

ABSTRACT

ZHENG, JUNYU. Quantification of Variability and Uncertainty in Emission Estimation: General Methodology and Software Implementation.(Under the supervision of Dr. H. Christopher Frey).

The use of probabilistic analysis methods for dealing with variability and uncertainty is being more widely recognized and recommended in the development of emission factor and emission inventory. Probabilistic analysis provides decision-makers with quantitative information about the confidence with which an emission factor may be used. Variability refers to the heterogeneity of a quantity with respect to time, space, or different members of a population. Uncertainty refers to the lack of knowledge regarding the true value of an empirical quantity. Ignorance of the distinction between variability and uncertainty may lead to erroneous conclusions regarding emission factor and emission inventory. This dissertation extensively and systematically discusses methodologies associated with quantification of variability and uncertainty in the development of emission factors and emission inventory, including the method based upon use of mixture distribution and the method for accounting for the effect of measurement error on variability and uncertainty analysis. A general approach for developing a probabilistic emission inventory is presented. A few example case studies were conducted to demonstrate the methodologies. The case studies range from utility power plant emission source to highway vehicle emission sources. A prototype software tool, AUVVEE, was developed to demonstrate the general approach in developing a probabilistic emission inventory based upon an example utility power plant emission source. A general software tool, AuvTool, was developed to implement all

methodologies and algorithms presented in this dissertation for variability and uncertainty analysis. The tool can be used in any quantitative analysis fields where variability and uncertainty analysis are needed in model inputs.

KEY WORDS: Variability, Uncertainty, Emission Factor, Emission Inventory, Software Implementation, Bootstrap Simulation, Monte Carlo Simulation, Two-Dimensional Simulation

**QUANTIFICATION OF VARIABILITY AND
UNCERTAINTY IN EMISSION ESTIMATION: GENERAL
METHODOLOGY AND SOFTWARE IMPLEMENTATION**

By

JUNYU ZHENG

A dissertation submitted to the graduate faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

CIVIL ENGINEERING

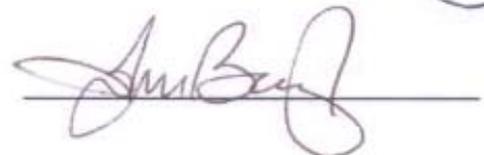
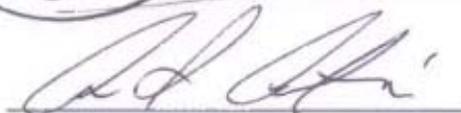
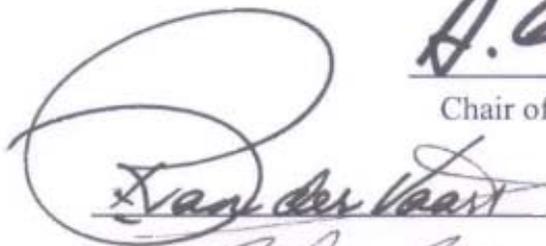
Raleigh

2002

APPROVED BY:



Chair of Advisory Committee



To my beloved parents, Yongnian Zheng and Guifang Zhang for their constant
understanding and support

To my wonderful wife, Yan Ouyang, for her love, concerns and encouragements

To my upcoming son, Alexander Zheng, for his inciting my responsibility and hope

BIOGRAPHY

Junyu Zheng received his Bachelor of Engineering degree in Water Supply and Drainage from Wuhan Urban Construction Institute, Wuhan, China, in 1991. During the period of August of 1991 through August of 1993, he worked as a water supply process manager in Power Plant, Beijing Yanshan Petrol & Chemical Corporation (BYPCC), where he was responsible for process management of industrial and civil water supply of BYPCC.

Junyu Zheng was admitted to the Department of Environmental Engineering of Tsinghua University, China, in September of 1993 for his pursuit of master degree, where he worked as research assistant at the Environmental System Lab and his research focused on the probabilistic analysis of water quality model. He earned his Master of Science degree in Environmental Engineering from Tsinghua University in 1996. After he received his M.S degree, he accepted a position of real-estate appraiser from Jingdu Certified Public Accountant (one member of Horwath International), Beijing, and part-time worked as an assistant general manager and software developer at Beijing Kelier Information Inc. to engage in the software development and network installation of Automatic Check Telephone Query System.

Junyu Zheng came to North Carolina State University, Raleigh, North Carolina, USA, in August 1998 for his Ph.D. in Environmental Engineering program. During his Ph.D study, he worked as a research assistant at the Computational Laboratory for Energy, Air and Risk (CLEAR).

Junyu Zheng's research interests include application of statistics and computer techniques to the quantification of variability and uncertainty in environmental data,

probabilistic risk or exposure assessment, and uncertainty analysis in air or water quality modeling.

ACKNOWLEDGEMENTS

This research was supported by U.S. EPA STAR Grants Nos. R826766 and R826790. The ORD of U.S. EPA funded the development of AuvTool via contract ID-S794-NTEX.

The author wishes to express his appreciation to Dr. H. Christopher Frey for his constant inspiration and guidance throughout the course of this research. Appreciation is also extended to Drs. L.A. Stefanski, E.D. Brill, J. W. Baugh, D. Vandervaart and A. Anton for their valuable suggestions. The author appreciates the guidance and encouragement of Dr. Jianping Xue and Dr. Haluk Ozkaynak of U.S. EPA during the development of the AuvTool.

The author thanks Alper Unal for his help in both research and life and his friendship. Thanks also go to all other members in the group of Computational Laboratory for Energy, Air and Risk.

Table of Contents

LIST OF TABLES	xii	
LIST OF FIGURES	xiv	
PART I	INTRODUCTION..... 1	
1.0	Introduction..... 3	
1.1	Variability	5
1.2	Uncertainty..... 6	
1.3	Distinctions Between Variability and Uncertainty	8
1.4	Examples of Probabilistic Analysis	10
1.5	Limitations of Current Studies in Variability and Uncertainty Analysis..	11
1.6	Available Software Tools in Probabilistic Analysis	14
1.7	Objectives	15
1.8	Overview of Research..... 16	
1.9	Organization..... 18	
1.10	References..... 20	
PART II	GENERAL METHODOLOGY OF QUANTIFICATION OF VARIABILITY AND UNCERTAINTY IN EMISSION ESTIMATION	25
2.0	General Methodology	27
2.1	General Approach for Developing a Probabilistic Emission Inventory ...	28
2.2	Data Preparation..... 29	
2.3	Emission Inventory Models	30
2.4	Numerical sampling techniques..... 32	
2.4.1	Monte Carlo Sampling..... 32	
2.4.3	Latin Hypercube Sampling	34
2.5	Visualization of Datasets Using Empirical Distributions	35
2.6	Definitions of Probability Distribution Models	38
2.6.1	Definition of Parametric Probability Distributions..... 39	

2.6.2	Empirical Distribution	39
2.7	Parameter Estimation of Parameter Distributions.....	40
2.7.1	Method of Matching Moments	43
2.7.2	Maximum Likelihood Estimation (MLE).....	43
2.8	Evaluation of Goodness-of-Fit of a Probability Distribution Model	47
2.8.1	Graphical Comparison of CDF of Fitted Distribution to the Data	50
2.8.2	Kolmogorov-Smirnov Test	51
2.8.3	Anderson-Darling Test.....	54
2.8.4	Graphical Comparison of Confidence Intervals for CDF of Fitted Distribution to the Data	56
2.8.5	Summary of Methods for Evaluating Goodness-of-Fit	57
2.9	Algorithms for Generating Random Samples from Probability Distributions.....	58
2.9.1	Pseudo Random Number Generator	59
2.9.2	Empirical Distribution	60
2.10	Characterization of Uncertainty in the Distribution for Variability.....	61
2.10.1	Bootstrap Method.....	63
2.10.2	Methods of Generating Bootstrap Samples	64
2.10.3	Methods of Forming Bootstrap Confidence Intervals	65
2.10.4	Two-Dimensional Simulation of Variability and Uncertainty.....	69
2.11	Probabilistic Approaches for Simulating Variability and Uncertainty in the Emission Inventories.....	72
2.12	Identification of Key Sources of Variability and Uncertainty	73
2.13	Summary	76
2.14	References.....	78
PART III	SOFTWARE IMPLEMENTATION	81
3.0	Software Implementation.....	83
3.1	Software Implementation of AuvTool	83
3.1.1	AuvTool Software Design Considerations	84
3.1.2	Development Environment and Tools	84
3.1.3	Structure Design of the AuvTool System	85

3.1.4	AuvTool Main Modules.....	85
3.2	Software Implementation of AUVEE.....	98
3.2.1	General Structure of the AUVEE Prototype Software	98
3.2.2	Databases in the AUVEE Prototype Software.....	98
3.2.3	Modules in the AUVEE Prototype Software	100
3.2.4	Software Development Tools	102
3.3	References.....	104
PART IV	QUANTIFICATION OF VARIABILITY AND UNCERTAINTY USING MIXTURE DISTRIBUTION: EVALUATION OF SAMPLE SIZE, MIXING WEIGHTS AND SEPARATION BETWEEN COMPONENTS	105
	Abstract.....	107
1.0	Introduction.....	108
2.0	Methodology.....	113
2.1	Mixture Distribution	113
2.2	Parameter Estimation of Mixture Distributions.....	115
2.3	Quantification of Variability and Uncertainty Using Mixture Distribution	118
3.0	Introduction to Study Design.....	122
4.0	Results and Discussion	123
4.1	Properties of Confidence Intervals of Cumulative Distributions....	123
4.2	Comparisons between Single Distribution and Mixture Distributions.....	125
4.3	Dependencies among Sampling Distributions of Parameters of Mixture Distributions.....	127
5.0	An Illustrative Case Study: NO _x Emission Factor for a Coal-Fired of Power Plant	128
5.1	Parameter Estimation for the Fitted Distribution.....	129
5.2	Variability and Uncertainty in the NO _x Emission Factor	129
5.3	Uncertainty in the Mean NO _x Emission Factor	130
6.0	Conclusion	131
	References.....	133

PART V	QUANTIFICATION OF VARIABILITY AND UNCERTAINTY WITH MEASUREMENT ERROR	151
	Abstract.....	153
1.0	Introductuion.....	154
1.1	Variability and Uncertainty.....	155
1.2	Limitations of Current Studies in Variability and Uncertainty Analysis.....	155
1.3	Measurement Error and Uncertainty.....	156
1.4	Classification of Measurement Errors	157
1.5	Purpose of This Study.....	158
2.0	Methodology.....	158
2.1	Measurement Error Models	159
2.2	Error Free Data Construction.....	160
2.3	Quantification of Variability and Uncertainty with Measurement Error	162
3.0	Introduction to Study Design.....	167
3.0	Case Study	167
4.0	Conclusion	172
	Acknowledgements.....	174
	References.....	175
PART VI	QUANTIFICATION OF VARIABILITY AND UNCERTAINTY IN AIR POLLUTANT EMISSION INVENTORIES: METHOD AND CASE STUDY FOR UTILITY NO _x EMISSIONS	187
	Abstract.....	189
	Implications.....	190
1.0	Introduction.....	190
2.0	Methodology.....	193
2.1	Compilation and Evaluation of a Database.....	193
2.2	Visualization of Data	196
2.3	Fitting, Evaluating, and Selecting Parametric Probabilistic Distribution Models	197
2.4	Characterization of Uncertainty in the Distributions for Variability	199

