Method</th>
<th>$B$</th>
<th>$m$</th>
<th>Lower</th>
<th>$x_{0.05}$</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Weibull</td>
<td>1</td>
<td>200</td>
<td>0.0958</td>
<td>0.1119</td>
<td>0.1248</td>
</tr>
<tr>
<td>Minimum lognormal</td>
<td>1</td>
<td>200</td>
<td>0.1727</td>
<td>0.1826</td>
<td>0.1913</td>
</tr>
<tr>
<td>Minimum Fréchet</td>
<td>1</td>
<td>200</td>
<td>0.1924</td>
<td>0.1991</td>
<td>0.2050</td>
</tr>
<tr>
<td>$\text{BL}_{\text{min}}$-Gumbel</td>
<td>10</td>
<td>20</td>
<td>0.0965</td>
<td>0.1325</td>
<td>0.1605</td>
</tr>
<tr>
<td>$\text{BL}_{\text{min}}$-GEV</td>
<td>4</td>
<td>50</td>
<td>0.0334</td>
<td>0.1452</td>
<td>0.1847</td>
</tr>
</tbody>
</table>

Table 2.1 ML estimates and likelihood-based approximate 95% confidence intervals for $\hat{x}_{0.05}$.

Table 2.1 lists the ML estimates and the corresponding lower and upper limits of the likelihood-based approximate 95% confidence intervals for the 0.05 quantile of the minimum distributions using different methods. Figure 2.12 displays the relative likelihood profile plots for the 0.05 quantile estimates using the block minima method. Figure 2.13 shows estimates of the parent and minimum distributions for the pipeline wall thickness data on Weibull probability paper under the different estimation methods when sample sizes are $n = 200$ (left) and $n = 1,000$ (right).

Recall that Figure 2.2 suggested that the parent distribution of wall thicknesses could be described by either a Weibull or a lognormal distribution, but that the Weibull distribution fits better. A similar plot for the population of 5,649 thickness values (not shown here) confirmed that the Weibull provides a better description than the lognormal distribution. Among many
distributions we tried, the Weibull distribution fits the 5,649 population thickness values best, especially in the lower tail. We then use the Weibull distribution to describe the population distribution. Using all of the $M = 5,649$ thickness values to estimate the parameters in the Weibull parent distribution, the ML estimate for the 0.05 quantile of the minimum Weibull distribution is 0.115. When comparing estimates for different models, we use this quantity as the “true quantile” being estimated. Also the actual minimum thickness of the overall 5,649 features in the population was 0.146 inches, indicated in the plots by a vertical line.

In the top two plots in Figure 2.13, comparing the distribution estimates for the Weibull and lognormal parent distributions, we see the strong divergence from the truth when extrapolating toward small probabilities. Also, the estimate of the distribution of the minimum based on the lognormal assumption has serious upward bias. The estimate based on the Weibull distribution assumption is, as expected, close to the population quantile.

In the middle row of plots in Figure 2.13 we see that the $BL_{\text{min}}$-Gumbel method provides estimates of the distribution of the minimum that are similar to those of the minimum Weibull method, but with less precision due to the smaller effective sample size (number of blocks). Although this method would be expected to provide more robustness, as we saw in Section 2.5.4, with blocks of size 10 (used for the $n = 200$ sample), the degree of robustness would be limited if the parent distribution is misspecified as a lognormal.

In the bottom row of Figure 2.13 we see that the $BL_{\text{min}}$-GEV method also provides estimates of the distribution of the minimum that are similar to those of the minimum Weibull method, but now with much less precision, again due to the smaller effective sample size. Nevertheless, as we saw in Section 2.5.4, this method will not be affected by the model-specification bias that could affect the other estimation methods.

In all plots in Figure 2.13, the “true quantile” falls within the likelihood-based approximate 95% confidence intervals for $\hat{x}_{0.05}$ under the three different estimation methods.

### 2.8 Conclusions, Recommendations, and Areas for Further Research

The observations from previous sections lead to the following conclusions:
• The direct distribution-of-minimum method provides the most precise quantile estimators when the parent distribution is correctly specified. Of course, in practice, we cannot expect, in many applications, to know the parent distribution exactly and we have seen that misspecification can lead to seriously biased answers.

• Sensitivity analysis showed that the performances of the incorrectly specified minimum distributions differ and depend highly on the lower tail behavior of the underlying parent distribution. For the normal parent distribution, the minimum distribution based on the SEV parent distribution performs poorly while the minimum distribution based on the LEV parent distribution provides fairly good quantile estimators. For the LEV parent distribution, although neither the SEV nor the normal minimum distributions provides good quantile estimators, the normal minimum distribution performs much better than the minimum distribution based on the SEV parent distribution. For the SEV parent distribution, both the normal and the LEV minimum distributions perform poorly. Compared with the performance of the LEV minimum distribution, the normal minimum distribution performs somewhat better. Generally, when extrapolating into the lower tail of the distribution, with the minimum distribution method using the SEV distribution assumption is conservative, relative to the normal and LEV distributions. This, of course, is not surprising given the lower tail behavior of these three distributions.

• In the block minima method, the choice of block size can be viewed as a trade-off between variance and bias and the trade-off is stronger for the BL_{min}-Gumbel method than it is for the BL_{min}-GEV method. For example, with a normal distribution parent, a large proportion of the MSE is contributed by the squared bias term, especially when the block sizes are small and the BL_{min}-Gumbel method is used. With a large number of blocks, the variance of the quantile estimator for the limiting distribution is relatively small. With a fixed amount of data, however, increasing the number of blocks will result in smaller blocks which will increase bias in the BL_{min}-Gumbel method (unless the parent is SEV), because the asymptotic extreme value theory assumes minima from large blocks.

• If there is a sufficiently large number of blocks and the blocks are of sufficient size, the BL_{min}-GEV method provides inferences on the distribution of the minimum that are robust
without need to specify the particular form of the parent distribution. If, however, the number of blocks is too small (say less than 30), ML estimation of the GEV parameter can fail to converge properly, no matter what the initial values are in the numerical optimization algorithm (Coles and Dixon 1999). Because of the unboundedness of the usual (product of densities) GEV likelihood, we used the “correct” (probability based) likelihood (as described, for example, by Barnard 1967, and Giesbrecht and Kempthorne 1976) to estimate the GEV parameters. As seen in Figure 2.9, even with this approach, large positive shape parameter estimates can arise and these result in the extremely small quantile estimates and in the BL$_{\text{min}}$-GEV method, the degree of this behavior depends strongly on the shape of the parent distribution.

- In the BL$_{\text{min}}$-Gumbel method, if the parent distribution is close to the lognormal or the Fréchet distribution (normal or LEV on the log scale), a large block size is needed to provide quantile estimates with small MSEs. If, however, the lower tail of the parent distribution is close to that of the Weibull distribution (SEV on the log scale), the BL$_{\text{min}}$-Gumbel method with a small block size performs well, even if the block sizes are small.

- Other simulation results (details not shown here) show that the BL$_{\text{min}}$-GEV method will also, in general, have some bias due to the finite block size but the bias is small relative to that in the BL$_{\text{min}}$-Gumbel method and unless the number of blocks is large, the MSE tends to be dominated by variance.

- In the BL$_{\text{min}}$-GEV method, if the parent distribution is close to the lognormal or the Fréchet distribution (normal or LEV on the log scale), using a large number of blocks (even with a small block size) provides good quantile estimates. If the parent distribution is close to the Weibull distribution (SEV on the log scale), the number of blocks needed to give reasonable precision is larger.

- For a given sample size and quantile of the minimum distribution, the ratio $M/n$ affects the effective amount of extrapolation. As the ratio $M/n$ increases (implying more extreme extrapolation into the distribution tail), bias and variance in the estimators of the quantile of interest will increase for both the BL$_{\text{min}}$-GEV method and the BL$_{\text{min}}$-Gumbel method.

Based on these conclusions, we have the following recommendations:

- The performance of ML estimators for quantiles based on the BL$_{\text{min}}$-GEV method can be
poor unless there is a large number of blocks (say greater than 50). One could use alternative point estimation methods. For example, the probability weighted moments method (PWM) described by Hosking, Wallis and Wood (1985) have been shown to have small sample superiority. It is not clear, however, that such alternative methods offer improvement when it is necessary to find confidence intervals for the quantile of interest (as is generally the case).

- Choosing an appropriate block size is essential for the successful use of the block minima method. Because the tail behaviors of different parent distributions could result in different shape parameter estimates in the GEV distribution, in order to make a decision between the BL$_{\text{min}}$-GEV method and the BL$_{\text{min}}$-Gumbel method, one should consider the degree of confidence that one has in the knowledge (perhaps due to physical knowledge or sampling considerations) of the parent distribution and the domain of attraction. After choosing a method, such knowledge is also important for choosing a block size.

- If the lower tail of the parent distribution can be appropriately described by the lognormal or the Fréchet distribution (normal or LEV on the log scale), one should choose a small block size with an adequately large number of the blocks when using the BL$_{\text{min}}$-GEV method. The BL$_{\text{min}}$-Gumbel method, however, requires a relatively larger block size to produce precise quantile estimators.

- If the parent distribution is closer to the Weibull distribution (SEV on the log scale), especially in the lower tail, the BL$_{\text{min}}$-Gumbel method with a small block size for the quantile estimation is recommended.

- With a large sample size (e.g., $n = 1,000$), the BL$_{\text{min}}$-GEV method provides an attractive method because of the robustness that it provides. For smaller sample sizes, the number of blocks used may not be large enough to provide a reasonable amount of precision. The BL$_{\text{min}}$-Gumbel method is recommended for a moderate sample size (e.g., $n = 200$), but distribution-specification bias may be large if block-size distributions are not large enough. When the sample size is small (e.g., $n = 20$), one should use the traditional statistical method to model the minimum, recognizing that serious bias could be an issue if the parent distribution is seriously misspecified.

Some areas for future research are:
• Our study focuses on the evaluation of the point estimates. In statistical inference, the accuracy of the quantile estimators using different estimation methods is another main issue. It would be of interest to compare confidence intervals of the quantile estimates in terms of the coverage probability.

• Methods that relax the restriction of equal block size in the block minima method could provide flexibility in modeling the minimum thickness of the pipeline and choosing the appropriate block sizes in the block minima extreme value theory method.

• The peaks over threshold (POT) method is an alternative to the block minima. A study similar to this one could be conducted to investigate threshold choice for the POT method.

• Bayesian methods, combining the pipeline wall thickness measurements with the prior information, especially, on the GEV shape parameter, have the potential to provide more precise quantile estimate for the minimum GEV distribution when legitimate prior information is available.

• A method for analyzing pipeline data taken over time to estimate corrosion rates and predict minimum at future points in time could be developed.

2.9 Acknowledgments

We would like to thank Luis Escobar, Ulrike Genschel, Yili Hong and Alyson Wilson for providing helpful comments on an earlier version of this paper.
Figure 2.12  Relative likelihood profile plots for the quantile $x_{0.05}$ under the BL$_{\text{min}}$-Gumbel and the BL$_{\text{min}}$-GEV methods for $n = 200$ and $n = 1,000$. The horizontal lines indicate corresponding approximate 50% and 95% confidence intervals.
Figure 2.13  Weibull probability plots of estimates of the parent distribution and distribution of minimum (DoM) for the pipeline wall thickness inspection data under the three different estimation methods for $n = 200$ (left) and $n = 1,000$ (right). The shorter vertical tick marks on the $p = 0.05$ line indicate likelihood-based 95% confidence interval for the $x_{0.05}$ quantile of the minimum distributions. The longer vertical tick mark on the $p = 0.05$ line indicates “true quantile” $x_{0.05}$ based on the 5,649 population thickness values.
Bibliography


CHAPTER 3. USING DEGRADATION MODELS TO ASSESS PIPELINE LIFE

A paper to be submitted
Shiyao Liu and William Q. Meeker
Department of Statistics
Iowa State University

Abstract

Longitudinal inspections of pipeline thickness at particular locations along the pipeline provide useful information to assess the lifetime of the pipeline. In applications with different mechanisms of corrosion processes, we have observed various types of general degradation paths. In one application, we used a degradation model to describe the corrosion initiation and growth behavior in the pipeline, and employed a Bayesian approach for parameter estimation for the degradation model. We also built a hierarchical model to quantify the pipeline corrosion rate for similar circuits within a single facility, under the assumption that the corrosion rates at particular locations are constant over time within a circuit in the facility. The failure-time and remaining lifetime distributions are derived from the degradation model, and we compute Bayesian estimates and credible intervals of the failure-time and remaining lifetime distribution.

Key Words: Bayesian; longitudinal data; pipeline reliability.
3.1 Introduction

3.1.1 Motivation and Purpose

Repeated measures of wall thickness across time at sampled locations along a pipeline circuit can be used to evaluate the reliability of a pipeline. Degradation models for longitudinal inspections of the pipeline thickness can be used to describe pipeline corrosion behavior, estimate the lifetime distribution of pipeline components, and predict the remaining lifetime of a pipeline circuit. There are two different purposes for such analyses: (1) estimating the lifetime distribution of pipeline segments to provide information that can be used to plan the construction of future pipelines and (2) to estimate the remaining life of existing pipelines. Depending on degradation and corrosion mechanisms, different statistical models and methods are needed to analyze pipeline data. In this paper, we analyze thickness data from two different pipelines and propose degradation models for each application. In some degradation models, it is computationally challenging to estimate parameters using the traditional likelihood-based method. Bayesian methods with appropriate prior distributions provide an alternative approach for estimating parameters of a complicated degradation model. In addition, evaluation of the failure time and remaining lifetime distributions is also computationally feasible and efficient when using the Bayesian method.