2.5	Propagation of Uncertainty and Variability through a Model	200
2.6	Identifying Key Sources of Uncertainty	201
3.0	Introduction to AUVVE	202
4.0	Case Study: A Probabilistic Emission Inventory for Utilities in a Single State.....	204
5.0	Conclusions.....	210
	Acknowledgments.....	213
	References.....	214
	About the Authors.....	226
PART VII	PROBABILISTIC ANALYSIS OF DRIVING CYCLE-BASED HIGHWAY VEHICLE EMISSION FACTORS	227
	Abstract.....	229
1.0	Introduction.....	230
1.1	Sources of Variability and Uncertainty.....	231
1.2	Variability and Uncertainty in Highway Vehicle Emission Factors.....	232
1.3	Modeling Assumptions and Input Data	232
1.4	Brief Review of the Mobile5b Model.....	232
1.5	Simplified Probabilistic Emission Factor Model.....	233
1.6	Collection of Emission Test Data	235
2.0	Quantification of Inter-Vehicle Variability in Correction Factors	236
3.0	Quantification of Uncertainty in Mean Correction Factors	240
4.0	Quantification of Variability and Uncertainty in the Emission Factors .	242
4.1	Inter-Vehicle Variability in Emission Factors.....	243
4.2	Uncertainty in Mean Emission Factors.....	245
4.3	Identifying Key Sources of Uncertainty	246
5.0	Results and Discussion	247
6.0	Acknowledgments.....	249
7.0	References.....	250
PART VIII	CONCLUSIONS AND RECOMMENDATIONS	275
8.1	Summary.....	277

8.1.1	Methodologies.....	278
8.1.2	Case Studies.....	280
8.1.3	Software Development.....	281
8.2	Conclusions.....	282
8.2.1	Methodological Conclusions	282
8.2.2	Conclusions Based Upon Case Studies.....	284
8.3	Recommendations for Future Work.....	286
8.3.1	Methodologies.....	286
8.3.2	Probabilistic Emission Inventories	287
8.3.3	Development of AuvTool	288
APPENDIX A.....		293
APPENDIX B.....		297

List of Tables

PART II

Table 2-1. The Definitions of Parametric Probability Distributions.....	40
Table 2-2. Critical Value of D_n^α the Kolmogorov-Smirnov Test.....	53
Table 2-3. The Critical Values for Anderson-Darling test for Normal, Lognormal and Weibull distributions.....	55
Table 2-4. The Critical Values for Anderson-Darling test for the Gamma Distribution..	55

PART III

Table 3-1. AuvTool Function Module Summarization Table	87
--	----

PART IV

Table 1. Selected Population Mixture Lognormal Distributions with Two Components	136
Table 2. Comparison of 95% Confidence Intervals of Selected Statistics of Single and Two Component Mixture Lognormal Distributions Fitted to Mixture Populations with Varying Component Separation and Standard Deviation for $n=100$ and $w=0.3$	137
Table 3. Comparison of 95% Confidence Intervals of Selected Statistics of Single and Two Component Mixture Lognormal Distributions Fitted to Mixture Populations with Varying Component Separation and Standard Deviation for $n=100$ and $w=0.5$	138
Table 4. Uncertainty of Estimated Parameters of the Fitted Mixture Lognormal Distribution	136

PART V

Table 1. The Observed Data Set and Estimated Error Free Data Sets under Different Measurement Error Models	178
Table 2. The Fitted Lognormal Distributions under Different Measurement Error Models	179
Table 3. Uncertainty in the Mean under Different Measurement Error Models	179

PART VI

Table 1. Summary of 12-Month NO_x Emission and Activity Factors and of Fitted Distributions for Five Power Plant Technology Groups	218
--	-----

Table 2. Summary of Uncertainty Results for the Emission Inventory Case Study.....	219
--	-----

PART VII

Table 1. Characterization of Inter-Vehicle Variability in Estimated Tailpipe CO Emission Factors for Technology Group 8.....	252
Table 2. Characterization of Fleet Average Uncertainty in Estimated Tailpipe CO Emission Factors for Technology Group 8.....	253
Table S-2. Input Uncertainty Assumptions for CO Emissions.....	257
Table S-3. Input Variability Assumptions for HC Emissions.....	258
Table S-4. Input Uncertainty Assumptions for HC Emissions.....	259
Table S-5. Input Variability Assumptions for NO _x Emissions.....	260
Table S-6. Input Uncertainty Assumptions for NO _x Emissions.....	261
Table S-7. Characterization of Inter-Vehicle Variability in Estimated Tailpipe HC Emission Factors for Technology Group 8.....	262
Table S-8. Characterization of Fleet Average Uncertainty in Estimated Tailpipe HC Emission Factors for Technology Group 8.....	263
Table S-9. Characterization of Inter-Vehicle Variability in Estimated Tailpipe NO _x Emission Factors for Technology Group 8.....	264
Table S-10. Characterization of Fleet Average Uncertainty in Estimated Tailpipe NO _x Emission Factors for Technology Group 8.....	265
Table S-11. Correlation of Uncertain CO Emission Factor with Input Uncertainties....	266
Table S-12. Correlation of Uncertain HC Emission Factor with Input Uncertainties....	267
Table S-13. Correlation of Uncertain NO _x Emission Factor with Input Uncertainties..	268

List of Figures

PART II

Figure 2-1. Plot Illustrating the 95 Percent Probability Range on a Cumulative Distribution Function.	36
Figure 2-2. Example Graph of Visualizing Data Using the Hazen’s Plotting Position Method (n=10).....	37
Figure 2-3. An example of an Empirical Distribution Represented a Step Function (n=10).....	40
Figure 2-4. Comparison of Fitted Beta Distribution to an Example Dataset.....	51
Figure 2-5. An Illustrative Example of Graphical Comparison of Confidence Intervals for CDF of Fitted Distribution to the Data	56
Figure 2-6. Flow Diagram For Bootstrap Simulation and Two-Dimensional Simulation of Variability and Uncertainty. (Where: B=number of Bootstrap Replications, q=Sample Size Used for Uncertainty, p=Sample Size Used of Variability).....	71

PART III

Figure 3-1. The Conceptual Structure Design and Context Diagram of AuvTool System	86
Figure 3-3a. Batch Analysis Module (1).....	91
Figure 3-3b. Batch Analysis Module (2)	91
Figure 3-4. Conceptual Structure Design of the Analysis of Uncertainty and Variability in Emissions Estimation (AUVEE) Prototype Software.....	99

PART IV

Figure 1. Simplified Flow Diagram for Quantification of Variability and Uncertainty Using Bootstrap Simulation based upon Mixture Distributions	139
Figure 2. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0, \sigma_1=0.5, \mu_2=1.5, \sigma_2=0.5$) for n=25,50 and 100, for w=0.1,0.3 and 0.5 Based on Bootstrap Simulation (B=500).....	140
Figure 3. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0, \sigma_1=0.5, \mu_2=2.0, \sigma_2=0.5$) for n=25,50 and 100, for w=0.1,0.3 and 0.5 Based on Bootstrap Simulation (B=500).....	141
Figure 4. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population	

	Distribution ($\mu_1=1.0, \sigma_1=0.5, \mu_2=3.0, \sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$).....	142
Figure 5.	95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0, \sigma_1=0.5, \mu_2=6.0, \sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$).....	143
Figure 6.	95 Percent Confidence Intervals of Cumulative Distributions of a Single Lognormal Distribution Fitted to a Mixture Population Distribution ($\mu_1=1.0, \sigma_1=0.5, \mu_2=1.5, \sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$).....	144
Figure 7.	95 Percent Confidence Intervals of Cumulative Distributions of a Single Lognormal Distribution Fitted to a Mixture Population Distribution ($\mu_1=1.0, \sigma_1=0.5, \mu_2=2.0, \sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$).....	145
Figure 9.	Scatter Plots of Bootstrap Simulation ($B=500$) Results for Parameters of Two Component Lognormal Mixture Distributions for $n=100, w=0.5$ with Moderately Separated Components ($\mu_1=1.0, \sigma_1=0.5, \mu_2=3.0, \sigma_2=0.5$).	147
Figure 10.	Scatter Plots of Bootstrap Simulation ($B=500$) Results for Parameters of Two Component Lognormal Mixture Distributions for $n=100, w=0.5$ with highly Separated Components ($\mu_1=1.0, \sigma_1=0.5, \mu_2=6.0, \sigma_2=0.5$).	148
Figure 11.	Mixture lognormal distribution fitted to six-month average NO_x emission factor data for T/LNC1 technology group ($n=41$).	149
Figure 12.	Probability band for fitted mixture lognormal distribution.	149

PART V

Figure 1.	Flow Diagram for Characterizing Variability and Uncertainty with Measurement Error by Using the Bootstrap Pair Technique	165
Figure 2.	The Fitted Lognormal Distributions to the Error Free Data Set under Different Measurement Errors and the Observed Data Set.....	180
Figure 3.	The Probability Band Based upon the Fitted Lognormal Distribution to the Observed Data Set (no measurement error is assumed).....	180
Figure 4.	The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=10.0$ without the Inclusion of Measurement Error	181
Figure 5.	The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=10.0$ with the Inclusion of Measurement Error	181
Figure 6.	The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=20.0$ without the Inclusion of Measurement Error	182
Figure 7.	The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=20.0$ with the Inclusion of Measurement Error	182

Figure 8. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=40.0$ without the Inclusion of Measurement Error	183
Figure 9. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=40.0$ with the Inclusion of Measurement Error	183
Figure 10. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=55.5$ without the Inclusion of Measurement Error	184
Figure 11. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_e=55.5$ with the Inclusion of Measurement Error	184
Figure 12. The Sampling Distributions for the Mean under Different Measurement Error Models.....	185

PART VI

Figure 1. Scatter plot of 12-month NO_x Emission Rate of 1997 and 1998 (No. of Data=390).....	220
Figure 2. Scatter Plot for 12-month Average Heat Rate versus 12-month Average Capacity Factor for Tangential-Fired Boilers Using Low NO_x Burners and OverfireAir Option 1. (n=36).....	220
Figure 3. Conceptual Design of the Analysis of Uncertainty and Variability in Emissions Estimation (AUVVE) Prototype Software.....	221
Figure 4. Probability Band for Distribution Fitted to Example Heat Rate Data for Dry Bottom Wall-fired Boilers Using Low NO_x Burners (n=98).....	222
Figure 5. Probability Band for Distribution Fitted to Example Capacity Factor Data for Dry Bottom Wall-fired Boilers Using Low NO_x Burners (n=98).....	222
Figure 6. Probability Band for Distribution Fitted to Example NO_x Emission Rate Data for Dry Bottom Wall-fired Boilers Using Low NO_x Burners (n=98)....	223
Figure 7. Probability Bands Based Upon Number of Units in the Emission Inventory (n=3) for the Example of NO_x Emission Rate.....	223
Figure 8. Uncertainty in a 12-Month NO_x Emission Inventory for an Individual Technology Group Comprised of 3 Units.....	224
Figure 9. Uncertainty in a 12-Month NO_x Emission Inventory Inclusive of Four Technology Groups.....	224
Figure 10. Relative Importance of Uncertainty in Emissions from Individual Technology Groups with Respect to Overall Uncertainty in the Total Emission Inventory.....	225

PART VII

Figure 1. Estimated Inter-Vehicle Variability in Temperature Correction Factor For CO Technology Group 8 254

Figure 2. Estimated Fleet Average in Temperature Correction Factor For CO Technology Group 8..... 254

Figure S-1. Estimated Inter-Vehicle Variability in Temperature Correction Factor For HC Technology Group 8 269

Figure S-2. Estimated Fleet Average in Temperature Correction Factor For HC Technology Group 8..... 269

Figure S-3. Estimated Inter-Vehicle Variability in Temperature Correction Factor For NO_x Technology Group 8 270

Figure S-4. Estimated Fleet Average in Temperature Correction Factor For NO_x Technology Group 8..... 270

Figure S-5. Estimated Inter-Vehicle Variability in RVP Correction Factor For CO Technology Group 8..... 271

Figure S-6. Estimated Fleet Average in RVP Correction Factor For CO Technology Group 8..... 271

Figure S-7. Estimated Inter-Vehicle Variability in RVP Correction Factor For NO_x Technology Group 8..... 272

Figure S-8. Estimated Fleet Average in RVP Correction Factor For NO_x Technology Group 8..... 272

Figure S-9. Estimated Inter-Vehicle Variability in RVP Correction Factor For HC Technology Group 8..... 273

Figure S-10. Estimated Fleet Average in RVP Correction Factor For HC Technology Group 8..... 273

APPENDIX A

Figure 1. Example of a Probability Plot for a Weibull Distribution (n=50)..... 295

PART I

INTRODUCTION

Junyu Zheng

1.0 Introduction

Air pollutant emission inventories (EIs) are a vital component of environmental decision-making. They are often used for short-term or long-term emission trend characterization, emission budgeting for regulatory and compliance purposes, and the predication of ambient pollutant concentrations using air quality models. If random errors and biases in emission inventories are not quantified, they can lead to erroneous conclusions regarding trends in emissions, source apportionment, compliance, and the relationship between emissions and ambient air quality (Frey *et al.*, 1999). For example, if resources are mistakenly devoted to reduce emissions for a source category where emissions are overestimated, or if resources are not applied to reduce emissions from a source category where emissions are under-estimated, then air quality objectives cannot be achieved in an efficient and cost-effective manner.