3.1.2 Pipeline Data

Figure 3.1 shows time plot of longitudinal pipeline data from Circuit G in Facility 3. Data were obtained from a sample thickness measurement locations (TMLs). For the first two inspections, only 12 TMLs were used. Subsequently, as perceived risk of failure increased, an additional 76 TMLs were used. Some of these TMLs correspond to elbows and the others correspond to straight pipes. For each TML, the thickness was measured at four different quadrants located at the 0, 90, 180, and 270 degree position (top, right, bottom, and left). The lines joining the points represent the degradation paths of the different combinations of location and quadrant. The first inspection was performed on February 11, 1995, a number of years after the pipeline had been installed.
The second pipeline data set is from a different facility. Figure 3.2 displays time plot for the pipeline data from Circuit Q in Facility 1. The data set consists of thickness values at 33 TMLs and each TML was measured at 4 times. Three component types of the pipeline in this data are elbow, straight pipe, and tee. In this facility, the first measurement was taken at the pipeline installation date. The time plot indicates that the original thicknesses vary from TML to TML. Also, the tee pipes are generally thicker than the elbow and straight pipes.

### 3.1.3 Related Work

Degradation models are often used to assess reliability of industrial products. Lu and Meeker (1993) illustrates that under some simple degradation path models, there can be a closed-form expression for the failure time distribution. Chapter 13 of Meeker and Escobar (1998) gives a general introduction to degradation models and described the relationship between the degradation and failure-time analysis methods of estimating a time-to-failure distribution. Chapter 8 of Hamada et al. (2008) provides an overview of Bayesian degradation...
models and uses several examples to illustrate how to estimate parameters of a degradation model. Nelson (2009) discusses a model for defect initiation and growth over time and uses maximum likelihood to estimate parameters in the model. Sheikh, Boah, and Hansen (1990) analyze data from water injection pipeline systems and use the Weibull distribution to model the time-to-first-leak. Pandey (1998) uses a probability model to estimate the lifetime distribution of a pipeline before and after repair due to the metal loss.

### 3.1.4 Overview

The rest of this paper is organized as follows. Section 3.2 proposes a degradation model for pipeline data from Circuit G in Facility 3 and uses the Bayesian approach to estimate the parameters in the degradation model. Section 3.3 derives failure time and remaining lifetime distributions for the circuit and computes the Bayesian estimates and the corresponding credible intervals. Section 3.4 analyzes pipeline data from Circuit Q in Facility 1. A degradation model is proposed to describe the corrosion initiation and growth behavior observed in this
pipeline. Section 3.5 evaluates the failure time distribution and predicts the remaining lifetime distribution of Circuit Q in Facility 1. In order to study the data needed for estimability, Section 3.6 analyzes simulated data for a single circuit having more than one inspection after corrosion initiation. Section 3.7 describes a hierarchical statistical model for the simulated pipeline data from several circuits within Facility 1. Section 3.8 contains the concluding remarks and areas for future research. An appendix provides OpenBUGS code of the hierarchical model that was used in Section 3.7.

3.2 Modeling Pipeline Data from Circuit G in Facility 3

In this section, we focus on the analysis of the pipeline data from Circuit G in Facility 3 as shown in Figure 3.1. We propose a degradation model and Bayesian estimation with diffuse prior distributions to estimate the parameters in the degradation model.

3.2.1 Degradation Model for Pipeline Data from Circuit G in Facility 3

We use $Y_{it_k}$ to denote the pipeline thickness at time $t_k$ for TML $i$ ($i = 1, 2, \ldots, 88; k = 1, 2, \ldots, 7$). We assume that the degradation path of Circuit G in Facility 3 is linear with respect to inspection time and has the form

$$Y_{it_k} = y_0 - \beta_{1i}(t_k - t_0) + \epsilon_{ik} \quad (3.1)$$

where $\beta_{1i}$ is $-1$ times the corrosion rate of location $i$ and $\epsilon_{ik}$ is the measurement error term. Here $y_0$ is the original thickness at installation time $t_0$. Specifically, the original thickness $y_0$ is 0.25 inches and the installation time $t_0$ is February 12, 1990. The precise dates of installation and beginning-use were not available and this date was obtained by extrapolating backwards in time. Because the corrosion rate defined as the thickness change per year varies from location to location and could only be negative, $\beta_{1i}$ in the degradation model (3.1) is a positive random variable. To guarantee a positive $\beta_{1i}$, we assume that $\beta_{1i}$ follows a lognormal distribution [i.e., $\beta_{1i} \sim \text{Lognormal} \ (\mu_{\beta_1}, \sigma_{\beta_1}^2)$] and that the measurement error is $\epsilon_{ik} \sim \text{NOR} \ (0, \sigma_{\epsilon}^2)$. Thus the parameters in the degradation model (3.1) are: $\theta = (\mu_{\beta_1}, \sigma_{\beta_1}, \sigma_{\epsilon})'$.
3.2.2 Bayesian Estimation of the Parameters in the Degradation Model

Bayesian estimation with the use of diffuse prior information is closely related to likelihood estimation (with a flat prior, the Bayesian joint posterior distribution is proportional to the likelihood). Bayesian methods provide a convenient alternative for estimating the parameters in the degradation model, particularly because we need to make inferences on complicated functions of the model parameters.

For the example, we use a normal distribution with mean zero and a large variance [i.e., \( \text{NOR} (0, 10^3) \)] as the prior distribution for the parameter \( \mu_{\beta_1} \). The prior distributions for \( \sigma_{\beta_1} \) and \( \sigma_\epsilon \) are Uniform \((0, 5)\). We obtain a large number of draws from the joint posterior distribution of the degradation model parameters using Markov Chain Monte Carlo (MCMC) implemented in OpenBUGS. Table 3.1 presents marginal posterior distribution summaries for the parameters in \( \theta \), including the mean and 95% credible intervals. Figure 3.3 shows the time plot of the fitted thickness values for Circuit G in Facility 3 with a 10-years extrapolation after the last inspection in January 20, 2003.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Posterior Mean</th>
<th>Posterior Std. Dev.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{\beta_1} )</td>
<td>-11.62</td>
<td>0.06263</td>
<td>-11.75 to -11.50</td>
</tr>
<tr>
<td>( \sigma_{\beta_1} )</td>
<td>0.5753</td>
<td>0.04743</td>
<td>0.4910 to 0.6768</td>
</tr>
<tr>
<td>( \sigma_\epsilon )</td>
<td>0.006045</td>
<td>2.537E-4</td>
<td>0.005575 to 0.006574</td>
</tr>
</tbody>
</table>

Table 3.1 Marginal posterior distribution summaries of the degradation model parameter estimates for pipeline data from Circuit G in Facility 3 using the degradation model (3.1).

3.2.3 Statistical Model for Different Quadrants

In Sections 3.2.1 and 3.2.2, we assumed that the corrosion rates of different quadrants from the same location follow the same distribution. In non-vertical pipes, however, the corrosion rate of locations in the upper quadrant might be expected to differ from that in the lower quadrant at the same TML. The degradation model in this section assumes that means of the logarithm of the corrosion rates vary from quadrant to quadrant. Assuming that the circuit with initial thickness 0.25 inches was installed on February 12, 1990, the degradation model is
Figure 3.3 Time plot showing the fitted thickness values for the pipeline data from Circuit G in Facility 3 using the degradation model (3.1).

\[ Y_{ij} t_k = y_0 - \beta_{1ij} (t_k - t_0) + \epsilon_{ijk} \]  

(3.2)

where \( \beta_{1ij} \) is the corrosion rate of quadrant \( j \) at TML \( i \) \((i = 1, 2, \ldots, 22; k = 1, 2, \ldots, 7; j = 1, \ldots, 4) \) and \( \epsilon_{ijk} \), as before, is the measurement error term. Similarly to model (3.1), \( \beta_{1ij} \) is also positive in model (3.2). We assume that \( \beta_{1ij} \) follows a lognormal distribution [i.e., \( \beta_{1ij} \sim \text{Lognormal} \left( \mu_{\beta_1j}, \sigma_{\beta_1j}^2 \right) \)] and \( \epsilon_{ijk} \sim \text{NOR} \left( 0, \sigma_{\epsilon}^2 \right) \). The parameters in model (3.2) are: 
\[ \theta = (\mu_{\beta_11}, \mu_{\beta_12}, \mu_{\beta_13}, \mu_{\beta_14}, \sigma_{\beta_1}, \sigma_{\epsilon})' \]. The Bayesian method is again used to estimate \( \theta \). Table 3.2 presents marginal posterior distribution summaries for the parameters in \( \theta \), including the mean and 95% credible intervals. Figure 3.4 shows the time plot of the fitted thickness values for different quadrants of this circuit.

The deviance information criterion (DIC) (defined in Gelman et al. 2003 on page 182-184), a measure of model goodness-of-fit and complexity, is used for the Bayesian model comparison. The values of DIC for models (3.1) and (3.2) are \(-2574.0\) and \(-2633.0\), respectively. Because
### Table 3.2 Marginal posterior distribution summaries of the degradation model parameter estimates for pipeline data from Circuit G in Facility 3 using the degradation model (3.2).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Posterior Mean</th>
<th>Posterior Std. Dev.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{\beta_1}$</td>
<td>-11.31</td>
<td>0.1182</td>
<td>-11.54 -11.08</td>
</tr>
<tr>
<td>$\mu_{\beta_2}$</td>
<td>-11.67</td>
<td>0.1179</td>
<td>-11.90 -11.45</td>
</tr>
<tr>
<td>$\mu_{\beta_3}$</td>
<td>-11.59</td>
<td>0.1167</td>
<td>-11.82 -11.36</td>
</tr>
<tr>
<td>$\mu_{\beta_4}$</td>
<td>-11.92</td>
<td>0.1219</td>
<td>-12.16 -11.68</td>
</tr>
<tr>
<td>$\sigma_{\beta_1}$</td>
<td>0.5416</td>
<td>0.04551</td>
<td>0.4619 0.6377</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>0.006033</td>
<td>2.556E-4</td>
<td>0.005556 0.006551</td>
</tr>
</tbody>
</table>

Model (3.2) has an importantly smaller DIC than model (3.1), we can conclude that there is a quadrant effect.

### 3.3 Models Relating Degradation and Pipeline Failure in Circuit G of Facility 3

#### 3.3.1 Bayesian Evaluation of the Failure Time Distribution

The degradation path over time is $D = D(t, \theta)$. The failure of an individual segment in a pipeline happens when the remaining pipeline thickness is less than the critical level $D_f$ (0.05 inches in our example). Such critical levels are determined through engineering judgment as the thickness below which there is risk of a leak. Because $\beta_{1ij} \sim \text{Lognormal} (\mu_{\beta_1j}, \sigma_{\beta_1}^2)$ in model (3.2), the failure time distribution $F(t)$ of individual segments in a population of segments of quadrant $j$ in the pipeline can be expressed in a closed form:
Figure 3.4  Time plot showing the fitted thickness values for different quadrants of pipeline data from Circuit G in Facility 3 using the degradation model (3.2).

\[
F(t) = \Pr(D(t) \leq D_f) \\
= \Pr(y_0 - \beta_{1_{ij}}(t_k - t_0) \leq 0.05) \\
= \Pr\left(\beta_{1_{ij}} \geq \frac{0.20}{t_k - t_0}\right) \\
= 1 - \Phi_{\text{nor}}\left(\frac{\log(0.20) - \log(t_k - t_0) - \mu_{\beta_{1_{ij}}}}{\sigma_{\beta_1}}\right) \\
= \Phi_{\text{nor}}\left(\frac{\log(t_k - t_0) - \log(0.20) + \mu_{\beta_{1_{ij}}}}{\sigma_{\beta_1}}\right).
\] (3.3)

where \(\Phi_{\text{nor}}\) is the standard normal cdf.
The failure time distribution, as a function of the degradation parameters, can be evaluated simply by using the Bayesian approach. For each draw from the joint posterior distribution, one can evaluate $F(t)$ in (3.3) to obtain draw from the posterior function of failure time distribution. Table 3.2 and Figure 3.4 suggest that the corrosion rate of quadrant 1 from the upper quadrant is the largest among these four different quadrants. Figure 3.5 (a) displays the estimate of the failure time distribution with two-sided 95% and 80% credible intervals for the pipeline data from quadrant 1 of Circuit G in Facility 3. One can also obtain the corresponding failure time distribution plots for other quadrants. But with the largest corrosion rate, the failure time plot for quadrant 1 is the most pessimistic. The failure time distribution we estimated using the Bayesian approach is the failure time for an individual pipeline segment. Although the primary interest is to estimate the lifetime of a pipeline viewed as a series system of many segments, the life time distribution of an individual pipeline segment provides useful information to plan the construction of future pipelines.