In practice, an emission inventory is often obtained from emission factors multiplied by activity factors. Emission factors are typically assumed to be representative of an average emission rate from a population of pollutant sources in a specific category (EPA,1995). For example, a power plant emission factor can be an average emission rate among all of the power plant units of the same type within a state. However, there may be uncertainty in the population average emission rates because of random sampling error, measurement errors, or possibly because the sample of power plants from which the emission factor was developed was not a representative sample. These errors in the estimation of the emission factor or activity factor can lead to biases in emission inventory estimation.

Some qualitative analysis methods have been presented in the past years, including conventional data quality rating method and the Data Attribute Rating System

(DARS) to address the errors in the estimation of the emission factor or activity factor. For example, in the data quality rating methods, qualitative “A” through “E” ratings are defined and reported in EPA’s Compilation of Air Pollutant Emission Factors (EPA, 1995). DARS is a method for combining data quality scores for both emission factor and activity data to develop an overall quality score for an emission inventory (Beck and Wilson, 1997). While DARS can be used to compare quality ratings for EIs, it can neither be used to quantify the precision of an inventory nor to evaluate the robustness of a decision to uncertainty. Other efforts have been focused on characterizing the mean and variance of emission estimates and using simplified approaches for combining uncertainties in activity and emission factor data to arrive at an aggregate uncertainty estimate (Dickinson and Hobbs, 1989; NRC 1991; Balentine and Dickinson, 1995). The applications of these approaches suffer from many shortcomings including: restrictive assumptions about the shape of probability distribution models; failure to deal with dependences between uncertainty estimates; failure to distinguish between variability and uncertainty estimates; inappropriate averaging times; improperly analyzed small sample data; and failure to use proper protocols in eliciting expert judgments.

The use of quantitative methods for dealing with variability and uncertainty is becoming more widely recognized and recommended for environmental modeling and assessment applications. For example, the National Research Council recently released a report on mobile source emissions estimation that calls for new efforts to quantify uncertainty in emissions (NRC, 2000). Quantification of variability and uncertainty has become widely accepted in human health risk assessment. The U.S. Environmental Protection Agency (US EPA), for example, has sponsored workshops regarding Monte

Carlo simulation methods (EPA, 1997, 1999a), has developed a guideline document on Monte Carlo methods (EPA, 1996) and has included guidance regarding probabilistic analysis in its most recent draft of Risk Assessment Guidance for Superfund (EPA, 1999b). One of the recommendations of the Emission Inventory Improvement Program (EIIP) of the US EPA is to encourage the use of quantitative methods to characterize variability and uncertainties in emission inventories (Radian, 1996). Uncertainty analysis is now part of the planning process for major assessments performed by EPA, such as the National Air Toxics Assessment.

The Intergovernmental Panel on Climate Change (IPCC) recently issues a good practice document regarding uncertainty analysis for greenhouse gas emission inventories (IPCC, 2000).

1.1 Variability

Variability is the heterogeneity of a quantity over time, space or members of a population. Thus variability indicates the range that a quantity can vary over. Variability exists in the every aspect in the quantification of emissions or exposure assessment; for example, in the quantification of highway emission factor, process variability leads to difference in emissions as a function of vehicle design (inter-vehicle variability) and operating conditions (intra-vehicle variability). For an individual vehicle, emissions may vary with time due to change in fuel composition, ambient temperature, engine load, random variation in throttle position, failure of pollution control systems, maintenance practice, and so on. In exposure assessments, common sources of variability are different in characteristics between individuals; for example, a given human individual has a body weight, intake rate, lifetime, exposure duration, and activity patterns that are different from that of other individuals. Variability can be represented

by a frequency distribution showing the variation in a characteristic of interest over time, space. Knowledge of the frequency distribution helps to assess whether a population needs to be subdivided into groups which are more nearly homogeneous. (Frey,1997; Cullen, Frey, 1999; Morgan and Henrion,1990)

1.2 Uncertainty

Uncertainty refers to a lack of knowledge about the true value of a quantity.

Uncertainty can be quantified as a probability distribution representing the likelihood that the unknown quantity falls within a given range of values (Cullen and Frey, 1999).

Uncertainty can be introduced in every step of analyzing a quantity. For example, it can come from measurement error because of biases in the measuring apparatus and the experimental procedures or from human error, such as random mistakes in recording or processing data; and random sampling error. Uncertainty exists in the whole model building process (Cullen and Frey, 1999). Draper et al. (1987) and Hodges (1987) pointed out that there are typically three main sources of uncertainty in any problems:

- (a) Uncertainty about the structure of the model;
- (b) Uncertainty about estimates of the model parameters (or model inputs), assuming that we know the structure of the model;
- (c) Unexplained random variation in observed variables even when we know the structure of the model and the values of the model parameters.

The structure of mathematical models employed to represent scenarios and phenomena of interest is often a key source of uncertainty, due to the fact that models are often only a simplified representation of a real-world system. In practice, though model uncertainty is a fact of life and likely to be more serious than other sources of uncertainty which have received far more attention from researchers or data analysts (Cullen, Frey,

1999), little study has been done for model uncertainty, even by statisticians (Chatfield, 1995). There are various sources of model uncertainty which include model structure, model detail, validation and verification, extrapolation, resolution in numerical analysis and model boundaries (Frey, 1992). Chatfield (1995) suggested that uncertainty about model structures can arise from: (1) model misspecification (e.g., omitting a variable by mistake); (2) specifying a general class of models of which the true model is a special, but unknown, case or (3) choosing between two or more models of quite different structures. More description about the sources of model uncertainty can be found in Cullen and Frey (1999).

One approach for addressing model uncertainty is to compare predictions made with alternative models. For example, Evans et al. (1994) present a probability tree in which alternative conceptual models are included. In general, a probability model with a better representative of a quantity will help reduce the uncertainty of the model. Therefore, the process for selection or evaluation of goodness-of-fit of a probability model fitted to a dataset is a way to improve the uncertainty analysis by reducing model uncertainty.

Model uncertainty can also be reduced by valiant simulation and replicate experiments or collecting additional data and the Bayesian model average approach. The Bayesian model average approach, presented by Chatfield (1995), which does not require choosing a single best model but rather averages over a variety of plausible competing models which are entertained with appropriate prior probabilities, might offer more promise; however, there are some limitations in applications for this approach (Chatfield, 1995). For example, (1) the efforts for trying the various plausible competing models are

very heavy; (2) it is not known how many plausible competing models are available. Maybe a safer way to proceed with model uncertainty is to replicate the study and to collect additional data to cope with model uncertainty, however this is not always possible, especially in situations where data are hard to collection or data collection are very expensive. Because the task of finding ways to address model uncertainty has only just begun, even for statisticians, this dissertation is not intended to focus on the uncertainty arising from model structure.

A number of different types of quantities are used in models. They can be empirical quantities which is measurable or at least in principle, defined constants, decision variables, value parameters which represents the preference or value judgments of a decision-maker and model domain parameters (Morgan and Henrion, 1990). Of all these model inputs, only empirical quantities are unambiguously subject to uncertainty. (Cullen and Frey, 1999). Thus, we focus here on identifying sources of uncertainty in empirical quantity model inputs.

1.3 Distinctions Between Variability and Uncertainty

Model inputs can be categorized according to the roles they serve (Cullen and Frey, 1999). For example, in emission inventory models, the model inputs can be divided into those that are variable, those that are uncertain, and those with some aspects of each (Bogen and Spear, 1987; IAEA, 1989; Morgan and Henrion, 1990; Finkel, 1990; Frey, 1992). In formal analysis, variability and uncertainty are characterized by two different probability spaces (Helton, 1996; Helton *et al.*, 1996). For example, variability can be represented by a frequency distribution showing the variation in a characteristic of interest over time, space (Frey, 1997). Uncertainty can be quantified as a probability

distribution representing the likelihood that the unknown quantity falls within a given range of values (Frey, 1997).

Variability and uncertainty should be treated separately because they each have different decision-making and policy implications (Frey *et al.*, 2002). For example, in risk assessment, uncertainty forces decision-makers to judge how probable it is that risks will be overestimated or underestimated for every member of the exposed population, whereas variability forces them to cope with the certainty that different individuals will be subjected to risks that both above and below any reference point one chooses. Information regarding key sources of uncertainty can be used to evaluate whether times series trends are statistically significant or not, to determine the likelihood that an emission budget will be met, and to prioritize additional data collection or research to improve estimates of emissions. Understanding variability can guide the identification of significant subpopulations that need more focused study. For example, knowledge of inter-unit variability in emissions from utility emission source will help to identify which unit makes most contribution, and hence need more improvement in control strategy and technology.

Characterization of variability and uncertainty in emission inventories using the probabilistic techniques will enable funding and regulatory agencies to target money and effort to activities that help identification of significant subpopulations contributing most to the total emissions and ones that will result in the greatest reductions in inventory uncertainties. The knowledge of variability and uncertainty in emission inventories can enables specific identification of the likelihood that a specific portion of an emission source population will comply with a particular standard or target. For example, the

result of a probabilistic analysis might indicate that there is a 95 percent probability that 90 percent of the total number of utility emission sources would comply with a proposed emission standard of 10 tons per year, or that a specific individual facility may have a 90 percent probability of compliance. This information could then be used to determine whether additional control measures might be required to increase the probability of compliance.

The National Research Council has recommended that the distinction between variability and uncertainty should be maintained rigorously at the level of individual components of a risk assessment as well as at the level of an integrated risk assessment (NRC, 1994).

1.4 Examples of Probabilistic Analysis

There is a growing track record of the demonstrated use of quantitative methods for characterizing variability and uncertainty applied to emission factors, emission inventories, air quality modeling, exposure assessment, and risk assessment. There have been a number of projects aimed at quantifying variability and uncertainty in highway vehicle emissions, including uncertainty estimates associated with the Mobile5a emission factor model and with the EMFAC emission factor model used in California (Kini and Frey, 1997; Frey, 1997; Frey *et al.*, 1999; Pollack *et al.*, 1999). Frey and Eichenberger (1997) and Frey *et al.* (2001) have quantified uncertainty in highway vehicle emission factors estimated based upon measured data collected using remote sensing and on-board instrumentation, respectively. There have been a number of efforts aimed at probabilistic analysis of various other emission sources, including power plants, non-road mobile sources, natural gas-fired engines, and specific area sources (Frey, Rhodes, 1996; Frey *et al.*, 1999; Frey and Zheng, 2000; Frey and Bammi, 2002a&b; Frey, and Zheng, 2002;

Frey and Bharvirkar, 2002; Li and Frey, 2002, Abdel-Aziz and Frey, 2002). Probabilistic analyses have also been applied to air quality models, such as the Urban Airshed Model (e.g., Hanna *et al.*, 2001). In the area of exposure and risk assessment, there have been a number of analyses in which variability and uncertainty were distinguished. These include, for example, Bogen and Spear (1987), Frey (1992), Hoffman and Hammonds (1996), Cohen *et al.* (1996), and others.

As an example of a probabilistic analysis in which variability and uncertainty were distinguished, Frey and Rhodes (1996) quantified variability and uncertainty in emissions of selected hazardous air pollutants from coal-fired power plants. Limited data were available regarding the concentration of trace species, such as arsenic in coal, and regarding the partitioning of the trace species in the major process areas of the plant, including the boiler, particulate matter control device, and flue gas desulfurization system. Parametric distributions were fitted to the available data that represented the inter-unit variability in plant performance. Bootstrap simulation was used to estimate confidence intervals for the fitted cumulative distribution function (CDF) for each input data set. Both variability and uncertainty were propagated through an emissions model to yield estimates of variability in emissions from one averaging time to another and uncertainty in emissions for any given simulated averaging period.