![Figure 3.5](image.png)

Figure 3.5 Degradation model estimates of (a) failure time distribution (years since pipeline installation) and (b) remaining lifetime distribution (years since the last inspection $t_c$) with two-sided 95% and 80% credible intervals on the lognormal paper for pipeline data from quadrant 1 of Circuit G in Facility 3.
3.3.2 Prediction of the Remaining Life of the Current Circuit

In the pipeline application, the remaining life of a particular segment of a circuit is an important quantity for assessing the lifetime of the pipeline. The distribution of the remaining lifetime $F_{RM}(t)$ conditional on surviving until the last inspection time (January 2003) is

$$F_{RM}(t) = \Pr(T \leq t | T > t_c) = \frac{F(t; \theta) - F(t_c; \theta)}{1 - F(t_c; \theta)}, \quad t \geq t_c$$

(3.4)

where $t_c$ is the last inspection time and $F(t)$ is the failure time distribution derived in Section 3.3.1. As before, evaluating (3.4) at posterior draws provides estimates and the corresponding credible intervals of the remaining lifetime distribution. Figure 3.5 (b) shows the posterior estimates of the remaining lifetime distribution after the last inspection in January 2003 with 95% and 80% credible intervals.

In the pipeline application, it is of great interest to estimate small quantiles of the minimum remaining lifetime of the population. To do this, one needs to extrapolate further into the tail of the remaining life distribution estimated for a given segment. Typically a TML segment is about one foot long. Suppose that the entire pipeline length has $M$ segments of this length. Then the distribution of the minimum remaining life among all of the $M$ segments along the pipeline can be expressed as

$$F_M(t) = \Pr[T_{min} \leq t] = 1 - [1 - F_{RM}(t)]^M$$

(3.5)

where $F_{RM}(t)$ is the remaining lifetime distribution for a single segment. If one wants to control $F_M(t)$, such that $F_M(t) = \Pr[T_{min} \leq t] = p$, then one would choose the threshold to be $t_p = F_M^{-1}(p)$, the $p$ quantile of the distribution of the minimum $T_{min}$ among the $M$ pipeline segments. The translation to the adjusted quantile in terms of the remaining lifetime distribution $F_{RM}(t)$ is as follows:

$$t_p = F_M^{-1}(p) = F_{RM}^{-1} \left(1 - (1 - p)^\frac{1}{M}\right).$$

(3.6)

This indicates that $p$ quantile of the minimum remaining lifetime distribution of the population of $M$ segments corresponds to the $1 - (1 - p)^{1/M}$ quantile of the remaining lifetime distribution.
for each segment. Figure 3.6 shows the posterior density of 0.1, 0.2, 0.3, and 0.4 quantiles of the minimum remaining lifetime distribution with the population size $M = 100$ using the degradation models (3.1) and (3.2) respectively. Model (3.2) is more conservative than model (3.1) as it generates the smaller quantile estimates.

Figure 3.6 Posterior density of the 0.1, 0.2, 0.3 and 0.4 quantiles of the minimum remaining lifetime distribution (years since the last inspection time $t_c$: January 2003) with the population size $M = 100$ of pipeline data from Circuit G in Facility 3 using the degradation models (3.1) and (3.2).

The small quantile estimates suggest that the Circuit G in Facility 3 could have leakage risks within one year after the last inspection. One should pay closer attention to this circuit. Careful examination, more frequent inspection at more TMLs, or retirement/replacement of the pipeline would protect against the unexpected pipeline leakage.
3.4 Modeling Pipeline Data from Circuit Q in Facility 1

Figure 3.7 is a trellis plot for the pipeline data from Circuit Q in Facility 1. Each panel of the trellis plot corresponds to thickness measurements for a specific TML. The trellis plot suggests an interesting pipeline corrosion process. For example, in the TML #1, #2, and #3, there is no detectable thickness loss in the first three inspections. Significant thickness losses, however, were detected at the fourth inspection time. This suggests that the corrosion process was initiated between the third and fourth inspection times. At some TMLs (e.g., TMLs #12, #13, and #33), the corrosion appears not to have initiated before the last inspection time.

![Trellis plot for pipeline data from Circuit Q in Facility 1.](image)

3.4.1 Degradation Model for Corrosion Initiation and Growth

We assume that after the corrosion initiation, the corrosion rate is constant for a particular location, but may differ from location to location. We propose a degradation model with a random corrosion initiation time and random corrosion rate to describe the overall corrosion
initiation and growth process. The degradation model for the pipeline thickness $Y_{it_j}$ at time $t_j$ for the TML $i$ ($i = 1, 2, \ldots, 33; j = 1, 2, \ldots, 4$) is:

$$Y_{it_j} = \begin{cases} Y_{0i} + \epsilon_{ij} & \text{for } t_j < T_{I_i} \\ Y_{0i} - \beta_{1i}(t_j - T_{I_i}) + \epsilon_{ij} & \text{for } t_j \geq T_{I_i}. \end{cases}$$ \hspace{1cm} (3.7)

In this model,

- $Y_{it_j}$ denotes the thickness measurement for TML $i$ at time $t_j$.
- $Y_{0i}$ is the original thickness of TML $i$. Because the distribution of the original thickness depends on the component type of the TML (elbow, tee, or straight pipe), we assume that the initial measurement $Y_{0i}$ follows a normal distribution with different means but a common standard deviation:
  - If the TML is an elbow, we assume that $Y_{0i} \sim \text{NOR} (\mu_{y0_{elbow}}, \sigma_{y0}^2)$;
  - If the TML is a pipe, we assume that $Y_{0i} \sim \text{NOR} (\mu_{y0_{pipe}}, \sigma_{y0}^2)$;
  - If the TML is a tee, we assume that $Y_{0i} \sim \text{NOR} (\mu_{y0_{tee}}, \sigma_{y0}^2)$.
- $\beta_{1i}$ is the corrosion rate for TML $i$ and we assume that $\beta_{1i} \sim \text{Lognormal} (\mu_{\beta1}, \sigma_{\beta1}^2)$ or $\beta_{1i} \sim \text{Weibull} (\nu_{\beta1}, \lambda_{\beta1})$.
- $T_{I_i}$ is the corrosion initiation time at TML $i$ and we assume that $T_{I_i} \sim \text{Lognormal} (\mu_{T_{I}}, \sigma_{T_{I}}^2)$.
- $\epsilon_{ij}$ is the measurement error and we assume that $\epsilon_{ij} \sim \text{NOR} (0, \sigma_{\epsilon}^2)$.
- $t_j$ is the time when the measurements were taken.

The model parameters are: $\theta = (\mu_{y0_{elbow}}, \mu_{y0_{pipe}}, \mu_{y0_{tee}}, \sigma_{y0}, \mu_{\beta1}, \sigma_{\beta1}, \mu_{T_{I}}, \sigma_{T_{I}}, \sigma_{\epsilon})'$ for the lognormal corrosion rate. When the corrosion rate follows a Weibull distribution, the model parameters are: $\theta = (\mu_{y0_{elbow}}, \mu_{y0_{pipe}}, \mu_{y0_{tee}}, \sigma_{y0}, \nu_{\beta1}, \lambda_{\beta1}, \mu_{T_{I}}, \sigma_{T_{I}}, \sigma_{\epsilon})'$.

### 3.4.2 Bayesian Estimation of the Parameters in the Degradation Model

In addition to the model, we need to specify prior distributions for the parameters in the degradation model (3.7). Gelman (2006) provided general suggestions for choosing proper prior
distributions for variance parameters in the hierarchical model. We use the following diffuse prior distributions for the standard deviations $\sigma_y$, $\sigma_\beta$, $\sigma_T$, and $\sigma_\epsilon$:

$$\sigma_y \sim \text{Uniform } (10^{-5}, 5),$$

$$\sigma_T \sim \text{Uniform } (10^{-5}, 10),$$

$$\sigma_\beta \sim \text{Uniform } (10^{-5}, 5),$$

$$\sigma_\epsilon \sim \text{Uniform } (10^{-5}, 0.25).$$

The fact that pipeline data of Circuit Q in Facility 1 has no more than one inspection after the corrosion initiation results in difficulty identifying the corrosion rate and initiation times in the degradation model. According to the knowledge from experts in the pipeline application, the corrosion rates of the TMLs fall into a certain range. Thus we specify informative prior distribution for the median of corrosion rates for TMLs $\beta_{1i0.5}$:

$$\beta_{1i0.5} \sim \text{Uniform } (10^{-6}, 0.022).$$

Here, the bounds of the uniform prior for $\beta_{1i0.5}$ are used such that the median of the $\beta_{1i}$ [i.e. $\exp(\mu_{\beta_i})$] falls between $10^{-6}$ and 0.022. Regarding the prior distributions for the parameters $\mu_{y\text{elbow}}$, $\mu_{y\text{pipe}}$, $\mu_{y\text{tee}}$, and $\mu_T$, we use the following priors by specifying the lower and upper bounds of the uniform distributions:

$$\mu_{y\text{elbow}} \sim \text{Uniform } (0.4, 0.47),$$

$$\mu_{y\text{pipe}} \sim \text{Uniform } (0.4, 0.47),$$

$$\mu_{y\text{tee}} \sim \text{Uniform } (0.5, 0.62),$$

$$\mu_T \sim \text{Uniform } (9.31, 10^6).$$

The lower bound of the uniform distribution for $\mu_T$ is determined by the assumption that the corrosion initiation can only occur after the installation. Similarly, if the corrosion rate follows a Weibull distribution, we specify the same independent prior distributions for the parameters $\sigma_y$, $\sigma_T$, $\sigma_\epsilon$, $\mu_{y\text{elbow}}$, $\mu_{y\text{pipe}}$, $\mu_{y\text{tee}}$, and $\mu_T$. For the Weibull corrosion rate distribution, we specify the prior distribution in terms of $\nu_{\beta_1}$, the Weibull shape parameter and $\beta_{1i0.5}$, the median
of corrosion rates for TMLs. The following are priors for the shape parameter $\nu_{\beta_1}$ and $\beta_{1i_0.5}$ in the Weibull distribution:

$$\nu_{\beta_1} \sim \text{Uniform}(1.5, 5),$$

$$\beta_{1i_0.5} \sim \text{Uniform}(10^{-6}, 0.022),$$

where $\lambda_{\beta_1} = \log_e(2)/\beta_{1i_0.5}$ is the alternative second parameter used in the OpenBUGS parameterization of the Weibull distribution. Tables 3.3 and 3.4 present the posterior distribution summaries of parameters in the degradation model using lognormal and Weibull corrosion rates respectively. Figures 3.8 and 3.9 show the trellis plot of the fitted thickness values for Circuit Q in Facility 1 using lognormal and Weibull corrosion rates with 10-years extrapolation after the last inspection on January 1, 2004.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Posterior Mean</th>
<th>Posterior Std. Dev.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{y_{0\text{elbow}}}$</td>
<td>0.4379</td>
<td>0.004035</td>
<td>0.4301 to 0.4461</td>
</tr>
<tr>
<td>$\mu_{y_{0\text{pipe}}}$</td>
<td>0.4315</td>
<td>0.003448</td>
<td>0.4246 to 0.4382</td>
</tr>
<tr>
<td>$\mu_{y_{0\text{tee}}}$</td>
<td>0.5215</td>
<td>0.006226</td>
<td>0.5093 to 0.5336</td>
</tr>
<tr>
<td>$\sigma_{y_0}$</td>
<td>0.01342</td>
<td>0.00192</td>
<td>0.01022 to 0.01767</td>
</tr>
<tr>
<td>$\beta_{1i_0.5}$</td>
<td>0.01815</td>
<td>0.003171</td>
<td>0.01804 to 0.02187</td>
</tr>
<tr>
<td>$\sigma_{\beta_1}$</td>
<td>0.5902</td>
<td>0.3129</td>
<td>0.1877 to 1.391</td>
</tr>
<tr>
<td>$\mu_{T_I}$</td>
<td>9.411</td>
<td>0.0122</td>
<td>9.389 to 9.438</td>
</tr>
<tr>
<td>$\sigma_{T_I}$</td>
<td>0.04408</td>
<td>0.01536</td>
<td>0.01804 to 0.07968</td>
</tr>
<tr>
<td>$\sigma_{\epsilon}$</td>
<td>0.004693</td>
<td>4.005E-4</td>
<td>0.003973 to 0.005552</td>
</tr>
</tbody>
</table>

Table 3.3 Marginal posterior distribution summaries of the parameters in the degradation model with lognormal corrosion rate for pipeline data from Circuit Q in Facility 1.

The deviance information criterion (DIC) is again used for the Bayesian model comparison. DIC values for models with lognormal and Weibull corrosion rates are $-1513.0$ and $-1016.0$, respectively. The model using the lognormal distribution for the corrosion rate has a smaller DIC, indicating a better fit. Figures 3.10 and 3.11 show the box plots of samples from the marginal posterior distributions of corrosion rates and initiation times for each TML in Circuit Q using lognormal corrosion rate. These plots indicate that for the TMLs where pipeline corrosion appears not to have initiated before the last inspection time, the posterior distribution of the initiation times is right skewed. Figure 3.12 compares plots of the marginal posterior
Table 3.4 Marginal posterior distribution summaries of the parameters in the degradation model with Weibull corrosion rate for pipeline data from Circuit Q in Facility 1.

distributions of the initiation times for TMLs with evidence of initiation and without initiation before the last inspections. There plots show that the marginal posterior distributions of the initiation times for the TMLs without initiation are right skewed and close to each other.