1.5 Limitations of Current Studies in Variability and Uncertainty Analysis

As described in Section 1.4, the quantitative methods for characterizing variability and uncertainty have been used in the development of air pollutant emission inventories from selected emission sources, including power plants, highway vehicle emission sources, non-road mobile sources, and natural gas-fired engines. However, there still exist some limitations in these methods and applications. For example, (1) The methods

for dealing with variability and uncertainty are focused on the use of single component distributions; (2) among main sources of uncertainty, only random sampling errors are characterized; the uncertainty arising from measurement error, another source of uncertainty, is not quantified; (3) most probabilistic developments of air pollutant emission inventories are done for a particular source category, there are no general frameworks which extensively and systematically summarize or introduce the associated methodological issues to address the variability and uncertainty in the development of a probabilistic emission inventory.

In most applications, a single distribution model such as normal or lognormal distribution is good enough to describe an emission factor or activity factor. However, in some cases, a single component distribution model might not provide a good representative to describe the variation or uncertainty of a quantity. Because the accuracy of quantifying variability and uncertainty in part depends on the goodness of fit of the distributions with respect to the available data, the use of single component distributions that are poor fits to data will lead to bias in the quantification of variability and uncertainty (Zheng and Frey, 2001). However, in these cases, an alternative is to use a finite mixture of distributions.

Mixture distributions have been extensively used as models in a wide variety of important practical situations because they can provide a powerful way to extend common parametric families of distribution to fit datasets not adequately fit by a single common parametric distribution. Mixture models have been used in the physical, chemical, social science, biological and other fields. As examples, Hariris (1983) applied mixture distributions to modeling crime and justice data, and Kanji (1985) described

wind shear data using mixture distributions. In human exposure and risk assessment, Burmaster (1994) used mixture lognormal models to re-analyze data sets collected by the U.S. EPA for the concentration of Radon²²² in drinking water supplied from ground water, and found that the mixture model yielded a high-fidelity fit to the data not achievable with any single parameter distributions.

However, there is little study on quantification of variability and uncertainty based on mixture distributions. Because a mixture distribution often has a more complicated mathematical form and has more parameters needed to be estimated, there are additional questions to address when developing an approach to address variability and uncertainty with the use of mixture distributions. For example, (1) how are parameters in mixture distributions estimated? Unlike single parameter distributions, there are often no analytical parameter estimators available for finite mixture distributions; (2) because no random sampling formulas and cumulative probability functions are available for any finite mixture distributions, how is a bootstrap sample drawn from a mixture distribution?

Any emission data must be collected by measuring instruments. Measurement instruments are created by humans, and every measurement is an experimental procedure. The results of measurements cannot be absolutely accurate. Quantitatively the measurement bias is characterized by the notion of either limits or uncertainty. Uncertainty of measurement is an interval within which a true value of the measurement lies with a given probability (Rabinovich, 1999).

As previously described, uncertainty arises from the random sampling errors and the imperfections in measurement techniques; the latter is often referred to as

measurement error. Measurement error is another main source of variability and uncertainty in emission estimation. However, there is little study relevant to the quantification of variability and uncertainty in emission estimation due to the error or bias caused by imperfections in measurement imperfections. Rabinovich (1999) pointed out that it is possible to separate these sources of variation or uncertainty due to the imperfections in measurement techniques or procedures from the observed values, and to propagate them separately through a model for the cases where the measurement error is known or can be reasonably estimated based upon expert judgment.

1.6 Available Software Tools in Probabilistic Analysis

A variety of programs have been developed that are capable of various types of probabilistic analysis. There are several commercially available software packages, such as Crystal Ball, @Risk, Analytica and RiskQ. Crystal Ball and @Risk both are Microsoft Excel-based add-in programs (Palisades, 1997; Decisioneering, 2001). Analytica is a stand-alone program for creating, analyzing, and communicating probabilistic models for risk and policy analysis (Lumina, 1996). RiskQ is implemented in Mathematica (Bogen, 1992). Capabilities to address both variability and uncertainty are available in Crystal Ball and RiskQ. While RiskQ has many powerful capabilities, it requires knowledge of programming in Mathematica (Murray and Burmaster, 1993). @Risk and Analytica do not provide convenient capabilities for simultaneous analysis of both variability and uncertainty. Crystal Ball uses a two-stage Monte Carlo simulation method as presented by Cohen *et al.*, 1996. The method of Cohen *et al.* (1996) is very similar to that of Frey (1992) and Frey and Rhodes (1996). The primary difference is that the approach of Cohen *et al.* discards many intermediate values during the simulation. While

this can reduce memory or storage requirements, it also results in the loss of useful information and the limitations in applications.

Frey and Rhodes (1996, 1998, 1999) developed a FORTRAN-based program at North Carolina State University referred to as "BOOTSIM." BOOTSIM featured two-dimensional probabilistic representations of variability and/or uncertainty for model inputs, propagation of the two-dimensional probabilistic information through a model, characterization of both variability and uncertainty in model results, and analysis of model results to identify key sources of variability and uncertainty.

BOOTSIM did not contain a capability to fit a parametric probability distribution to a data set or to compare alternative fitted distributions to data, and did not have Graphical User Interface (GUI) to allow users to input data and visually select a good fit, and did not include the use of statistical goodness-of-fit tests nor capabilities for dealing with the issues associated with mixture distributions and measurement errors.

1.7 Objectives

Based on limitations discussed above in the methodologies, applications and the software tools to address variability and uncertainty, the objectives of this dissertation are described below:

1. To summarize and further develop general methodologies on quantification of variability and uncertainty in emission inventories. These methods will feature the use of bootstrap simulation, and address the issues associated with mixture distribution and measurement errors in variability and uncertainty analysis;

2. To evaluate and investigate the properties of quantification of variability and uncertainty based on mixture distribution with respect to sample size, mixing weight and separation between components;
3. To develop a user-friendly prototype software tool to demonstrate the development of a probabilistic emission inventory for a selected emission source. This prototype tool is named as AUVVEE in this dissertation;
4. To develop a user-friendly software tool with graphic user interface which is generally applicable for quantifying variability and uncertainty in model inputs for emission estimation, risk or exposure assessment and other quantitative analysis fields. The software tool can be capable of fitting distributions (including single, empirical and mixture distributions) to datasets (including datasets with known measurement error); of characterizing uncertainty in the distribution for variability by featuring the use of bootstrap simulation. The software tool is named as AuvTool (Analysis of Variability and Uncertainty Tool) in this study;
5. To apply the methodologies developed in this study to some case studies such as quantification of variability and uncertainty in highway emission factors, development of probabilistic emission inventories for utility power plant emission sources.

1.8 Overview of Research

The research of this dissertation focused on three aspects: methodologies, software implementation of methodologies and example case studies used to demonstrate the use of methodologies and software tools developed in this dissertation. The

methodologies developed in this research include a general approach for calculating a probabilistic emission inventory, which is based upon the work done by Frey and Bharvirkar (Frey *et al.*, 1999); the methods for quantifying variability and uncertainty analysis based on the use of mixture distributions and methods for improving variability and uncertainty if there are known measurement errors in a dataset.

Two software tools were developed in this research, one is AUVÉE, and another is AuvTool. AUVÉE is a prototype software tool for demonstrating the use of the general approach to develop a probabilistic emission inventory. AUVÉE is developed based on the BOOTSIM (Frey and Rhodes, 1998). However, BOOTSIM did not contain a capability to fit a parametric probability distribution to a data set or to compare alternative fitted distributions to data, and did not have GUI to allow users to input data and visually select a good fit, and did not have the internal database and user databases to support the development of a probabilistic emission inventory for the utility power plant emission source category. AuvTool is a general tool for doing variability and uncertainty analysis for model inputs. Its purposes are to implement all methods developed in this study and to make it generally applicable for any applications where variability and uncertainty are needed.

This dissertation features new methodological contributions regarding mixture distribution and measurement errors. Part IV of this dissertation deals with the use of mixture distributions for doing variability and uncertainty analysis if a mixture distribution is used to represent a data set. The properties of quantification of variability and uncertainty were evaluated with respect to variation in sample size, mixing weight and separation between components. Part V of this dissertation demonstrates the

methods for improving variability and uncertainty estimates if there are known measurement error in an observed dataset. The effect of measurement error on quantification of variability and uncertainty was evaluated and investigated with respect to the size of measurement error. Part IV and V are based in large part upon the independent contribution of the author.

A case study was done to demonstrate the development of probabilistic emission inventories for utilities in a single state. The case study was partly based upon the work done by Frey and Bharvirkar (Frey *et al.*, 1999). However, a lot of improvements have been made in this research. These include improvement for the completeness of the general approach, update of database and software implementation of a prototype software tool, AUVÉE, to develop a probabilistic emission inventory. The case study for quantifying variability and uncertainty in highway vehicle emission factors was an extension of the work done by Kini and Frey (1997). The new work reported in the case study dealt with the quantification of variability and uncertainty in temperature correction factor and fuel Reid vapor pressure correction factor in the MOBILE5b model. Inter-vehicle variability and average fleet uncertainty in HC, CO and NO_x emission factors based upon driving cycles, with the incorporation of variability and uncertainty from the two correction factors, was investigated.

1.9 Organization

This dissertation will first presented general methodologies for quantifying variability and uncertainty in emission estimation, which is given in Part 2 of this dissertation, and then introduce software implementations of the accompanying software tools AUVÉE and AuvTool developed in this research, which is documented in Part 3 of this dissertation. Four manuscripts that the author has submitted or plans to submit for

publication in peer-reviewed journals will be presented in Part 4 through Part 7 of this dissertation, respectively.

The paper given in Part 4 of this dissertation provides a discussion on the properties in quantifying variability and uncertainty with the use of mixture distributions. The manuscript presented in Part 5 presents the methodologies on quantification of variability and uncertainty if there are known measurement errors in an observed data set; a case study is used to illustrate the use of the methods and the effect of measurement error on variability and uncertainty analysis.

The manuscript given in Part 6 of this dissertation demonstrated the methodology in developing a probabilistic air pollutant emission inventory and an example case study based on utility NO_x emission was presented. Part 7 of this dissertation demonstrated a probabilistic analysis approach for quantifying inter-vehicle variability and fleet average uncertainty in highway vehicle emission factors. Finally, the conclusions of this study, and the recommendations for future studies are presented in Part 8.

Each part of this manuscript has its own list of references cited.

1.10 References

Abdel-Aziz, A., and H.C. Frey, "Quantification of Variability and Uncertainty in Hourly NO_x Emissions from Coal-Fired Power Plants," *Proceedings, Annual Meeting of the Air & Waste Management Association*, Pittsburgh, PA, June 2002 (in press).

Balentine, H.W., and Dickson, R.J., 1995, "Development of Uncertainty Estimates For the Grand Canyon Visibility Transport Commission Emissions Inventory," In *The Emission Inventory: Programs and Progress, The Proceedings of A Specialty Conference*, Air & Waste Management Association: Pittsburgh, PA, pp. 407-425.

Beck, L.; Wilson, D. (1997), "EPA's Data Attribute Rating System," In *Emission Inventory: Planning for the Future, The Proceedings of A Specialty Conference, Air & Waste Management Association*, Pittsburgh, PA, pp. 176-189.

Bogen, K.T., and Spear, R.C., 1987, "Integrating Uncertainty and Interindividual Vaiability in Environmental Risk Assessment," *Risk Analysis*, 7(4): 427-436.

Bogen, K.T., 1992, *RiskQ: An Interactive Approach to Probability, Uncertainty, and Statistics for Use with Mathematica (Reference Manual)*, Lawrence Livermore National Laboratory, Livermore, CA.

Burmaster, D.E., R.H. Harris, 1994, "The Magnitude of Compounding Conservatism in Superfund Risk Assessments", *Risk Analysis*, 13(2):131-143.

Chatfield, C., 1995, "Model Uncertainty, Data Mining and Statistical Inference," *J. of the Royal Statistical Society, Series A*, 158 (3), pp419-466.

Cohen, J.T., M.A., Lampson, and S. Bowers, 1996, "The Use of Two-Stage Monte Carlo Simulation Techniques to Characterize Variability and Uncertainty in Risk Analysis," *Human and Ecological Risk Assessment*, 2(4): 939-971.