3.5 Models Relating Degradation and Failure of Pipeline Data from Circuit Q in Facility 1

3.5.1 Bayesian Evaluation of the Failure Time Distribution

As in the analysis of the pipeline data from Circuit G in Facility 3, there are two main purposes for using the degradation model. The first is to assess the lifetime distribution of individual pipeline components or segments. The second is to predict the remaining lifetime of the entire circuit. The degradation path over time is $D = D(t, \theta)$. A soft failure is defined to be the time at which the remaining pipeline thickness is less than 20% of the mean of the thickness at the installation date. Suppose that $T_I \sim \text{Lognormal} (\mu_{T_I}, \sigma_{T_I}^2)$, $Y_0 \sim \text{NOR} (\mu_{y_0_{\text{elbow}}}, \sigma_{y_0}^2)$, $Y_0 \sim \text{NOR} (\mu_{y_0_{\text{pipe}}}, \sigma_{y_0}^2)$, $Y_0 \sim \text{NOR} (\mu_{y_0_{\text{tee}}}, \sigma_{y_0}^2)$, and $\beta_1 \sim \text{Lognormal} (\mu_{\beta_1}, \sigma_{\beta_1}^2)$. Then the cumulative distribution giving the proportion of pipeline segments as a function of operating time is

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Posterior Mean</th>
<th>Posterior Std. Dev.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{y_0_{\text{elbow}}}$</td>
<td>0.4380</td>
<td>0.004059</td>
<td>0.4299</td>
</tr>
<tr>
<td>$\mu_{y_0_{\text{pipe}}}$</td>
<td>0.4314</td>
<td>0.003478</td>
<td>0.4245</td>
</tr>
<tr>
<td>$\mu_{y_0_{\text{tee}}}$</td>
<td>0.5215</td>
<td>0.006039</td>
<td>0.5096</td>
</tr>
<tr>
<td>$\sigma_{y_0}$</td>
<td>0.01347</td>
<td>0.001942</td>
<td>0.01027</td>
</tr>
<tr>
<td>$\nu_{\beta_1}$</td>
<td>2.890</td>
<td>0.9342</td>
<td>1.556</td>
</tr>
<tr>
<td>$\beta_{1_{\text{init}}}$</td>
<td>0.0197</td>
<td>0.001908</td>
<td>0.01488</td>
</tr>
<tr>
<td>$\mu_{T_I}$</td>
<td>9.412</td>
<td>0.01221</td>
<td>9.392</td>
</tr>
<tr>
<td>$\sigma_{T_I}$</td>
<td>0.04744</td>
<td>0.01431</td>
<td>0.02657</td>
</tr>
<tr>
<td>$\sigma_e$</td>
<td>0.004715</td>
<td>4.016E−4</td>
<td>0.004007</td>
</tr>
</tbody>
</table>
Figure 3.8 Trellis plot of the fitted thickness values for Circuit Q in Facility 1 using the lognormal corrosion rate distribution. The dotted lines indicate extrapolation.

\[ F(t) = \Pr(D(t) \leq D_f) \]

\[ = \Pr(Y_0 - \beta_1(t - T_I)I(t \geq T_I) \leq D_f) \]

\[ = \Pr(Y_0 - \beta_1(t - T_I) \leq D_f \cap t \geq T_I) + \Pr(Y_0 \leq D_f \cap t < T_I) \]

\[ = \Pr(Y_0 - \beta_1(t - T_I) \leq D_f \cap t \geq T_I) + \Pr(Y_0 \leq D_f)\Pr(t < T_I) \]

\[ = \int \int \int_{y_0 - \beta_1(t - T_I) \leq D_f, t \geq T_I} \frac{1}{\sigma_{y_0}} \phi_{\text{NOR}}(z_{y_0}) \times \frac{1}{T_I \sigma_{T_I}} \phi_{\text{NOR}}(z_{T_I}) \times \frac{1}{\beta_1 \sigma_{\beta_1}} \phi(z_{\beta}) dy_0 dT_I d\beta_1 \]

\[ + \Phi_{\text{NOR}} \left( \frac{D_f - \mu_{y_0}}{\sigma_{y_0}} \right) \times [1 - \Phi_{\text{NOR}}(z_{T_I})], \]

where \( z_{y_0} = (y_0 - \mu_{y_0})/\sigma_{y_0} \), \( z_{T_I} = (\log(T_I) - \mu_{T_I})/\sigma_{T_I} \), and \( z_{\beta} = (\log(\beta_1) - \mu_{\beta_1})/\sigma_{\beta_1} \). With the lognormal corrosion rate, \( \phi(z_{\beta}) = \phi_{\text{NOR}}(z_{\beta}) \) is the standard \((\mu = 0, \sigma = 1)\) normal probability density function (pdf). When the corrosion rate follows a Weibull distribution, \( \phi(z_{\beta}) = \phi_{\text{SEV}}(z_{\beta}) = \exp(z_{\beta} - \exp(z_{\beta})) \) is the standardized smallest extreme value pdf. Because
Table 3.2. Fitted Wall Thickness for Elbow, Pipe and Tee Segments

<table>
<thead>
<tr>
<th>Date</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ap</td>
<td>1Jan04</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nov 2000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jul 2001</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jan 2004</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 3.9  Trellis plot of the fitted thickness values for Circuit Q in Facility 1 using the Weibull corrosion rate distribution. The dotted lines indicate extrapolation.

\[ F(t) \text{ in (3.8)} \] does not have a closed form, the Monte Carlo simulation method described in Section 13.5.3 of Meeker and Escobar (1998) is used to evaluate failure time distributions, using 1,000 simulation trials for the evaluation. Figure 3.13 shows failure time distributions for elbows, pipes and tees using normal, lognormal and Weibull corrosion rates. The plots suggest that in the degradation model (3.7), the lognormal corrosion rate provides the most conservative results compared with the other two model assumptions on the corrosion rate. Figure 3.14 shows failure time distributions for elbow, pipe and tee segments using the lognormal corrosion rate distribution with two-sided 95% and 80% credible intervals.

### 3.5.2 Predication of the Remaining Life of the Current Circuit

Figure 3.15 compares the remaining lifetime distributions with normal, lognormal and Weibull corrosion rates for elbows, pipes and tees. The plots suggest that a lognormal distribution for the corrosion rate in the degradation model (3.7) provides the most conservative
estimates. This is due to the long upper tail of the lognormal distribution. Figure 3.16 shows estimates of the remaining lifetime distributions using the lognormal corrosion rate in the degradation model (3.7) and the corresponding two-sided 95% and 80% credible intervals.

As in Section 3.3.3, we are primarily interested in estimating small quantiles of the minimum remaining lifetime distribution for Circuit Q in Facility 1. Figure 3.17 shows the posterior density of 0.1, 0.2, 0.3, and 0.4 quantiles of the minimum remaining lifetime distribution from the degradation model (3.7) with the population size \( M = 100 \) using the lognormal distribution for corrosion rate. The larger quantile estimates for the tee components indicate that tees have a longer remaining lifetime. The results are consistent with what we observed previously in Figures 3.14 and 3.16.
3.6 Effect of Additional Inspections on Identifiability

In Section 3.4.3, in the analysis of the pipeline data from the Circuit Q in Facility 1, we used a moderately informative prior distribution to describe prior knowledge about the median of the corrosion rates, alleviating the identifiability problem that was caused by having no more than one inspection after any of the observed corrosion initiation events. The results of that analysis showed a large amount of uncertainty in predictions of remaining life. To investigate this identifiability problem, in this section, we simulate data from model (3.7) such that there is more than one inspection after an initiation (i.e., data that is similar to those from Circuit Q in Facility 1 but with additional inspections). We continue to use a lognormal corrosion rate distribution. Figure 3.18 displays the time plot for the simulated pipeline data from a single circuit with 33 TMLs and three components: elbow, straight pipe and tee pipe. Corrosion was measured at each TML at 5 times.
Figure 3.12  Posterior densities of the initiation times for each TML in Circuit Q in Facility 1.

We use the same diffuse prior distributions used in Section 3.4.2 for all parameters except for the median of the corrosion rates $\beta_{1,0.5}$, $i = 1, 2, \ldots, 33$. Because there is more than one inspection after the corrosion initiation in the simulated data, the identifiability problem no longer exists. Therefore, rather than restrict the upper bound of the prior distribution of $\beta_{1,0.5}$ to 0.022, we can relax the upper bound to 0.10 providing a diffuse prior for $\beta_{1,0.5}$ [i.e. $\beta_{1,0.5} \sim \text{Uniform} \ (10^{-6}, 0.10)$]. For these simulated data, the Bayesian parameter estimates are close to the true parameter values from which the data were simulated. Figure 3.19 shows the trellis plot of the fitted thickness values for the simulated pipeline data using the diffuse prior distributions.

As in Section 3.5, we used the Monte Carlo simulation method to evaluate the marginal posterior distributions of the failure time distribution at chosen points in time. Figure 3.20 shows the failure time distributions for the simulated pipeline data of a single circuit by using the diffuse priors. Compared with the results in Figure 3.14 for the pipeline data from Circuit
Q in Facility 1, the credible intervals in Figure 3.20 are much narrower. The reason is that with more inspections after the corrosion initiation in the simulated data, the identifiability problem that caused the wide intervals is no longer present. From a practical perspective, having several inspections that occur after an initiation time provides a more effective estimation of pipeline segment lifetime distributions.

3.7 Hierarchical Model with Corrosion Initiation and Growth Along a Pipeline

The pipeline data we analyzed in Sections 3.4 and 3.5 were from a particular circuit in Facility 1. In real applications, there will often be several similar parallel circuits in a facility having the application(s) and that would therefore be expected to have similar corrosion behavior. The corrosion rates of TMLs within a circuit can be expected to follow a common distribution with a fixed location parameter. The location parameters of the distributions for the corrosion rates however could vary somewhat across different circuits due to unrecorded factors such as raw materials variability or variability in operational factors such as temperature or pressure. It could also be that the similar circuits are in different facilities. Here we extend model in (3.7) to develop a hierarchal model to describe corrosion behavior for several circuits. Because such multiple circuit data were not available to us, we use simulated data to illustrate the use of the model. As an extension to (3.7), we use

\[
Y_{ijt_k} = \begin{cases} 
Y_{0ij} + \epsilon_{ijk} & \text{for } t_k < T_{Iij} \\
Y_{0ij} - \beta_{1ij}(t_k - T_{Iij}) + \epsilon_{ijk} & \text{for } t_k \geq T_{Iij}
\end{cases}
\]

(3.9)

where

- \(Y_{ijt_k}\) denotes the thickness measurement for TML \(j\) in circuit \(i\) at time \(t_k\) \((i = 1, 2, \ldots, 12; j = 1, 2, \ldots, 33; k = 1, 2, \ldots, 5)\).

- \(Y_{0ij}\) is the original thickness of TML \(j\) in circuit \(i\). Because the original thickness depends on the component type of the TML, we assume that \(Y_{0ij}\) follows a normal distribution with different means but a common standard deviation:
– If the TML is an elbow, we assume that $Y_{0ij} \sim \text{NOR} \left( \mu_{y0\text{elbow}}, \sigma_{y0}^2 \right)$;

– If the TML is a pipe, we assume that $Y_{0ij} \sim \text{NOR} \left( \mu_{y0\text{pipe}}, \sigma_{y0}^2 \right)$;

– If the TML is a tee, we assume that $Y_{0ij} \sim \text{NOR} \left( \mu_{y0\text{tee}}, \sigma_{y0}^2 \right)$.

\begin{itemize}
  \item $\beta_{1ij}$ is the corrosion rate for TML $j$ in circuit $i$ and we assume that $\beta_{1ij} \sim \text{Lognormal} \left( \mu_{\beta_1}, \sigma_{\beta_1}^2 \right)$.
  \item The mean of the logarithm of the corrosion rates differ from circuit to circuit and the mean rate in circuit $i$ has a distribution $\mu_{\beta_1i} \sim \text{NOR} \left( \mu_{\beta_1}, \sigma_{\beta_1}^2 \right)$.
  \item $T_{I_{ij}}$ is the corrosion initiation time at TML $j$ in circuit $i$ and $T_{I_{ij}} \sim \text{Lognormal} \left( \mu_{T_I}, \sigma_{T_I}^2 \right)$.
  \item $\epsilon_{ijk}$ is the measurement error and $\epsilon_{ijk} \sim \text{NOR} \left( 0, \sigma_{\epsilon}^2 \right)$.
  \item $t_k$ is the time when the measurements at inspection $k$ were taken.
\end{itemize}

The model parameters are: $\theta = (\mu_{y0\text{elbow}}, \mu_{y0\text{pipe}}, \mu_{y0\text{tee}}, \sigma_{y0}, \mu_{\beta}, \sigma_{\beta}, \sigma_{\beta_1}, \mu_{T_I}, \sigma_{T_I}, \sigma_{\epsilon})'$. Figure 3.21 shows time plots of the simulated pipeline data from Circuits 1 to 4 in the same Facility.