Cullen, A.C., H.C. Frey, 1999, *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press: New York.

Decisioneering, <http://www.decisioneering.com> (Accessed 01/20/2001).

Dickson, R.J. and Hobbs, A.D. (1989), "Evaluation of Emission Inventory Uncertainty Estimation Procedures," Paper No. 89-24.8, In *82nd Annual Meeting*, Air & Waste Management Association: Anaheim, CA.

Draper, D., Hodges, J.S., Leamer, E.E., Morris, C.N. and Rubin, D.B., 1987, "A Research Agenda for Assessment and Propagation of Model Uncertainty," *Report N-2683-RC*, Rand Corporation, Santa Monica.

- EPA, 1995, *Compilation of Air Pollutant Emission Factors 5th Ed., AP-42 and Supplements*, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC.
- EPA, 1996, *Summary Report for the Workshop on Monte Carlo Analysis, EPA/630/R-96/010, Risk Assessment Forum*, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC.
- EPA, 1997, *Guiding Principles for Monte Carlo Analysis*, EPA/630/R-97/001, U.S. Environmental Protection Agency, Washington, DC.
- EPA, 1999a, *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessment*, EPA/630/R-98/004, U.S. Environmental Protection Agency, Washington, DC.
- EPA, 1999b, *RAGS 3A - Process for Conducting Probabilistic Risk Assessment*, Draft, U.S. Environmental Protection Agency, Washington, DC.
- Evans, J.S., Graham, J.D., Gray, G.M., and Sielken, R.L., 1994a, "A Distributional Approach to Characterizing Low-Dose Cancer Risk," *Risk Analysis*, 14(1):25-34.
- Finkel, A.M., 1990, *Confronting Uncertainty in Risk Assessment: A Guide for Decision Makers*, Center for Risk Management, Resources for the Future, Washington, DC.
- Frey, H.C., 1992, "Quantitative Analysis of Uncertainty and Variability in Environmental Policy Making," Directorate for Science and Policy Programs, American Association for the Advancement of Science, Washington, DC.
- Frey, H.C., 1997, "Variability and Uncertainty in Highway Vehicle Emission Factors," in *Emission Inventory: Planning for the Future* (held October 28-30 in Research Triangle Park, NC), *Air and Waste Management Association*, Pittsburgh, Pennsylvania, pp. 208-219.
- Frey, H.C., and S. Bammi, 2002a, "Quantification of Variability and Uncertainty in Lawn and Garden Equipment NO_x and Total Hydrocarbon Emission Factors," *Journal of the Air & Waste Management Association*, accepted January 2002 for publication.
- Frey, H.C., and S. Bammi, 2002b, "Probabilistic Nonroad Mobile Source Emission Factors," *Journal of Environmental Engineering*, tentatively accepted pending revisions.
- Frey, H.C., and R. Bharvirkar, 2002, "Quantification of Variability and Uncertainty: A Case Study of Power Plant Hazardous Air Pollutant Emissions," Chapter 10 in *Human and Ecological Risk Analysis*, D. Paustenbach, Ed., John Wiley and Sons: New York. (In press).

Frey, H.C., R. Bharvirkar, J. Zheng, 1999, "Quantitative Analysis of Variability and Uncertainty in Emissions Estimation," Final Report, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Frey, H.C., and D.A. Eichenberger,,1997, *Remote Sensing of Mobile Source Air Pollutant Emissions: Variability and Uncertainty in On-Road Emissions Estimates of Carbon Monoxide and Hydrocarbons for School and Transit Buses*, FHWY/NC/97-005, Prepared by North Carolina State University for North Carolina Department of Transportation, Raleigh.

Frey, H.C., D.S. Rhodes, 1996, "Characterizing, Simulating, and Analyzing Variability and Uncertainty: An Illustration of Methods Using an Air Toxics Emissions Example," *Human and Ecological Risk Assessment*, 2(4):762-797.

Frey, H.C., D.S. Rhodes, 1998, "Characterization and simulation of uncertain frequency distributions: Effects of Distribution Choice, Variability, Uncertainty, and Parameter Dependence," *Human and Ecological Risk Assessment*, 4(2):423-468.

Frey, H.C., N.M. Roupail, A. Unal, and J.D. Colyar, "Emissions Reduction Through Better Traffic Management: An Empirical Evaluation Based Upon On-Road Measurements," FHWY/NC/2002-001, Prepared by Department of Civil Engineering, North Carolina State University for North Carolina Department of Transportation, Raleigh, NC. December 2001

Frey, H.C., J. Zheng, 2000, "Methods and Example Case Study for Analysis of Variability and Uncertainty in Emissions Estimation (AUVVE)," Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Frey, H.C., J. Zheng, 2002, "Quantification of Variability and Uncertainty in Utility NO_x Emission Inventories," *J. of Air & Waste Manage. Assoc.*, accepted for publications.

Harris, C.M., 1983, "On finite mixtures of geometric and negative binomial distributions," *Commun, Statist.-Ther. Meth.* 12:987-1007.

Helton, J.,1996, "Probability, Conditional Probability and Complementary Cumulative Distribution Functions in Performance Assessment for Radioactive Waste Disposal," Sandia National Laboratories, Albuquerque, NM.

Helton et al., 1996, "Computational Implementation of a System Prioritization Methodology for the Waste Isolation Pilot Plant: A Preliminary Example," Sandia National Laboratories, Albuquerque, NM.

Hodges, J.S. (1987), "Uncertainty, Policy Analysis and Statistics," *Statistical Science*, 2, 259-291.

Hoffman, F.O., J.S. Hammonds, 1994, "Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability," *Risk Analysis*, 14(5):707-712.

IAEA 1989, "Evaluating the Reliability of Predictions Made Using Environmental Transfer Models," *Safety Series*, No. 100, International Atomic Energy Agency Vienna, Austria.

IPCC, 2000, *Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories*. National Greenhouse Gas Inventories Program, Intergovernmental Panel on Climate Change (IPCC), 2000.

Kanji, G.K., 1985, "A mixture model for wind shear data," *J. Appl. Statist.*, 12:49-58

Kini, M.D., and H.C. Frey, 1997, *Probabilistic Evaluation of Mobile Source Air Pollution: Volume 1, Probabilistic Modeling of Exhaust Emissions from Light Duty Gasoline Vehicles*, Prepared by North Carolina State University for Center for Transportation and the Environment, Raleigh, December.

Li, S., and H.C. Frey, 2002, "Methods and Example for Development of a Probabilistic Per-Capita Emission Factor for VOC Emissions from Consumer/Commercial Product Use", *Proceedings, Annual Meeting of the Air & Waste Management Association*, Pittsburgh, PA, June 2002 (in press).

Lumina, 1996, *Analytica User Guide*, Lumina Decision Systems, Los Altos, CA.

Morgan, M.G., and M. Henrion, 1990, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York.

Murray, D.M., D.E. Burmaster, 1993, "Review of RiskQ: An Interactive Approach to Probability, Uncertainty, and Statistics for Use with Mathematica", *Risk Analysis*, 13(4): 479-482.

NRC, 1991, *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, National Academy Press, Washington, D.C.

NRC, 1994, *Science and Judgment in Risk Assessment*, National Research Council, National Academy Press, Washington, D.C.

NRC, 2000, *Modeling Mobile Source Emissions*, National Research Council, National Academy Press, Washington, D.C.

Palisades, <http://www.palisade.com> (Accessed 01/20/2001).

Pollack, A.K., P. Bhave, J. Heiken, K. Lee, S. Shepard, C. Tran, G. Yarwood, R.F. Sawyer, and B.A. Joy, 1999, *Investigation of Emission Factors in the California EMFAC7G Model*, PB99-149718INZ, Prepared by ENVIRON International Corp, Novato, CA, for Coordinating Research Council, Atlanta, GA.

Radian, 1996, *Evaluating the Uncertainty of Emission Estimates, Final Report*, Prepared by Radian Corporation for the Emission Inventory Improvement Program, State and Territorial Air Pollution Control Officers' Association, Association of Local Air Pollution Control Officers, and U.S. Environmental Protection Agency, Research Triangle Park, NC.

Rubinstein, R. Y., 1981, *Simulation and the Monte Carlo Method*, John Wiley & Sons: New York.

Rabinovich, S., *Measurement Errors and Uncertainties: Theory and Practice*, Springer-Verlag, New York, 1999

Zheng, J., and H.C. Frey, 2001, "Quantitative Analysis of Variability and Uncertainty in Emission Estimation: An Illustration of Methods Using Mixture Distributions," *Proceedings, Annual Meeting of the Air & Waste Management Association*, Orlando, FL.

PART II

GENERAL METHODOLOGY OF QUANTIFICATION OF VARIABILITY AND UNCERTAINTY IN EMISSION ESTIMATION

Junyu Zheng

2.0 General Methodology

A general methodology for the quantification of both variability and uncertainty in emission factors, activity factor, and emission inventories is described in this part. The methods illustrated here are based upon the assumption that the data to be used to make variability and uncertainty analysis have been compiled and evaluated.

In practice, the two basic components of all kinds of emission inventories are the emission factor, which addresses the amount of pollutant from a given operation for specific source categories, and the activity factor, which quantifies the number of the operations for the specific source categories (Beck and Wilson, 1997). The product of the emission factor and activity factor produces an inventory of emissions from a certain population of sources. Emission factors are typically assumed to be representative of an average emission rate from a population of pollutant sources in a specific category (EPA,1995). However, there may be uncertainty in the population average emissions because of random sampling error, measurement errors, or possibly because the sample of the pollutant sources from which the emission factor was developed was not a representative sample. These first two factors typically lead to imprecision in the estimate of the population average, whereas the third factor may lead to possible biases or systematic errors in the estimated average. In order to avoid errors in inferences made based upon emission inventories, it is important to understand and account for the uncertainty in the inventory.

For different pollutant sources categories or pollution sources, though there exists some differences in calculating emissions or there are different characteristics for different sources categories; however, as discussed above, an emission inventory composes of two basic components: emission factor component and activity factor

component. Therefore it is possible to develop a general framework for producing a probabilistic emission inventory for any source categories.

This part will present a general approach to develop a probabilistic emission inventory, and introduce in detail the methods involved in the approach related to the development of a probabilistic emission inventory. These methods include, for example, database compilation and evaluation, statistical analysis of variability and uncertainty in emission factors and activity factors, the propagation of variability and uncertainty in emission inventory model inputs through emission inventory models, and the methods for calculation of the relative importance of input uncertainties with respect to uncertainty in the total emission inventory.

2.1 General Approach for Developing a Probabilistic Emission Inventory

An emission inventory could also be both variable and uncertain. Initially, probability distributions are developed for the emission factor data set and the activity factor data set. These probability distributions typically represent inter-plant variability for a specified averaging time.

There is uncertainty regarding the true value of each individual data point. Consequently, there is also uncertainty regarding the true value of the frequency distribution regarding variability among sources within the population. As a result, there is uncertainty in any estimate of any statistic of the population, such as the mean emission rate.

The variable and uncertain emission and activity factors are then propagated through the emission inventory model to simulate the uncertainty in the estimate for the total emissions from a population of emission sources. However, the true value of the emission and activity factors for each source are unknown. Hence, uncertainty in

emission and activity factors applied to individual sources is reflected by a distribution of uncertainty for the total emissions.

Based on the guidance of emission inventory development of EPA EIIP program (ERG, 1997), a general approach employed to develop a probabilistic emission inventory can be summarized as the following major steps:

1. Data preparation. It includes the assessment of data needs, data collection plans, and compilation or evaluation of existing databases for the specific sources categories.
2. Selection or development of emission inventory models.
3. Statistical analysis of variability in emission inventory model inputs. It includes visualization of data by developing empirical cumulative distribution functions for model inputs; fitting, evaluation, and selection of alternative parametric probability distribution models for representing variability in model inputs.
4. Characterization of uncertainty in the distributions for variability.
5. Propagation of uncertainty and variability in model inputs through emission inventory models to estimate uncertainty in category-specific emissions and/or total emissions from a population of emission sources.
6. Calculation of importance of variability and uncertainty.