Table 3.5 presents the marginal posterior distribution summaries for the model parameters using the simulated pipeline thickness data with 12 circuits and 33 TMLs within each circuit. We use diffuse priors in the Bayesian analysis. For example, we use the normal distribution with mean zero and a large variance [i.e, NOR $(0, 10^3)$] for the parameters $\mu_{y0\text{elbow}}, \mu_{y0\text{pipe}}, \mu_{y0\text{tee}}, \mu_{T_I}$. The prior distributions for $\sigma_{y0}, \sigma_{\beta}, \sigma_{\beta_1}, \sigma_{T_I}$ and $\sigma_{\epsilon}$ are Uniform $(0, 5)$. The results in Table 3.5 suggest that with the diffuse priors, the Bayesian parameter estimates are close to the true values. Although the true value of $\sigma_{\beta}$ falls into the 95% credible interval of $\hat{\sigma}_{\beta}$, the Bayesian approach slightly overestimates $\sigma_{\beta}$.

Figure 3.22 shows the box plots of the corrosion rate estimates for the 33 TMLs within each of the 12 circuits. Because $\sigma_{\beta}$ is almost two-times greater than $\sigma_{\beta_1}$, the within circuit TMLs corrosion rates variability is less than the variability of medians of TMLs corrosion rates across circuits.

### 3.8 Concluding Remarks and Areas for Future Research

In this paper, we developed degradation models to describe the pipeline corrosion behaviors for two particular pipeline data set. The Bayesian approach with appropriate prior distributions
<table>
<thead>
<tr>
<th>Parameters</th>
<th>True Value</th>
<th>Posterior Mean</th>
<th>Posterior Std. Dev.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{y_{\text{elbow}}}$</td>
<td>0.4392</td>
<td>0.4471</td>
<td>0.003769</td>
<td>0.4391 - 0.4545</td>
</tr>
<tr>
<td>$\mu_{y_{\text{pipe}}}$</td>
<td>0.4313</td>
<td>0.4344</td>
<td>0.003280</td>
<td>0.4279 - 0.4408</td>
</tr>
<tr>
<td>$\mu_{y_{\text{tee}}}$</td>
<td>0.5350</td>
<td>0.5337</td>
<td>0.005854</td>
<td>0.5222 - 0.5451</td>
</tr>
<tr>
<td>$\sigma_{y_0}$</td>
<td>0.04349</td>
<td>0.04518</td>
<td>0.001630</td>
<td>0.04212 - 0.04851</td>
</tr>
<tr>
<td>$\mu_{\beta}$</td>
<td>-3.730</td>
<td>-3.709</td>
<td>0.04546</td>
<td>-3.799 - -3.619</td>
</tr>
<tr>
<td>$\sigma_{\beta}$</td>
<td>0.1309</td>
<td>0.1523</td>
<td>0.03896</td>
<td>0.09712 - 0.2474</td>
</tr>
<tr>
<td>$\sigma_{\beta_1}$</td>
<td>0.07130</td>
<td>0.07032</td>
<td>0.003849</td>
<td>0.06315 - 0.07813</td>
</tr>
<tr>
<td>$\mu_{T_I}$</td>
<td>7.091</td>
<td>7.100</td>
<td>0.01923</td>
<td>7.062 - 7.138</td>
</tr>
<tr>
<td>$\sigma_{T_I}$</td>
<td>0.3539</td>
<td>0.3686</td>
<td>0.01436</td>
<td>0.3417 - 0.3976</td>
</tr>
<tr>
<td>$\sigma_{\epsilon}$</td>
<td>0.004460</td>
<td>0.004476</td>
<td>1.096E-4</td>
<td>0.004267 - 0.004697</td>
</tr>
</tbody>
</table>

Table 3.5 Marginal posterior distribution summaries of the degradation model parameter estimates of the simulated pipeline data with 12 circuits and 33 TMLs within each circuit in Facility 1.

is useful for estimating parameters in the degradation models. The Bayesian method, as an alternative to the likelihood approach, provides a convenient method to estimate and compute credible bounds for functions of the degradation model parameters, even when a closed-form expression of the function does not exist. The failure time and the remaining lifetime distributions and small quantile estimates of the minimum remaining lifetime distribution provide useful information to evaluate the life of a pipeline. There remains, however, a number of areas for future research. These include:

- In the degradation model for corrosion initiation and growth, test planning methods (see Section 9.6 of Hamada et al. 2008) could be developed to choose an appropriate number of inspections after the corrosion initiation to obtain more precise estimate of the failure time distribution.

- The model with linear degradation paths and the constant corrosion rate can be extended to the models having nonlinear relationships between pipeline thickness and time.

- Each pipeline circuit within a facility, viewed as a series system of many segments, could be considered as a component in a complex system. In some applications, the life time of the pipeline system could be particularly important.
In some pipeline application, it may be possible to obtain dynamic covariate information such as temperature, flow, and type of material. The degradation models could then be generalized by incorporating this dynamic covariate information into the modeling and analysis.

3.9 Appendix: OpenBUGS Code for the Hierarchical Model

model
{
  for (k in 1: NumCircuit) {
    for(i in 1 : N) {
      thickness.init.elbow[(k-1)*N+i] ~ dnorm(mu.y0.elbow, tau.y0)
      thickness.init.pipe[(k-1)*N+i] ~ dnorm(mu.y0.pipe, tau.y0)
      thickness.init.tee[(k-1)*N+i] ~ dnorm(mu.y0.tee, tau.y0)
      thickness.init[(k-1)*N+i] <- thickness.init.elbow [(k-1)*N+i]*step(N.elbow-i)+
        thickness.init.pipe [(k-1)*N+i]*step(N.pipe+N.elbow-i)*step(i-N.elbow-1) +
        thickness.init.tee [(k-1)*N+i]*step(i-N.elbow-N.pipe-1)
      log.init [(k-1)*N+i] ~ dnorm(mu.TI, tau.TI)
      init[(k-1)*N+i] <- exp(log.init [(k-1)*N+i])
      log.beta[(k-1)*N+i] ~ dnorm(mu.beta[k], tau.beta)
      beta[(k-1)*N+i] <- exp(log.beta[(k-1)*N+i])
      for (j in 1: T) {
        Y[(k-1)*N+i, j] ~ dnorm (mu[(k-1)*N+i,j], tau)
        mu[(k-1)*N+i,j] <- thickness.init[(k-1)*N+i] - beta[(k-1)*N+i] * 
          step(time[j]-init[(k-1)*N+i]) * (time[j]-init[(k-1)*N+i])/365
      }
    }
  }
  mu.beta[k] ~ dnorm(mu.beta0, tau.beta0)
}

# prior distribution for mu.y0
mu.y0.elbow ~ dnorm(0.0, 1.0E-6)
mu.y0.pipe ~ dnorm(0.0, 1.0E-6)
mu.y0.tee ~ dnorm(0.0, 1.0E-6)
# prior distribution for sigma.y0
sigma.y0 ~ dunif(0.00001, 5)
sigma.y0.sq <- pow(sigma.y0, 2)
tau.y0 <- 1/sigma.y0.sq
# prior distribution for mu.TI
mu.TI ~ dnorm(0, 1.0E-6)
# prior distribution for sigma.TI
sigma.TI ~ dunif(0.00001, 5)
sigma.TI.sq <- pow(sigma.TI, 2)
tau.TI <- 1/sigma.TI.sq
# prior distribution for sigma
sigma ~ dunif(0, 5)
sigma.sq <- pow(sigma, 2)
tau <- 1/sigma.sq
# Prior distribution for the corrosion rate
mu.beta0 <- dnorm(0.0, 1.0E-6)
sigma.beta0 ~ dunif(0, 5)
sigma.beta0.sq <- pow(sigma.beta0, 2)
tau.beta0 <- 1/sigma.beta0.sq
sigma.beta ~ dunif(0, 5)
sigma.beta.sq <- pow(sigma.beta, 2)
tau.beta <- 1/sigma.beta.sq
}

Figure 3.13  Degradation model estimates of failure time distributions for pipeline components from Circuit Q in Facility 1 comparing normal, lognormal and Weibull corrosion rate distributions in the degradation model (3.7).
Figure 3.14 Degradation model estimates (the center lines) of failure time distributions for pipeline components from Circuit Q in Facility 1 with the lognormal corrosion rate distribution in the degradation model (3.7) and two-sided 95% and 80% credible intervals.
Figure 3.15 Degradation model estimates of remaining lifetime distributions for pipeline components from Circuit Q in Facility 1 comparing normal, lognormal and Weibull corrosion rate distributions in the degradation model (3.7).
Figure 3.16  Degradation model estimates (the center lines) of remaining lifetime distributions for pipeline components from Circuit Q in Facility 1 with the lognormal corrosion rate distribution in the degradation model (3.7) and two-sided 95% and 80% credible intervals.
Figure 3.17 Posterior density of the 0.1, 0.2, 0.3, and 0.4 quantiles of the minimum remaining lifetime distribution (years after the last inspection time $t_c$: January 2004) with the population size $M = 100$ using the lognormal corrosion rate distribution of pipeline data from Circuit Q in Facility 1.
Figure 3.18  Time plot for the simulated pipeline data from a single circuit with 33 TMLs.
Figure 3.19  Trellis plot of the fitted thickness values for the simulated pipeline data in a single circuit with 33 TMLs using the diffuse prior distributions.
Figure 3.20  Degradation model estimates of failure time distributions for the simulated pipeline data in a single circuit with 33 TMLs using the lognormal corrosion rate distribution and diffuse priors in the degradation model (3.7) and two-sided 95% and 80% credible intervals.
Figure 3.21  Time plot for simulated pipeline data from Circuits 1 to 4 in the same Facility.

Figure 3.22  Box plots of the corrosion rates within each of the 12 circuits.
Bibliography


CHAPTER 4. UNDERSTANDING AND ADDRESSING THE UNBOUNDED “LIKELYHOOD” PROBLEM

A paper to be submitted
Shiyao Liu, Huaiqing Wu, William Q. Meeker
Department of Statistics
Iowa State University

Abstract

The joint probability density function, evaluated at the observed data, is commonly used as the likelihood function to compute maximum likelihood estimates. For some models, however, there exist paths in the parameter space along which this density-approximation likelihood goes to infinity and maximum likelihood estimation breaks down. In applications, all observed data are discrete due to the round-off or grouping error of measurements. The “correct likelihood” based on interval censoring can eliminate the problem of an unbounded likelihood. This paper categorizes the models leading to unbounded likelihoods into three groups and illustrates the density breakdown with specific examples. We also study the effect of the round-off error on estimation, and provide a sufficient condition for the joint density to provide the same maximum likelihood estimate as the correct likelihood, as the round-off error goes to zero.

Key Words: Density approximation; Interval censoring; Maximum likelihood; Round-off error; Unbounded likelihood.
4.1 Introduction

4.1.1 Background

Because of inherent limitations of measuring instruments, all continuous numerical data are subject to the round-off or grouping error of measurements. This has been described, for example, by Kempthorne (1966), Barnard (1967), Kempthorne and Folks (1971), Giesbrecht and Kempthorne (1976), Cheng and Iles (1987), and Vardeman and Lee (2005). For convenience, such discrete observations are often modeled on a continuous scale. Usually, when the round-off error is small, the likelihood for a sample of independent observations is defined as the product of the probability densities evaluated at each of the “exact” observations. For some models, however, such a likelihood may be unbounded along certain paths in the parameter space, causing numerical and statistical problems in maximum likelihood (ML) estimation. As has been suggested in the references above, using the correct likelihood based on small intervals (e.g., implied by the data’s precision) instead of the density approximation will eliminate the problem of an unbounded likelihood. Practitioners should know about the potential problems of an unbounded likelihood and how to use the correct likelihood to avoid the problems. The purpose of this paper is to review and consolidate previous results concerning this problem, to provide a classification of models that lead to an unbounded “likelihood,” and to present related theoretical results.

4.1.2 An Illustrative Example

Example 11.17 of Meeker and Escobar (1998) illustrates ML estimation for the three-parameter lognormal distribution using the diesel generator fan data given in Nelson (1982, page 133). The likelihood function based on the usual density approximation for exact failures at time $t_i, i = 1, \ldots, n$ has the form

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta; t_i) = \prod_{i=1}^{n} f(t_i; \theta),$$

(4.1)

where
Figure 4.1 Three-parameter lognormal profile log-likelihood plots of the threshold parameter $\gamma$ for the diesel generator fan data using (a) the unbounded density-approximation likelihood $L$ and (b) the correct likelihood $\mathcal{L}$ (with the round-off error $\Delta = 5$).

\[
f(t_i; \theta) = \frac{1}{\sigma(t_i - \gamma)} \phi_{nor} \left[ \frac{\log(t_i - \gamma) - \mu}{\sigma} \right] I_{(t_i > \gamma)}
\]

(4.2)

is the probability density function (pdf) of the three-parameter lognormal distribution and $\theta = (\mu, \sigma, \gamma)'$. Here $\phi_{nor}$ is the pdf for the standard normal distribution and $\exp(\mu), \sigma$ and $\gamma$ are the scale, shape and threshold parameters, respectively. As the threshold parameter $\gamma$ approaches the smallest observation $t_{(1)}$, the profile log-likelihood for $\gamma$ in Figure 4.1 (a) increases without bound (i.e., $L(\theta) \to \infty$), indicating the breakdown in the density approximation. For the diesel generator fan data, there is a local maximum that corresponds to the maximum of the correct likelihood. For some data sets, the local maximum is dominated by the unbounded behavior.

### 4.1.3 A Simple Remedy

The unboundedness of the likelihood leads to computational difficulty. As suggested in the references in Section 4.1.1, using the “correct likelihood” based on interval censoring instead of the density approximation will eliminate the problem of an unbounded likelihood. Because
probabilities can not be larger than 1, the correct likelihood based on small intervals (implied by the data’s precision) will always be bounded. The correct likelihood can be expressed as

\[ L(\theta) = \prod_{i=1}^{n} L_i(\theta; t_i) = \prod_{i=1}^{n} \frac{1}{2\Delta_i} \int_{t_i-\Delta_i}^{t_i+\Delta_i} f(x; \theta) \, dx = \prod_{i=1}^{n} \frac{1}{2\Delta_i} [F(t_i + \Delta_i; \theta) - F(t_i - \Delta_i; \theta)], \]  

(4.3)

where, for the example,

\[ F(t_i; \theta) = \Phi_{\text{nor}} \left[ \frac{\log(t_i - \gamma) - \mu}{\sigma} \right] I(t_i > \gamma) \]

is the three-parameter lognormal cumulative distribution function (cdf). Here \( \Phi_{\text{nor}} \) is the cdf for the standard normal distribution.

The values of \( \Delta_i \) reflect the round-off error in the data and may depend on the magnitude of the observations. For the diesel generator fan data, because the life times were recorded to a precision of \( \pm 5 \) hours, we choose \( \Delta_i = 5 \) for all of the \( t_i \) values. Figure 4.1 (b) shows that, with the correct likelihood, the profile plot is well behaved with a clear maximum at a value of \( \gamma \) that is a little less than 400.

### 4.1.4 R. A. Fisher’s Definition of Likelihood

Fisher (e.g., 1912, page 157) suggests that a likelihood defined by a product of densities should be proportional to the probability of the data (which we now know is often, but not always true). In particular, he says “... then \( P' \) [the joint density] is proportional to the chance of a given set of observations occurring.” Fisher (1922, page 327) points out that “Likelihood [expressed as a joint probability density] also differs from probability in that it is not a differential element, and is incapable of being integrated: it is assigned to a particular point of the range of variation, not to a particular element of it.”

### 4.1.5 Related Literature

Cheng and Iles (1987) summarize several alternative methods of estimation that have been proposed to remedy the unbounded likelihood problem. Cheng and Amin (1983) suggest the
maximum product of spacings (MPS) method. This method can be applied to any univariate distribution. Wong and Li (2006) use the MPS method to estimate the parameters of the maximum generalized extreme value (GEV) distribution and the generalized Pareto distribution (GPD), both of which have unbounded density-approximation likelihood functions. Cheng and Traylor (1995) point out the drawbacks of the MPS method owing to the occurrence of the tied observations and numerical effects involved in ordering the cdf when there are explanatory variables in the model. Harter and Moore (1965) suggest that one can use the smallest observation to estimate the threshold parameter and then estimate the other two parameters using the remaining observations. This method has been further studied by Smith and Weissman (1985). Although the smallest observation is the ML estimator of the threshold parameter, under this method, the ML estimators of the other parameters are no longer consistent. Kempthorne (1966) and Barnard (1967) suggest a method that is similar to the interval-censoring approach; their method groups the observations into non-overlapping cells, implying a multinomial distribution in which the cell probabilities depend on the unknown parameters.

Cheng and Traylor (1995) describe the unbounded likelihood problem as one of the four types of non-regular maximum likelihood problems with specific examples including the three-parameter Weibull distribution and discrete mixture models. Cheng and Amin (1983) point out that in the three-parameter lognormal, Weibull and gamma distributions, there exist paths in the parameter space where as the threshold parameter tends to the smallest observation, the density-approximation likelihood function approaches infinity. Giesbrecht and Kempthorne (1976) show that the unbounded likelihood problem of the three-parameter lognormal distribution can be overcome by using the correct likelihood instead of the density approximation. Atkinson, Pericchi, and Smith (1991) apply the grouped-data likelihood approach to the shifted power transformation model of Box and Cox (1964). Kulldorff (1957, 1961) argues that ML estimators based on the correct likelihood for grouped data are consistent and asymptotically efficient. Other examples of unbounded density-approximation likelihood functions are given in Section 4.2.
4.1.6 Overview

The remainder of this paper is organized as follows. Section 4.2 divides the models leading to unbounded likelihoods into three categories and for each category, illustrates the density-approximation breakdown with specific examples frequently encountered in practice. Section 4.3, using the minimum GEV distribution and the mixture of two univariate normal distributions (both of which have unbounded likelihood functions) as examples, studies the effect that different amounts of the round-off error have on estimation. Section 4.4 describes the equicontinuity condition, which is a sufficient condition for the product of densities to provide the same maximum likelihood estimate as the correct likelihood, as the round-off error goes to zero. Section 4.5 provides some conclusions.

4.2 A Classification of Unbounded “Likelihoods”

In this section, we divide the models that have an unbounded density-approximation likelihood into three categories.

- Continuous univariate distributions with three or four parameters, including a threshold parameter.
- Discrete mixture models of continuous distributions for which at least one component has both a location and a scale parameter.
- Minimum-type (and maximum-type) models for which at least one of the marginal distributions has both a location and a scale parameter.

The classification we provide includes all of the unbounded likelihood situations that we have observed or found in the literature. Our classification, however, may not be exhaustive.

4.2.1 Continuous Univariate Distributions with Three or Four Parameters, Including a Threshold Parameter

For $n$ independent and identically distributed (iid) observations $x_1, x_2, \ldots, x_n$ from a certain distribution with a threshold parameter that shifts the distribution by an amount $\gamma$, the pdf is
f(x) > 0 for all x > γ and f(x) = 0 for x \leq γ. Generally, there exist paths in the parameter space where the “likelihood” function grows without bound as the threshold parameter tends to the smallest observation. For example, in the log-location-scale distributions (e.g., Weibull, Fréchet, loglogistic and lognormal) with a threshold parameter, the pdf is

\[
\frac{1}{\sigma(x - \gamma)} f_0 \left( \frac{\log(x - \gamma) - \mu}{\sigma} \right) I_{(x>\gamma)}, \tag{4.4}
\]

where \( f_0(x) > 0 \) for all \( x \) is the pdf of a location-scale distribution. Here \( \exp(\mu) \) is a scale parameter, \( \sigma > 0 \) is a shape parameter, and \( \gamma \) is a threshold parameter. The density-approximation likelihood function \( L(\mu, \sigma, \gamma) \to \infty \) as \( \mu = \log(x_{(1)} - \gamma), \sigma^2 = \mu^2, \) and \( \gamma \to x_{(1)} \) (with \( \gamma < x_{(1)} \)).

The simple example in Section 4.1.2 using the three-parameter lognormal distribution is an example in this category. The profile plot in Figure 4.1 (a) indicates the breakdown of the density approximation as the threshold parameter approaches the smallest observation. The unboundedness can be eliminated with the correct likelihood \( L(\mu, \sigma, \gamma) \) as shown in Figure 4.1 (b).

The three-parameter gamma and Weibull distributions also fall into this category. Cheng and Traylor (1995) point out that one can extend the three-parameter Weibull distribution to the maximum generalized extreme value (GEV) distribution by letting the power parameter become negative. Hirose (1996) gives details about how to obtain the minimum GEV distribution by reparameterizing the three-parameter Weibull distribution. The minimum GEV family has the cdf

\[
G(x) = 1 - \exp \left\{ - \left[ 1 - \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\},
\]

\[
1 - \xi \left( \frac{x - \mu}{\sigma} \right) \geq 0, \quad \xi \neq 0. \tag{4.5}
\]

Here \( \xi \) is a shape parameter and \( \mu \) and \( \sigma \) are, respectively, location and scale parameters. Coles and Dixon (1999) suggest that when the number of observations is small (say less than 30), no matter what the initial values are in the numerical optimization algorithm, ML estimation of the maximum GEV parameters can fail to converge properly. In the minimum GEV distribution,
for any location parameter $\mu$, one can always find a path along which the values of $\sigma$ and $\xi$ change and the density-approximation log-likelihood increases without bound. There is a similar result for the maximum GEV distribution. To illustrate this behavior Figure 4.2 (a) plots the density-approximation log-likelihood for a simulated sample of $n = 20$ observations from a minimum GEV distribution with $\mu = -2.2$, $\sigma = 0.5$, and $\xi = -0.2$ as a function of $\mu$ for three different combinations of fixed $\sigma$ and $\xi$. Instead of using the profile log-likelihood plot with respect to the location parameter $\mu$ that blows up at any $\mu$, an alternative density log-likelihood plot is used to present the unbounded behavior. This plot indicates that when the shape parameter $\xi < -1$, as the location parameter $\mu$ approaches $x_{(1)} - \sigma/\xi$ and $\sigma > 0$ and $\xi < -1$ are fixed, the density-approximation log-likelihood increases without bound. When the shape parameter $-1 < \xi < 0$, as the location parameter $\mu$ approaches $x_{(1)} - \sigma/\xi$ and $\sigma > 0$ and $-1 < \xi < 0$ are fixed, the density-approximation log-likelihood decreases without bound. The local maximum is close to the true $\mu$ at $-2.2$. If the shape parameter $\xi > 0$, as the location parameter $\mu$ approaches $x_{(20)} - \sigma/\xi$, the density-approximation log-likelihood again increases without bound. For all of these cases, the unboundedness can be eliminated by using the correct likelihood as shown by the profile log-likelihood plot in Figure 4.2 (b).

The Box-Cox (1964) transformation family of distributions with a location shift provides another example of an unbounded density-approximation likelihood in this category, as described in Chapters 6 and 9 of Atkinson (1985), Atkinson, Pericchi and Smith (1991), and Section 4 of Cheng and Traylor (1995). For a sample $x_1, x_2, \ldots, x_n$, Box and Cox (1964) give the shifted power transformation as

$$y_i(x_i; \gamma, \lambda) = \begin{cases} \frac{(x_i + \gamma)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log(x_i + \gamma), & \text{if } \lambda = 0. \end{cases}$$

Suppose that $y_i \overset{iid}{\sim} \text{Normal } (\mu, \sigma^2), i = 1, \ldots, n$. Then the likelihood function for the original observations $x_i, i = 1, \ldots, n$ using the density approximation is (4.1) where the density is given by
Figure 4.2  The minimum GEV log-likelihood plots of the location parameter $\mu$ for data with sample size $n = 20$ using (a) the density-approximation likelihood $L$ and (b) the correct likelihood $\mathcal{L}$.

$$f(x_i; \theta) = \begin{cases} 
\frac{1}{\sigma} \phi_{nor} \left( \frac{(x_i + \gamma)^\lambda - \mu \lambda - 1}{\sigma \lambda} \right) |(x_i + \gamma)^\lambda - 1|, & \text{if } \lambda \neq 0; \\
\frac{1}{\sigma |x_i + \gamma|} \phi_{nor} \left( \frac{\log(x_i + \gamma) - \mu}{\sigma} \right), & \text{if } \lambda = 0,
\end{cases}$$

and $\theta = (\mu, \sigma, \gamma, \lambda)^\prime$.

Section 9.3 of Atkinson (1985) shows that the profile log-likelihood $\log L^*(\gamma) = \max_{\lambda, \mu, \sigma} \{ \log L(\mu, \sigma, \gamma, \lambda) \}$ goes to $\infty$ as $\gamma \to -x_{(1)}$. Atkinson, Pericchi and Smith (1991) illustrate that the correct likelihood can be used to avoid the unbounded likelihood problem for this distribution.

### 4.2.2 Discrete Mixture Models Where at Least One Component Has Both a Location and a Scale Parameter

Suppose there are $n$ iid observations $x_1, x_2, \ldots, x_n$ from the $m$-component discrete mixture distribution with the pdf
\[ f(x; \theta) = \sum_{i=1}^{m} p_i f_i(x; \theta_i), \] (4.6)

where \( \theta = (p_1, p_2, \ldots, p_m, \theta'_1, \theta'_2, \ldots, \theta'_m)' \), \( p_i \) is the proportion of component \( i \) with \( \sum_{i=1}^{m} p_i = 1 \) and for at least one \( i \), the pdf for component \( i \) can be expressed as

\[ f_i(x; \theta_i) = \frac{1}{\sigma_i} \phi \left( \frac{x - \mu_i}{\sigma_i} \right). \]

That is, at least one component \( i \) belongs to the location-scale family with an unknown location parameter \( \mu_i \) and scale parameter \( \sigma_i \). Then if one sets a component location parameter equal to one of the observations, fixes the component proportion parameter at a positive value (less than 1), and allows the corresponding scale parameter to approach zero while fixing other parameter values, the likelihood increases without bound. Section 1.2.3 of Zucchini and MacDonald (2009) shows that replacing the density-approximation likelihood with the correct likelihood can again avoid the problem of unboundedness. Of course the same problem arises in mixtures of the corresponding log-location-scale distributions for which \( \exp(\mu_i) \) is a scale parameter and \( \sigma_i \) is a shape parameter.