The technologies and methodologies associated with the steps are described in the following sections.

2.2 Data Preparation

A starting point of developing a probabilistic emission inventory is to collect and handle data. In data collection, an important step is to assess data needs and to make data

collection plans. This involves in the assessment of the scope and objective of inventories, the evaluation of information contained in the existing emissions inventories and necessary emission calculations. For example, emissions are estimated for area and mobile sources by selecting representative subsets of individual sources from which emission calculation can be derived and then scaled up to reflect the population of these sources. Data handling will include the Quality Assurance (QA)/ Quality Control (QC) checking, data combining, data screening and data evaluating. QA is the management of the data to ensure that the data quality objectives are met, and ensures that adequate protocols are followed and independent testing of data. QC is the management of the collection and analysis of data to ensure they meet data quality objectives, and routinely check the calibration of laboratory equipment. Therefore, the QA/QC checking refers to if or not data collection and analysis of data strictly follows the protocols to ensure the data quality. Data combining refers to the situation in which data from different sources might need to be combined. The purpose of data screening is to eliminate the data that do not have enough information or the data are not needed for the specific emission inventory. Data evaluating assesses if or not the data can be representatives of emission factors and activity factors for specific source categories. These steps might be a little different for different source categories. The resulting data via data handling can be used to form a database, which will be used in the development of probabilistic emission inventories. An example of database compiling and evaluation in developing a statewide utility NO_x emission inventory are described in the Part VI.

2.3 Emission Inventory Models

As mentioned previously, emission estimates can be obtained by multiplying an emission factor with an activity factor that represents the extent of the emissions-

generating activity, therefore, a general emission inventory model, which can cover different emission sources, can be described as

$$E = A \times EF \quad (2-1)$$

Where,

E = emissions (e.g., lb of NO_x as NO₂)

A = activity factor (e.g., tons of coal burned), and

EF = emission factor (e.g., lb of NO_x as NO₂ per ton of coal burned).

In most cases, the emission factor and activity factor are often the outputs of other models describing emission factor or activity factors. For example, for a power plant unit, the activity factor is the product of the unit heat rate (BTU of fuel input required to produce one kWh of electricity), unit capacity factor (average capacity utilization for a given time), and unit capacity (MW). For highway vehicles emission sources, emission factor depends on mileage, deterioration rate, temperature, Reid vapor pressure and other parameters. A known model to describe the highway vehicle emission factors is MOBILE_x

There are some known models that EPA recommends to calculate emission factor or emission inventory for some emission source categories. For example, MOBILE_x estimates emission factors for on-road mobile vehicles. An example of utilizing MOBILE5b to quantify the variability and uncertainty in highway emission factors are shown in the Part V of this dissertation. NONROAD model is used to estimate the emissions from non-road engines or equipments; and LAEEM model is designed to estimate the air emissions from landfills for state and local regulatory agencies (EPA, 2002). Selection of emission inventory models depends on the specific emission source

categories. In general, if available, a known emission inventory model or an emission inventory model recommended by EPA for the specific source categories is encouraged to use in developing a probabilistic emission inventory.

2.4 Numerical sampling techniques

In order to analyze variability and uncertainty in model input or output, a probabilistic modeling environment is required. Various techniques have been proposed for performing variability and uncertainty analysis, for example, internal analysis, analytic method, numerical approximation approximations such as Taylor series expansion, and numerical sampling methods. Based on the limits, robustness, and mathematical complexity of these methods, numerical sampling techniques are the most widely used since it is the most robust and least restrictive with respect to model design and model input specification. Two numerical sampling techniques, Monte Carlo Sampling and Latin Hypercube Sampling are presented in the next subsequent sections.

2.4.1 Monte Carlo Sampling

Monte Carlo methods have been used for centuries, but only in the past several decades has the technique gained the status of a full-fledged numerical method capable of addressing the most complex applications. The name "Monte Carlo" was coined by Metropolis (inspired by Ulam's interest in poker), who applied these techniques to simulate the diffusion of neutrons during the Manhattan Project of World War II. Monte Carlo simulation is now used routinely in many diverse fields, from the simulation of complex physical phenomena such as radiation transport in the earth's atmosphere and the simulation of the esoteric sub-nuclear processes in high energy physics experiments, to the mundane.

In Monte Carlo method, a probability distribution is specified for each model input. Each distribution can be represented as a cumulative distribution function (CDF). Numerical methods based upon the use of a pseudo random number generator (PRNG) are used to generate random values from the assigned probability distribution model. There are several approaches for generating random numbers, for example, the method of inversion, the method of convolution, the method of the composition, and the acceptance –rejection method (Law and Kelton,1991). Of these methods, conceptually the most straight-forward is based upon the use of an inverse CDF. To generate a random number from a specified probability distribution model, first a pseudo random number is generated from a uniform distribution over the range of zero to 1. The random number is then used as an estimate of cumulative probability as an input to the inverse CDF for the assigned probability distribution model. The output of the inverse CDF is a numerical value of the random variable of interest. This process is repeated n times, where n is the desired simulation sample size, to produce many estimates of a model input. This process is also conducted simultaneously, with different random values from the PRNG, for each of many probabilistic inputs to a model. The random values for each model input are used to calculate a model output values. The propagation of multiple model inputs through a model leads to a distribution of model output values that reflect the uncertainty or variability in the model inputs (Ang and Tang,1984). More information regarding random Monte Carlo method may be found in Morgan and Henrion (1990), Cullen and Frey (1999) and others.

One advantage of using Monte Carlo sampling is that, with a sufficient sample size, it can provide an excellent approximation of the output distribution. Its primary

disadvantage is that it may be necessary to use large sample sizes to obtain smooth approximation of the CDF of a model output. Random Monte Carlo simulation is a desirable method when the objective is to simulate random sampling error. Other methods, such as Latin Hypercube Sampling, can produce smoother estimates of the empirical CDF of a model output with fewer samples than are needed using random Monte Carlo (Frey, et al., 1998b).

2.4.3 Latin Hypercube Sampling

As an alternative to random Monte Carlo sampling, Latin Hypercube Sampling (LHS), developed by McKay, Beckman, and Conover (1979), is a stratified sampling technique which can ensure that samples are taken from the entire range of a distribution. In LHS, the range of each input distribution is divided into n intervals of equal marginal probability. One value of the random variable is selected from each interval. The sample taken from each interval may be selected at random from within the interval, or from the median of the interval. The former is referred to as random LHS while the later is called median LHS. In both median and random LHS, the n values from each distribution are randomly paired with values generated for other model inputs. The stratification of the input distributions into n equal probability intervals ensures that samples are taken from the entire range of the distributions even with a relatively small sample size compared to random Monte Carlo sampling.

The disadvantage of LHS is that it is not a purely random sampling technique and, therefore, the results may not be subject to analysis by standard statistics (McKay *et al.*, 1979). Therefore, one cannot decide in advance how many samples are needed for a desired degree of convergence, as is possible for random Monte Carlo sampling.

However, for some applications, LHS is a more precise numerical simulation method than random Monte Carlo for a given simulation sample size.

2.5 Visualization of Datasets Using Empirical Distributions

Some of the key purposes of visualizing data sets include: (1) evaluation of the central tendency and dispersion of the data; (2) visual inspection of the shape of the empirical distribution of the data as a potential aid in selecting parametric probability distribution models to fit to the data; and (3) identification of possible anomalies in the data set (e.g., outliers). Specific techniques for evaluating and visualizing data include calculation of summary statistics, and plotting a data set as an empirical Cumulative Distribution Function (CDF).

Three key characteristics of a CDF are its central tendency, dispersion, and shape. There are several measures of central tendency, which include the mean, median, and mode. The dispersion, or the spread, of a distribution is measured by the standard deviation or the variance of the distribution. The relative standard deviation (RSD), also known as the coefficient of variation (CV), is the standard deviation divided by the mean. For a non-zero mean, the CV provides a normalized indication of the dispersion of data values, with a large CV indicating relatively large variability in the data set. The shape of the distribution is reflected by quantities such as skewness and kurtosis. The skewness is the asymmetric of a distribution, and the kurtosis refers to the peakedness of a distribution. These statistics can be used to aid in the selection of a parametric probability distribution model to fit to the data (Cullen and Frey, 1999).

A CDF is a relationship between “cumulative probability” and values of the random variable. Cumulative probability is the probability that the random variable has values less than or equal to a specific numerical value of the random variable. CDFs

provide a relationship between fractiles and quantiles. A fractile is the fraction of values that are less than or equal to a specific value of a random variable. Fractiles expressed on a percentage basis are referred to as percentiles. A quantile is the value of a random variable associated with a given fractile (Frey, Bharvirkar and Zheng, 1999). For example, the range of data values enclosed by the 0.025 and 0.975 fractiles (2.5 and 97.5 percentiles) is often of particular interest, since this provides an indication of the dispersion of a distribution as reflected by the 95 percent probability range of values. An example of a CDF is illustrated in Figure 2-1

Empirical estimation of a fractile from data requires rank ordering of the data. There are several possible methods for estimating the percentile of an empirically observed data point.

These methods are referred to as “plotting positions.” The plotting position is an estimate of the cumulative probability of a data point. As described by Cullen and Frey (1999), Harter (1984) provides an overview of the various types of plotting positions.

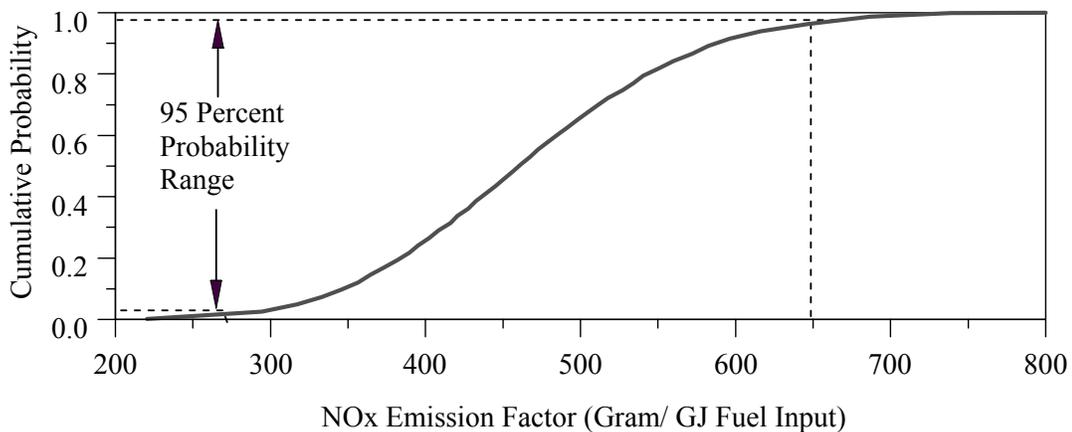


Figure 2-1. Plot Illustrating the 95 Percent Probability Range on a Cumulative Distribution Function.

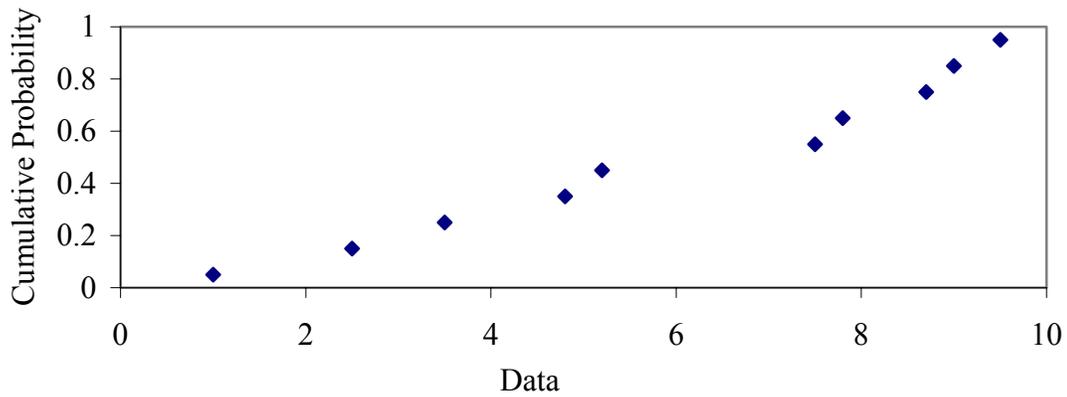


Figure 2-2. Example Graph of Visualizing Data Using the Hazen's Plotting Position Method (n=10)

$$F_x(x_i) = \Pr(X < x_i) = \frac{i - 0.5}{n}, \text{ for } i = 1, 2, \dots, n \text{ and } x_1 < x_2 < \dots < x_n \quad (2-2)$$

where,

i = Rank of the data point when the data set is arranged in an ascending order

n = number of data points

$x_1 < x_2 < \dots < x_n$ are data points in the rank-ordered data set

$\Pr(X < x_i)$ = Cumulative probability of obtaining a data point whose value is less than x_i

A commonly used plotting position, proposed by Hazen (1914), is used in AuvTool for displaying data points in comparison to fitted parametric distributions:

An example graph of visualizing data using the Hazen's plotting position method is shown in Figure 2-2. The figure depicts the plotting position of each of 10 data points for a small data set.