We use a simulated example to illustrate that replacing the density-approximation likelihood with the correct likelihood will eliminate the unbounded likelihood problem for the finite mixture models. We simulated one hundred observations \( x_1, x_2, \ldots, x_{100} \) (following Example 1 of Yao 2010) from a mixture of two univariate normal components with proportions \( p_1 = 0.7, p_2 = 0.3 \), means \( \mu_1 = 1, \mu_2 = 0 \), and variances \( \sigma_1^2 = 1, \sigma_2^2 = 0.25 \). Let \( \delta = \min\{\sigma_1, \sigma_2\}/\max\{\sigma_1, \sigma_2\} \), so that \( \delta \in (0, 1] \). The original observations were rounded to one digit after the decimal point and thus the corresponding value of \( \Delta \) used in the correct likelihood is 0.05. Figure 4.3 (a) shows that as \( \delta \) approaches 0, the density-approximation profile log-likelihood of \( \delta \) increases without bound. The counterpart in Figure 4.3 (b) using the correct likelihood solves the unboundedness problem. This plot also shows that the correct log-likelihood profile for \( \delta \) tends to be flat for small values of \( \delta \). This is due to the fact that as \( \delta \) approaches 0, the log-likelihood function is dominated by two parts: one part comes from the point mass at the smallest observation \( x_{(1)} \), the other part is the log-likelihood of \( x_{(2)}, \ldots, x_{(100)} \)
that follow the other normal distribution.

\[ (a) \quad (b) \]

Figure 4.3 Mixture distribution of two univariate normal profile log-likelihood plots of \( \delta \) using (a) the unbounded density-approximation likelihood and (b) the correct likelihood (with \( \Delta=0.05 \)).

The switching regression model, described in Quandt (1972) and Quandt and Ramsey (1978), provides another example in this mixture category in which there is a finite mixture of regression models. For example, when there are \( m = 2 \) components, the pdf is given by (4.6) with

\[
f_i(y; \theta_i) = \frac{1}{\sigma_i} \phi_{nor} \left( \frac{y - x_i' \beta_i}{\sigma_i} \right).
\]

The switching regression model is a special case of the \( m \)-component discrete mixture distribution with \( \mu_i = x_i' \beta_i \).

Protheroe (1985) proposes a new statistic to describe the periodic ultra-high energy \( \gamma \)-ray signal source represented by a moving point on a circle. As explained there, events observed in time are caused by a background process in which the noise occurs according to a Poisson process with a constant intensity. Events from the “signal” occur according to a process with an intensity that is periodic with a known period \( P \). For the event times \( T_1, T_2, \ldots \), the
transformation $X_i = \text{mod}(T_i, P)/P$ maps the periodic event time into a circle with a unit circumference. As shown in Meeker and Escobar (1994), the pdf of $X$ can be written as

$$f_X(x; \theta) = p + (1 - p)f_N(x; \mu, \sigma),$$

where $\theta = (\mu, \sigma, p)'$, $0 \leq x \leq 1$, $0 < p < 1$, $0 \leq \mu \leq 1$, $\sigma > 0$, and

$$f_N(x; \mu, \sigma) = \frac{1}{\sigma} \sum_{j=-\infty}^{\infty} \phi_{\text{nor}} \left( \frac{x + j - \mu}{\sigma} \right),$$

where $\phi_{\text{nor}}$ is the standard normal pdf. Here, $f_N(x; \mu, \sigma)$ is the pdf for a “wrapped normal distribution.” For any $0 < p < 1$, let $\mu = x_i$ for any $i$, where $x_1, x_2, \ldots, x_n$ are iid transformed observations with the pdf $f_X(x; \theta)$, and then as $\sigma \to 0$, the product of the pdf’s approaches $\infty$. Section 8.3.2 of Meeker and Escobar (1994) shows how to use ML to estimate the parameters of the wrapped normal distribution by using the bounded correct likelihood.

Another example arises from a mixture model, described by Vardeman (2005), in which a proportion $p$ has a uniform $(0, 1)$ distribution and the remaining proportion $(1 - p)$ has a uniform $(\alpha - \beta, \alpha + \beta)$ distribution, where $0 \leq \alpha - \beta < \alpha + \beta \leq 1$. The pdf for this mixture is

$$f(x; \alpha, \beta, p) = pI(0 \leq x \leq 1) + \frac{1 - p}{2\beta} I(\alpha - \beta \leq x \leq \alpha + \beta),$$

where $\alpha$ is a location parameter and $\beta$ is a scale parameter. For the iid sample $x_1, x_2, \ldots, x_n$, if we set $\alpha = x_{(1)}$, constrain $0 < \beta < \min(x_{(1)}, x_{(2)} - x_{(1)})$, and fix $0 < p < 1$, then

$$L(\alpha, \beta, p) = p^{n-1} \left( p + \frac{1 - p}{2\beta} \right) \to \infty,$$

as $\beta \to 0$.

4.2.3 Minimum-Type (and Maximum-Type) Models Where at Least One of the Marginal Distributions Has Both a Location and a Scale Parameter

For $m \geq 2$ independent random variables $X_1, X_2, \ldots, X_m$ with cdf $F_i(x; \theta_i)$ and pdf $f_i(x; \theta_i)$, the cdf $F_{\text{min}}(x)$ of the minimum $X_{\text{min}} = \min\{X_1, X_2, \ldots, X_m\}$ can be expressed as
\[ F_{\text{min}}(x; \theta) = \Pr(X_{\text{min}} \leq x) = 1 - \prod_{i=1}^{m} [1 - F_i(x; \theta_i)]. \] (4.7)

Then, the pdf of \( X_{\text{min}} \) is

\[ f_{\text{min}}(x; \theta) = \sum_{i=1}^{m} \left\{ f_i(x; \theta_i) \prod_{j \neq i}^{m} [1 - F_j(x; \theta_j)] \right\}. \]

Suppose again that for at least one \( i \),

\[ f_i(x; \theta_i) = \frac{1}{\sigma_i} \phi \left( \frac{x - \mu_i}{\sigma_i} \right), \]

which is the pdf belonging to the location-scale family with location parameter \( \mu_i \) and scale parameter \( \sigma_i \). Then for \( n \) iid observations \( x_1, x_2, \ldots, x_n \) with the pdf \( f_{\text{min}}(x; \theta) \), if one sets the location parameter for one of the components equal to the largest observation and allows the corresponding scale parameter to approach zero while fixing other parameter values, the likelihood increases without bound.

Friedman and Gertsbakh (1980) describe a minimum-type distribution (MTD) that is a special case of (4.7) with two random failure times: \( t_A \sim \text{Weibull} (\alpha_A, \beta_A), \) \( t_B \sim \text{Exp} (\alpha_B) \). The failure time of the device is \( T = \min\{t_A, t_B\} \) with cdf

\[ P(T \leq t) = 1 - \exp \left[ -\frac{t}{\alpha_B} - \left( \frac{t}{\alpha_A} \right)^{\beta_A} \right]. \] (4.8)

The cdf for \( Y = \log (T) \) can be written as

\[ P(Y \leq y) = 1 - \exp \left[ -\exp(y - \mu_B) - \exp \left( \frac{y - \mu_A}{\sigma_A} \right) \right], \]

where \( \mu_B = \log(\alpha_B) \) and \( \mu_A = \log(\alpha_A) \) are location parameters and \( \sigma_A = 1/\beta_A \) is the scale parameter of the smallest extreme value distribution. Friedman and Gertsbakh (1980) show that if all three parameters are unknown, there exists a path in the parameter space along which the likelihood function tends to infinity for any given sample and suggest an alternative method of estimation.
We simulated 100 observations \( t_1, t_2, \ldots, t_{100} \) from the MTD in (4.8) with \( \alpha_A = 2, \beta_A = 4 \), and \( \alpha_B = 1 \). Figure 4.4 (a) shows that when \( \mu_A = \log(t_{(100)}) \), \( \mu_B \) is fixed and \( \sigma_A \) approaches zero, the log-likelihood function increases without bound. The unboundedness problem can again be resolved by using the correct likelihood as shown in Figure 4.4 (b).

The correct likelihood has a global maximum in situations where a sufficient amount of data is available from each component of the minimum process. When, however, one or the other of the minimum process dominates in generating the data, due to particular values of the parameters or right censoring, there can be an identifiability problem so that a unique maximum of the three-parameter likelihood will not exist.

\[
(a) \hspace{2cm} (b)
\]

Figure 4.4 An MTD corresponding to a minimum of two independent random variables having the Weibull and exponential distributions. The profile log-likelihood plots of \( \sigma_A \) using (a) the density-approximation likelihood and (b) the correct likelihood (with \( \Delta=0.05 \)).

The simple disequilibrium model illustrated in Griliches and Intriligator (1983) is another special case of the minimum-type model. They consider two random variables \( X_1 \) and \( X_2 \) from a normal distribution leading again to the likelihood function in (4.1) with
\[ f(y_i; \theta) = \frac{1}{\sigma_1} \phi_{\text{nor}} \left( \frac{y_i - X_{1i} \beta_1}{\sigma_1} \right) \left[ 1 - \Phi_{\text{nor}} \left( \frac{y_i - X_{2i} \beta_2}{\sigma_2} \right) \right] \]
\[ + \frac{1}{\sigma_2} \phi_{\text{nor}} \left( \frac{y_i - X_{2i} \beta_2}{\sigma_2} \right) \left[ 1 - \Phi_{\text{nor}} \left( \frac{y_i - X_{1i} \beta_1}{\sigma_1} \right) \right], \]

where \( \theta = (\beta_1', \beta_2', \sigma_1, \sigma_2)' \). An argument very similar to that employed in the case of the switching regression model can be used to show that the simple disequilibrium model has an unbounded likelihood function. Again, using the correct likelihood avoids this problem.

Maximum-type models for which at least one of the marginal distributions has both a location and scale parameter have the same problem as the minimum-type models.

4.3 The Effect of the Round-off Error on Estimation

In this section we use two examples to explore the effect that the round-off error \( \Delta \) has on estimation with the correct likelihood. Of course, if one knows the precision of the measuring instrument and the rule that was used for rounding one’s data, the choice of \( \Delta \) is obvious. Sometimes, however, the precision of the measuring instrument and the exact rounding scheme that was used for a data set are unknown.

4.3.1 Background

Giesbrecht and Kempthorne (1976) present an empirical study that investigates the effect of the round-off error \( \Delta \) for estimation of the parameters for the three-parameter lognormal distribution. As \( \Delta \) decreases, the asymptotic variances and covariances of the MLE of the parameters rapidly approach the values for the no-censoring case provided in Harter and Moore (1965). Atkinson, Pericchi and Smith (1991) suggest that care is needed in choosing \( \Delta \) and recommend that one could examine profile likelihood plots for different values of \( \Delta \) and choose the value of \( \Delta \) that makes the profile likelihood smooth near \( \mu = -y(1) \) for the Box-Cox (1964) shifted power transformation model. Vardeman and Lee (2005) discuss how to use the likelihood function based on the rounded observations to make inferences about the parameters of the underlying distribution. They emphasize that the relationship between the range of the
rounded observations and the rounding rule, defined by $\Delta$, can have an important effect on inferences when the ratio of $\Delta/\sigma$ is large (say more than 3).

### 4.3.2 Numerical Examples

In the first example, we investigate the effect of the round-off error on estimation of the location parameter for the minimum GEV distribution. Vardeman and Lee (2005) suggest (for a different distribution) that the round-off error $\Delta$ should be sufficiently large (equivalently, the number of digits after the decimal point should be sufficiently small), relative to the range of the rounded observations, so that the maximum of the correct likelihood function exists. We consider four different minimum GEV distributions with a common location parameter $\mu = -2.2$ and shape parameter $\xi = -0.2$ but different scale parameters $\sigma = 0.3, 0.5, 1, \text{and } 2$. For each minimum GEV distribution, we generated 15 samples, each with $n = 20$ observations. For a given rounding scheme, the correct likelihood of the minimum GEV distribution (proportional to the probability of the data) for a sample $x_1, x_2, \ldots, x_{20}$ can be written as

$$L(\theta) \propto \prod_{i=1}^{20} \Pr(l_{x_{ik}} < x_i < u_{x_{ik}})$$

$$= \prod_{i=1}^{20} \left[ G(u_{x_{ik}}) - G(l_{x_{ik}}) \right]$$

$$= \prod_{i=1}^{20} \left( \exp \left\{ - \left[ 1 - \xi \left( \frac{l_{x_{ik}} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} - \exp \left\{ - \left[ 1 - \xi \left( \frac{u_{x_{ik}} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \right),$$

where $l_{x_{ik}}$ and $u_{x_{ik}}$ are respectively the lower and upper endpoints of the interval with $k$ digits after the decimal point for the $i$th observation and $\theta = (\mu, \sigma, \xi, k)'$.

For different values of $\sigma$, the plots in Figure 4.5 compare the location parameter estimates $\hat{\mu}$ when $k$, the number of digits after the decimal point, varies from 1 to 4 (i.e., the corresponding round-off error $\Delta$ changes from 0.05 to 0.00005). We did not use the same y-axis range in these four plots to allow focus on the stability of the estimates as a function of $k$. Figure 4.5 indicates that when $\sigma$ is small, more samples provide different parameter estimates $\hat{\mu}$ for different numbers of digits after the decimal point $k$ than in the situation with large $\sigma$. This is true especially when the number of digits after the decimal point $k$ is small (i.e., $\Delta$ is large).
In this example, when $\sigma$ is 2, the choice of $k$ does not significantly affect the estimation of the location parameter $\mu$.

![Figure 4.5 Plots of the minimum GEV location parameter estimates $\hat{\mu}$ versus the number of digits after the decimal point $k$ with $\sigma$ equal to 0.3, 0.5, 1, and 2. The $\Delta$ values on the top of each plot correspond to the round-off errors.](image)

For the second example, we return to the mixture of two univariate normal distributions in Section 4.2.2 to study the effect that $\Delta$ in the correct likelihood has on estimation. The plots in Figure 4.3 for $\Delta = 0.05$ are relatively smooth. Rounding the same original observations $x_1, x_2, \ldots, x_{100}$ to two digits after the decimal point, the profile log-likelihood functions with $\Delta = 0.005$ in Figure 4.6 are more wiggly than the corresponding profile log-likelihood functions, in Figure 4.3 with $\Delta = 0.05$, especially in the right part of the plots. The occurrence of multiple bumps in the profile log-likelihood curves, for $\Delta = 0.005$ is due to the fact that increasing the data precision will result in more distinct clusters with data points close together. As in Figure 4.3, we note that the profile log-likelihood plot using the correct likelihood is flat as $\delta$ approaches
Figure 4.6 Mixture distribution of two univariate normal profile log-likelihood plots of $\delta$ using (a) the unbounded density-approximation likelihood and (b) the correct likelihood (with $\Delta = 0.005$).

4.4 A Sufficient Condition for Using the Density-Approximation Likelihood

As discussed in Sections 4.1.2 and 4.1.3, the likelihood function based on the usual density approximation for $n$ iid observations $t_i, i = 1, \ldots, n$ has the form

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta; t_i) = \prod_{i=1}^{n} f(t_i; \theta).$$

The correct likelihood based on small intervals $(t_i - \Delta, t_i + \Delta)$ (implied by the data’s precision) can be expressed as

$$L_\Delta(\theta) = \prod_{i=1}^{n} L_i(\theta; t_i) = \left( \frac{1}{2\Delta} \right)^n \prod_{i=1}^{n} \int_{t_i - \Delta}^{t_i + \Delta} f(t; \theta) dt.$$
Here, for simplicity we assume that $\Delta_i = \Delta$ for all $i$. Kempthorne and Folks (1971, page 259) state that a sufficient condition for $L(\theta)$ to be proportional to the probability of the data (i.e., to be a proper likelihood) is to satisfy the Lipschitz condition which states that for all $\epsilon \neq 0$ in the interval $(-\Delta/2, \Delta/2)$ ($\Delta > 0$ is fixed), there exists a function $h(t; \theta)$ such that

$$\left| \frac{f(t + \epsilon; \theta) - f(t; \theta)}{\epsilon} \right| < h(t; \theta).$$

Kempthorne and Folks (1971) did not provide a proof, and we have been unable to find one. Also, we have been unable to construct or find a pdf $f(t; \theta)$ that satisfies the Lipschitz condition but has different ML estimates when using the density-approximation and the correct likelihoods as $\Delta \to 0$. We, therefore, look at this problem from a different perspective and provide an alternative sufficient condition through Theorem 1 and its corollary.

**Theorem 1**: For $n$ iid observations $t_1, t_2, \ldots, t_n$ from a distribution with the pdf $f(t; \theta)$ and $\theta \in \Theta \subset \mathbb{R}^p$, if $\{f(t; \theta)\}_{\theta \in \Theta}$ is equicontinuous at $t_1, t_2, \ldots, t_n$, then the correct likelihood $\{L_\Delta(\theta)\}$ converges uniformly to the density-approximation likelihood $L(\theta)$ as $\Delta \to 0$.

The proof is given in Appendix 6.1. By the property of a uniformly convergent sequence described by Intriligator (1971), if $\{L_\Delta(\theta)\}$ converges uniformly to $L(\theta)$ as $\Delta \to 0$, then $\{\theta^*_\Delta\}$ converges to $\theta^*$, where $\theta^*_\Delta$ and $\theta^*$ are the unique maximizers of $L_\Delta(\theta)$ and $L(\theta)$, respectively, assuming that $\theta^*_\Delta$ and $\theta^*$ exist in the parameter space $\Theta$. We now have the following corollary.

**Corollary 1**: For $n$ iid observations $t_1, t_2, \ldots, t_n$ from a distribution with the pdf $f(t; \theta)$ for $\theta \in \Theta \subset \mathbb{R}^p$, if $\{f(t; \theta)\}_{\theta \in \Theta}$ is equicontinuous at $t_1, t_2, \ldots, t_n$, then $\{\theta^*_\Delta\}$ converges to $\theta^*$ as $\Delta \to 0$, where $\theta^*_\Delta$ and $\theta^*$ are the unique maximizers of $L_\Delta(\theta)$ and $L(\theta)$, respectively, assuming that $\theta^*_\Delta$ and $\theta^*$ exist in the parameter space $\Theta$.

The equicontinuity condition, however, is not necessary. An example is given in Appendix 6.2.

### 4.5 Concluding Remarks

In this paper, we used the correct likelihood based on small intervals instead of the density approximation to eliminate the problem of an unbounded likelihood. We explored several classes
of models where the unbounded “likelihood” arises when using the density-approximation likelihood and illustrated how using a correctly expressed likelihood can eliminate the problem. We investigated the effect that the round-off error has on estimation with the correct likelihood, especially under the circumstance having unknown precision of the measuring instrument. The equicontinuity condition is sufficient for the density-approximation and correct likelihoods to provide the same MLE’s as the round-off error $\Delta \to 0$.

4.6 Appendix

4.6.1 Proof of Theorem 1

If \( \{ f(t; \theta) \}_{\theta \in \Theta} \) is equicontinuous at \( t_1, t_2, \ldots, t_n \), then for each \( t_i \), for any \( \epsilon > 0 \), there exists a \( \delta > 0 \) such that if \( |t - t_i| < \delta \), then \( |f(t; \theta) - f(t_i; \theta)| < \epsilon \) for all \( \theta \in \Theta \). Hence, for \( 0 < \Delta < \delta \),

\[
\frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} f(t; \theta) \, dt - f(t_i, \theta) \leq \frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} |f(t; \theta) - f(t_i; \theta)| \, dt \leq \epsilon.
\]

Therefore, for every \( t_i \), \( \{ \frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} f(t; \theta) \, dt \} \) converges uniformly to \( f(t_i, \theta) \), and thus \( \{ L_\Delta(\theta) \} \) converges uniformly to \( L(\theta) \) as \( \Delta \to 0 \).

4.6.2 An Example Showing the Equicontinuity Condition Is Not Necessary

The pdf \( \{ f(x; \theta) \}_{\theta \in \Theta} \) below is not equicontinuous at \( x = 0 \) or \( 2 \), but the ML estimates using the density-approximation likelihood and the correct likelihood are the same as \( \Delta \to 0 \). For \( 0 < \theta < 2 \), the univariate pdf has the form

\[
f(x, \theta) = \begin{cases} 
1 - \frac{x}{\theta}, & \text{if } 0 < x < \theta; \\
\frac{x - \theta}{2 - \theta}, & \text{if } \theta \leq x < 2.
\end{cases}
\]
Also, $f(x, 0) = x/2$ for $0 \leq x \leq 2$ and $f(x, 2) = 1 - x/2$ for $0 \leq x \leq 2$. Note that $\{f(x, \theta)\}_{\theta \in \Theta}$ is not equicontinuous at $x = 0$ or 2, because $f(x, \theta)$ is not even continuous at $x = 0$ or 2 for $\theta \in [0, 2]$. Suppose that $n = 1$ and we have one observation $x_1$. When using the density-approximation likelihood, if $0 \leq x_1 < 1$, the ML estimate is $\hat{\theta} = 2$; if $1 < x_1 \leq 2$, the ML estimate is $\hat{\theta} = 0$; if $x_1 = 1$, the ML estimate is $\hat{\theta} = 0$ or 2 (not unique). For $0 < \theta < 2$, the cdf has the form

$$F(x, \theta) = \begin{cases} \frac{2 - \frac{x}{\theta}}{2}, & \text{if } 0 \leq x < \theta; \\ \frac{\theta}{2} + \frac{(x - \theta)^2}{2(2 - \theta)}, & \text{if } \theta \leq x \leq 2. \end{cases}$$

Also, $F(x, 0) = x^2/4$ for $0 \leq x \leq 2$, and $F(x, 2) = x - x^2/4$ for $0 \leq x \leq 2$.

For $0 < x_1 < 2$, consider $0 < \Delta < \min\left(\frac{x_1}{8}, \frac{2 - x_1}{8}\right)$. Then the correct likelihood is

$$L_\Delta(\theta) = \frac{1}{2\Delta} \left[ F(x_1 + \Delta, \theta) - F(x_1 - \Delta, \theta) \right]$$

$$= \begin{cases} \frac{x_1}{2}, & \text{if } \theta = 0; \\ \frac{x_1 - \theta}{2 - \theta}, & \text{if } 0 < \theta < x_1 - \Delta; \\ \frac{1}{4\Delta} \left[ \frac{(x_1 + \Delta - \theta)^2}{2 - \theta} + \frac{(x_1 - \Delta - \theta)^2}{\theta} \right], & \text{if } x_1 - \Delta \leq \theta \leq x_1 + \Delta; \\ 1 - \frac{x_1}{\theta}, & \text{if } x_1 + \Delta < \theta < 2; \\ 1 - \frac{x_1}{2}, & \text{if } \theta = 2. \end{cases}$$

Note that for $x_1 - \Delta \leq \theta \leq x_1 + \Delta$,

$$L_\Delta(\theta) \leq \frac{1}{4\Delta} \left[ \frac{(2\Delta)^2}{2 - \theta} + \frac{(2\Delta)^2}{\theta} \right] \leq \frac{\Delta}{(2 - x_1) - \Delta} + \frac{\Delta}{x_1 - \Delta}$$

$$\leq \frac{\Delta}{8\Delta - \Delta} + \frac{\Delta}{8\Delta - \Delta} = \frac{2}{7};$$

for $0 < \theta < x_1 - \Delta$,

$$L_\Delta(\theta) < \frac{x_1 - \frac{x_1}{2 \theta}}{2 - \theta} = \frac{x_1}{2} = L_\Delta(0);$$

and for $x_1 + \Delta < \theta < 2$,
Thus the ML estimate using the correct likelihood is

\[
\hat{\theta}_{\Delta} = \begin{cases} 
2, & \text{if } 0 < x_1 < 1; \\
0, & \text{if } 1 < x_1 < 2; \\
0 \text{ or } 2, & \text{if } x_1 = 1.
\end{cases}
\]

This is exactly the same as the \( \hat{\theta} \) we obtained before when using the density-approximation likelihood.

For \( x_1 = 0 \), we have

\[
\mathcal{L}_{\Delta}(\theta) = \frac{1}{2\Delta} F(\Delta, \theta) = \begin{cases} 
\frac{\Delta}{8}, & \text{if } \theta = 0; \\
\frac{1}{2\Delta} \left[ \frac{\theta}{2} + \frac{(\Delta - \theta)^2}{2(2\theta - \theta)} \right], & \text{if } 0 < \theta \leq \Delta; \\
\frac{1}{2} - \frac{\Delta}{8}, & \text{if } \Delta < \theta < 2; \\
\frac{1}{2} - \frac{\Delta}{8}, & \text{if } \theta = 2.
\end{cases}
\]

Thus, \( \hat{\theta}_{\Delta} = 2 \) (for \( 0 < \Delta < \frac{1}{8} \)), the same as \( \hat{\theta} \). Similarly, for \( x_1 = 2 \), we also have \( \hat{\theta}_{\Delta} = \hat{\theta} = 0 \) (for \( 0 < \Delta < \frac{1}{8} \)).
Bibliography


CHAPTER 5. GENERAL CONCLUSIONS

In this dissertation, we established statistical models to estimate pipeline integrity and reliability. The main part of this dissertation consists of three technical papers.

In Chapter 2, we estimated the minimum thickness along a pipeline using several different methods including the traditional minimum method, the $BL_{\min}$-Gumbel method and the $BL_{\min}$-GEV method. We conducted a simulation study to explore the impact that block size has on the performance of the block-minima extreme value method for estimating small quantiles of a distribution of a minimum. The simulation results suggest that if the parent distribution is known, the traditional minimum method provides quantile estimates with the most precision. In the $BL_{\min}$-Gumbel method, how to choose an appropriate block size depends on the parent distribution. As the most robust method, the $BL_{\min}$-GEV method provides inference on the distribution of the minimum without need to specify the particular form of the parent distribution. Use of the $BL_{\min}$-GEV method, however, requires using a large number of blocks; at least 30 blocks are recommended.

In Chapter 3, for different longitudinal pipeline data, we proposed degradation models to describe the pipeline corrosion behaviors. As an alternative to the likelihood approach, we estimated the parameters in the degradation models by using a Bayesian approach with appropriate prior distributions. For the purpose of assessing the life time of the pipelines, we derived the failure time and the remaining lifetime distributions. We also simulated a data with pipeline thickness measurements at a larger number of circuits within a facility and built a hierarchical model that could be used to estimate the corrosion rate of TMLs for the simulated data.

In Chapter 4, we suggested that instead of using the density-approximation likelihood, the correct likelihood based on small intervals could eliminate the problem of unbounded like-
lihoods. We studied three classes of models with the “unbounded likelihood” problem and illustrated that using a correctly expressed likelihood can eliminate the problem. We also investigated the effect that round-off error has on estimation with the correct likelihood. We proposed the equicontinuous condition as a sufficient condition for the density approximation and the correct likelihoods to provide the same MLE’s as the round-off error approaches 0.