2.6 Definitions of Probability Distribution Models

Probability distribution models used in this study include the normal, lognormal, Weibull, gamma, beta, uniform, symmetric triangle parametric distributions and empirical distribution. Ang and Tang (1984), Hahn and Shapiro (1967), Morgan and Henrion (1990), Cullen and Frey (1999) and others review the theoretical basis underlying each of these distributions. The normal and lognormal distributions have an underlying theoretical basis in the central limit theorem (CLT) when applied to additive or multiplicative processes, respectively. For example, a process of pollutant dispersion generated by the sum of many random variations can be described by the Gaussian plume model (Seinfeld, 1986). Although the normal distribution is not appropriate for representing non-negative quantities because it has an infinite negative tail, it is often used to represent non-negative quantities, such as weight or length, so long as the coefficient of variation is less than about 0.2 (Morgan and Henrion, 1990).

The lognormal, gamma and Weibull distributions are useful for representing non-negative and positively-skewed data. The two-parameter beta distribution is bounded by zero and one, and has flexibility to represent data with a variety of central tendency and skewness. The uniform and symmetric triangle distributions are most commonly used to represent expert judgments made in the absence of data. Empirical distributions can be used instead of parametric distributions. A comparison of the use of empirical and parametric distribution is described in EPA (1999a) and in Section 2.7.

More discussion of distribution selection criteria can be found in Hahn and Shapiro (1967), Ang and Tang (1984), Morgan and Henrion (1990), Hattis and Burmaster (1994), and Alvarez (1996), and Cullen and Frey (1999), among others.

2.6.1 Definition of Parametric Probability Distributions

The definitions of the seven parametric distributions included in the accompanying software tool AuvTool developed in this study are presented in Table 2-1. The definitions are based upon the probability density function (PDF). In Table 2-1, for the normal distribution, μ is the arithmetic mean, and σ is the arithmetic standard deviation. For the lognormal distribution, $\mu_{\ln x}$ is the mean of the $\ln x$, and $\sigma_{\ln x}$ is the standard deviation of $\ln x$. For the beta distribution, α and β are shape parameters, and $B(\alpha, \beta)$ is the beta function. For the gamma distribution, α is the shape parameter, β is the scale parameter, and $\Gamma(\cdot)$ is the gamma function. For the Weibull distribution, k is the scale parameter, and c is the shape parameter. For the uniform distribution, a and b are the smallest and largest possible values. For the symmetric triangle distributions, a and b determine the range within which the variable can vary.

2.6.2 Empirical Distribution

An empirical distribution is defined as a discrete distribution, F , that gives equal probability, $1/n$, to each value x_i in the dataset, \mathbf{x} (Efron, 1979). The CDF for this function is therefore a step function of original data set, \mathbf{x} , where each value x_i is assigned a cumulative probability of i/n for $i = \{1, 2, \dots, n\}$. An example of an empirical distribution representing a step function is provided in Figure 2-3.

Table 2-1. The Definitions of Parametric Probability Distributions

Name of Distribution	Probability Density Function (PDF)	
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	
Lognormal	$f(x) = \frac{1}{x\sqrt{2\pi\sigma_{\ln x}^2}} e^{-\frac{(\ln x - \mu_{\ln x})^2}{2\sigma^2}}$	$(0 < x < \infty)$
Beta	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$	$(0 \leq x \leq 1)$
Gamma	$f(x) = \frac{\beta^{-\alpha} x^{\beta-1} e^{-x/\beta}}{\Gamma(\alpha)}$	$(0 \leq x < \infty)$
Weibull	$f(x) = \frac{c}{k} (x/k)^{c-1} \exp(-(x/k)^c)$	$(0 \leq x < \infty)$
Uniform	$f(x) = \frac{1}{b-a}$	$(a \leq x \leq b)$
Symmetric Triangle	$f(x) = \frac{b- x-a }{b^2}$	$(a-b \leq x \leq a+b)$

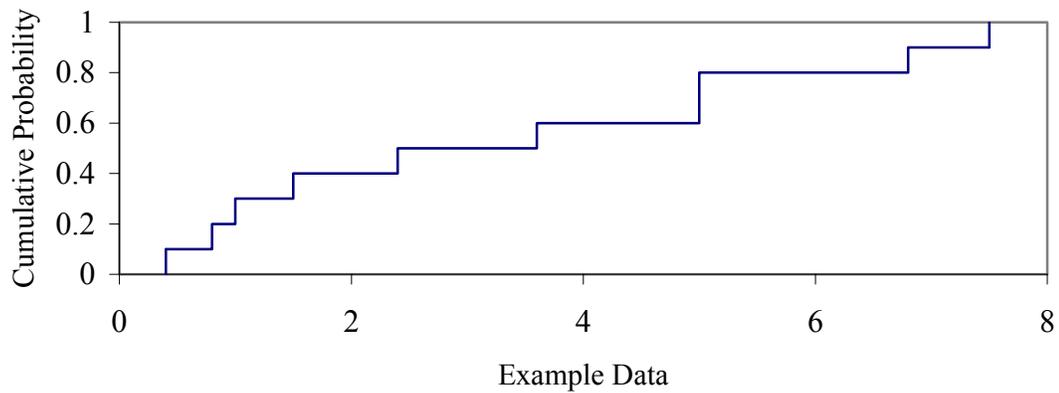


Figure 2-3. An example of an Empirical Distribution Represented a Step Function (n=10)

2.7 Parameter Estimation of Parameter Distributions

A probability distribution model is a description of the probabilities of all possible values in a sample space. A probability distribution model is typically represented as a

PDF or a CDF for a continuous random variable. The PDF for a continuous random variable indicates the range and relative likelihood of values. The CDF is obtained by integrating the PDF (Cullen and Frey, 1999).

Probability distribution models may be empirical, parametric, or combinations of both. A parametric probability distribution model is a model described by parameters. The power of using parametric probability distribution models is that data sets, which may contain large numbers of data points, can be described in a compact manner based on a particular type of parametric distribution function and the values of its parameters. For example, a normal distribution is fully specified if its mean and standard deviation are known. Another potential advantage of parametric probability distributions compared to empirical distributions is that it is possible to make predictions in the tails of the distribution beyond the range of observed data. In contrast, using conventional empirical distributions, the minimum and maximum values of the distribution are limited to their minimum and maximum values, respectively, of the data set. These values typically change as more data are collected. EPA (1999a) presents a discussion of the use of empirical versus parametric distributions.

Based upon visual inspection of an empirical distribution of data as described in Section 2.5, and consideration of processes that generated the data, the analyst can make a judgment regarding selection of one or more candidate parametric distributions to fit to the data set. Once a particular parametric distribution has been selected, a key step is to estimate the parameters of the distribution. The method of Maximum Likelihood Estimation (MLE) and the Method of Matching Moments (MoMM) are among the most typical techniques used for estimating the parameters.

In order to estimate values of the parameters of a parametric probability distribution, statistical estimation methods must be used. Using such estimation methods, inferences are made from an available data set regarding a single best estimate of the parameter values. Usually, there are alternative methods available to estimate parameter values. Thus, it is necessary to choose a parameter estimation method. Small (1990) has discussed the following six desirable characteristics of estimators. These characteristics are useful when comparing and selecting an estimation method:

Consistency: A consistent estimator converges to the “true” value of the parameter as the number of samples increases.

Lack of Bias: On average over many applications to many different data sets, an unbiased estimator yields an average value of the parameter estimate that is equal to that of the population value.

Efficiency: An efficient estimator has minimum variance in the sampling distribution of the estimate. A sampling distribution is a probability distribution for a statistic (e.g., mean, standard deviation, distribution parameters).

Sufficiency: An estimator that makes maximum use of information contained in a data set is said to be sufficient.

Robustness: A robust estimator is one that works well even if there are departures of the data from the underlying distribution. In other words, such as estimator will yield reasonable values of the parameters even if there are some anomalies in the data set.

Practicality: A practical estimator is one that satisfies the needs for the preceding five characteristics while remaining computationally efficient.

For small sample sizes, the MLE method does not always yield minimum variance or unbiased estimates (Holland and Fitz-Simmons, 1982). However, for larger sample sizes, the MLE method tends to better satisfy the first five criteria for statistical estimation than other methods. Compared to MLE, MoMM estimators tend to be more robust but less efficient. MLE can be extended to estimate parameters for distributions fitted to censored data. In the present study, both MLE and MoMM are included as options for estimation of parameters of parametric probability distributions. The MoMM and MLE methods are described in more detail in the next subsections.

2.7.1 Method of Matching Moments

The Method of Matching Moments (MoMM) is based upon matching the moments or central moments of a parametric distribution (e.g., mean, variance) to the moments or central moments of the data set. MoMM estimators are often easy to calculate. The method is usually the most straightforward to implement. Therefore it typically satisfies the criteria of practicality, for example, there are convenient solutions for MoMM parameter estimates for normal, lognormal, gamma, and beta distributions (Hahn and Shapiro, 1967). However, it may not fully satisfy the other criteria. The estimators for estimating parameters for the common parametric distributions using the method of matching moments can be found in the Appendix A.

2.7.2 Maximum Likelihood Estimation (MLE)

The MLE methods involves the selection of parameter values that characterize a distribution which was most likely to yield the observed data set (Cohen and Whitten, 1993). A likelihood function for independent samples is defined as the product of the

PDF evaluated at each of the sample values. For a continuous random variable, for which independent samples have been obtained, the likelihood function is:

$$L(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^n f(x_i | \theta_1, \theta_2, \dots, \theta_k) \quad (2-3)$$

where,

$\theta_1, \theta_2, \dots, \theta_k$ = Parameters of the parametric probability distribution model.

k = Number of parameters for the parametric probability distribution model.

x_i = Values of the random variable, for, $i = 1, 2, \dots, n$

n = Number of data points in the data set.

f = Probability density function.

The general idea behind MLE is to choose values of the parameters of the fitted distribution so that the likelihood that the observed data is a sample from the fitted distribution is maximized. The likelihood is calculated by evaluating the probability density function for each observed data point, conditioned upon assumed values for the parameters, and multiplying the results. The parameter values may be changed, such as by using an optimization method, to change the value of the likelihood function until a maximum is reached. More commonly, the log-transformed version of the likelihood function is used, which is based upon the sum of the natural log of the probability density function evaluated for each data point, conditioned upon assumed values or the parameters. The MLE parameter estimators can be obtained by varying the parameter values so as to find the maximum of the log-likelihood function.

The log-likelihood function of a univariate (describing one data set) two-

parameter distribution is given by:

$$L = \sum_{i=1}^n \ln[f(x_i | \theta_1, \theta_2)] \quad (2-4)$$

where,

n = number of data points.

L = Log-likelihood function

f = Probability density function

θ_1, θ_2 = parameters of a two-parameter distribution

For definitions of the probability density function $f(x | \theta_1, \theta_2)$ for different parametric distributions, see Table 2-1 in Section 2.6. For some parametric probability distributions, such as the normal and lognormal distributions, analytical solutions for the maximum likelihood estimators of the parameters are available by setting the first partial derivatives of the likelihood function equal to zero. However, in many cases, an analytical solution is not readily available. In these cases, the maximum likelihood parameter estimates can be found using numerical optimization techniques. For the uniform distribution, since the density function is a constant, no MLE solution is available. Except for the uniform distribution, the estimation of the maximum likelihood parameter values for the distributions in Table 2-1 can be formulated as the following optimization problem:

$$\text{Maximize} \quad L = \sum_{i=1}^n \ln[f(x_i | \theta_1, \theta_2)] \quad (2-5)$$

Subject to

$\theta_1 > 0$ for beta ($\theta_1 = \alpha$), gamma ($\theta_1 = \alpha$), Weibull ($\theta_1 = k$)

$\theta_2 > 0$ for beta ($\theta_2 = \beta$), gamma ($\theta_2 = \beta$), Weibull ($\theta_2 = c$)

where,

n= number of samples

The optimization problem here is a multidimensional constrained one. A variety of methods are available to solve such problems. These methods include the downhill simplex method; the direction-set method, of which Powell's method is the prototype; the penalty function method; and others (Press, *et al.*, 1992). In this study, Powell's method is employed. This method is relatively easy to program, it does not require calculations of derivatives, and it typically provides good results.

Optimization solutions for the MLE parameter estimates are used in the accompanying software AuvTool for the gamma, Weibull, beta, and symmetric triangle distributions. The MLE estimators for the normal and lognormal distributions are as follows (Morgan and Henrion, 1990):

MLE Parameter Estimators for the Normal Distribution

$$\mu = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2-6)$$

$$\sigma^2 = s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (2-7)$$

MLE Parameter Estimators for the Lognormal Distribution

$$\mu_{\ln x} = \frac{1}{n} \sum_{i=1}^n (\ln x_i) \quad (2-8)$$

$$\hat{\sigma}_{\ln x} = \left(\frac{1}{n} \sum_{i=1}^n (\ln x_i - \hat{\mu}_{\ln x})^2 \right)^{1/2} \quad (2-9)$$

2.8 Evaluation of Goodness-of-Fit of a Probability Distribution Model

There are many goodness-of-fit tests available from which to evaluate the goodness of fit of an assumed distribution model with respect to the data. Two general types of approaches for evaluating goodness of fit include probability plots and statistic tests (Cullen and Frey, 1999).

Probability plots are widely recognized to be a subjective method for determining whether or not data contradict an assumed model based upon visual inspection (Cullen and Frey, 1999). A graphical technique used in this study is to compare the CDF of the fitted distribution with the original data set plotted using the Hazen plotting position method (Hazen, 1914) that was introduced in Section 2.5.

Statistical goodness-of-fit tests provide a quantitative measure of the goodness-of-fit of the assumed probability distributions, but many only apply to parametric distributions. An empirical distribution is an exact representation of the data in which each data point is assigned a probability of $1/n$; therefore, a statistical goodness-of-fit test is not needed in this case. Three common goodness-of-fit tests for parametric distributions include the chi-square test, the Kolmogorov-Smirnov (K-S) test, and the Anderson-Darling (A-D) test. However, these tests may only be employed if a minimum amount of data is available (Cullen and Frey, 1999). For example, for the chi-square test, at least 25 data points should be available. The K-S test can be used with as few as five data points. The A-D test is valid if the number of samples is greater than or equal to eight.

The chi-square test involves calculating a test statistic that approximately follows a chi-square distribution only if the hypothesized model cannot be rejected as a poor fit to the data. The advantage of chi-square test is its flexibility; it can be used to test any

distribution. However, a disadvantage of this method is that it has lower power than other statistical tests (Cullen and Frey, 1999). This is because the chi-square test involves binning of the data. In binning the data, some of the information associated with individual data points is lost. Thus, the chi-square test is less discriminatory than a test that makes more sufficient use of all data points, such as the K-S test.

The K-S test involves a comparison between a stepwise empirical CDF and the CDF of a hypothesized distribution. This test is based upon evaluation of the maximum difference in the cumulative probability of the fitted distribution versus that of a data point. An attractive feature of K-S test is that it is a distribution-free test of goodness of fit. An advantage of K-S test over the chi-square test is that it can be used with smaller sample sizes. However, K-S test tends to be more sensitive to deviations of a good fit near the center of the distribution compared to at the tails (Stephens, 1974; D'Agostino and Stephens, 1986).

The A-D test is a “quadratic” test that is based upon a weighted square of the vertical distance between the empirical and fitted distributions (Cullen and Frey, 1999). The A-D test gives more weight to the tails than does the K-S test and therefore is more sensitive to deviations in the fit at the tails of a distribution (Stephens, 1974). However, the A-D test is not distribution-free test. Therefore, the critical values must be calculated specifically for each type of parametric distribution. Therefore, the A-D test is often used as a supplement to other goodness-of-fit tests.

Because the chi-square test requires at least 25 data points, and because it is not as powerful as other methods, the chi-square test was not included in this study and not implemented in the accompanying software tool AuvTool.

It must be pointed out that there are some limitations with the use of statistical goodness-of-fit tests. For example, they address only one possible criterion for determining goodness-of-fit, and could imply acceptance of a fit that might be poor for reasons not addressed by the criterion, or imply rejection of a fit that might be acceptable for reasons not addressed by the criterion. For example, it is possible that a normal distribution might not be rejected by a goodness-of-fit test. However, if the normal distribution is used to represent a quantity that must be non-negative, and if the probability of predicting negative values using a normal distribution is not negligible, then the use of a normal distribution will not make physical sense. Therefore, an uncritical application of a goodness-of-fit test can lead to an inappropriate choice of parametric distribution. Conversely, the goodness-of-fit test may imply rejection of a non-negative distribution, such as a lognormal, which might be theoretically consistent with the basis of the data.

The graphical comparison of the CDF of the fitted distribution to the original data set plotted using the Hazen plotting position is more informative when confidence intervals are estimated for the fitted CDF, and when the frequency with which data are enclosed by the confidence intervals is taken into account. This approach is discussed in more detail in Section 2.10 on bootstrap simulation.

In the following subsections, methods for evaluating the adequacy of the fit of a parametric distribution with respect to the data are explained in more detail. These include: (1) visually comparing the CDF of the fitted distribution with the data; (2) using the K-S test; (3) using the A-D test; (4) and visually comparing confidence intervals for the CDF of the fitted distribution with the data.

2.8.1 Graphical Comparison of CDF of Fitted Distribution to the Data

The goodness-of-fit of a parametric distribution compared to the data can be visually inspected. This is accomplished by plotting the CDF of the fitted distribution versus the data. The data can be plotted using the Hazen plotting position introduced in Section 2.5.

Since analytical solutions are not readily available for CDFs for all of the parametric distributions used in this study, the CDFs are estimated using numerical simulation. The construction of a numerically stable representation of CDF of the fitted distribution is based on statistical theory. The CDF is estimated by generating a large number of random samples from the parametric distribution and plotting them using the Hazen plotting position. With a large number of samples, the numerically simulated CDF will look as if it is a continuous smooth curve. The sample size chosen for numerical simulation of the CDF for purposes of graphical display is based upon the statement in Casella and Berger (1990) that if the sample size is large enough (e.g., $\geq 2,000$), then the sample can be assumed to be a very good representation of population distribution. Therefore, in the accompanying software tool AuvTool, 2,000 random numbers are generated for the distribution and are used to construct an empirical CDF using the Hazen plotting position. The numerically simulated CDF is considered to be a very good representation of the actual CDF of the fitted distribution, and it is plotted in the same graph with the original data set.

An example of a graphical comparison of a numerically simulated CDF for a parametric probability distribution and of the data to which the distribution was fit is shown in Figure 2-4. The data are depicted by open circles. The numerically simulated CDF is depicted by a solid line. The example shown in Figure 2-4 is for a beta.

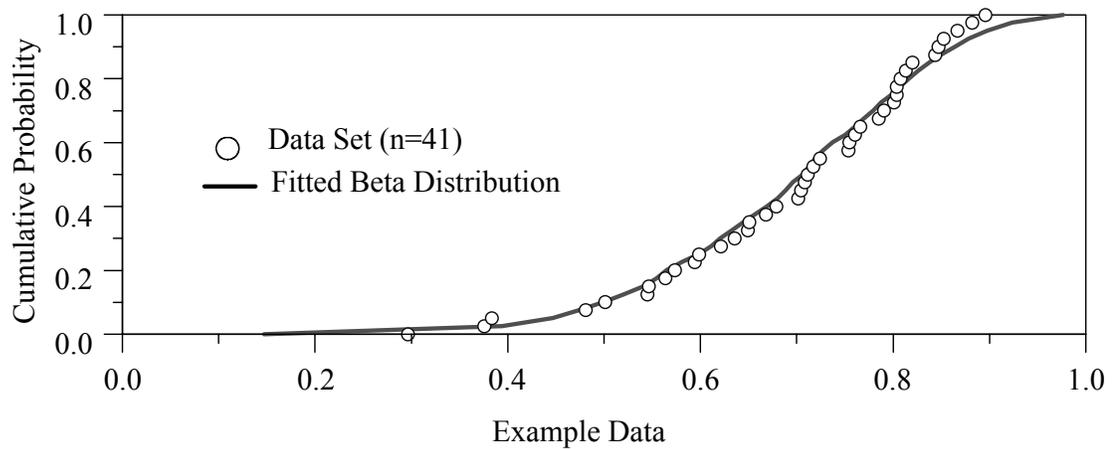


Figure 2-4. Comparison of Fitted Beta Distribution to an Example Dataset

distribution fit to a data set for a quantity that is bounded by zero and one. The beta distribution corresponds very closely with the data over most of the range of the observed values

2.8.2 Kolmogorov-Smirnov Test

As previously noted, the K-S test is based on comparison of the CDF of the fitted distribution to an empirical CDF of the data. The maximum discrepancy in the estimated cumulative probabilities for the two CDFs is identified. The maximum discrepancy is then compared to a critical value of the test statistic. If the maximum discrepancy is larger than the critical value, the hypothesized distribution is rejected (Cullen and Frey, 1999). This method is also discussed by Ang and Tang (1984), D'Agostino and Stephens (1986), and others.

The algorithm for performing the K-S test is described here:

- (1) Rank the original data in an ascending order to have an ordered dataset \mathbf{X} in which $X_k < X_{k+1}$, where, $k = 1, 2, \dots, n$.
- (2) Develop a stepwise cumulative density function as follows:

$$S_n(x) = \begin{cases} 0 & x < x_1 \\ k/n & x_k \leq x \leq x_{k+1} \\ 1 & x \geq x_n \end{cases} \quad (2-10)$$

where,

$S_n(x)$ = The stepwise cumulative density function

n = The number of data points in a data set

x_k = The data

(3) Calculate the maximum difference between $S_n(x)$ and the CDF of the fitted distribution over the entire range of X . The maximum difference is denoted by:

$$D_n = \max|F(x) - S_n(x)| \quad (2-11)$$

where,

D_n = The maximum difference

$F(x)$ = The CDF of the fitted distribution

(4) Compares the calculated maximum difference from Equation (2-11) with the critical value D_n^α at a significance level of α .

The often-used significance level is 0.05. The critical values of D_n^α at a significance level of $\alpha=0.05$ are tabulated in the Table 2-2.

Table 2-2 lists two kinds of critical values at a significant level of $\alpha=0.05$. One is marked as “Specified”, another is marked as “Unspecified”. “Specified” implies that the underlying distribution type representing a data set is known, while “Unspecified” means

Table 2-2. Critical Value of D_n^α the Kolmogorov-Smirnov Test

n	$\alpha=0.05$ (Specified)	n	$\alpha=0.05$ (Unspecified)
5	0.56	5	0.337
10	0.41	8	0.285
15	0.34	10	0.258
20	0.29	12	0.242
25	0.27	15	0.220
30	0.24	16	0.213
35	0.23	18	0.200
40	0.21	20	0.190
45	0.20	25	0.180
50	0.19	30	0.161
>50	$1.36/\sqrt{n}$	>30	$0.886/\sqrt{n}$

(Massey, 1951; Lilliefors, 1967)

that the information involving the underlying distribution for a data set is unknown. For example, if there is a sample for which the true values of the parameters of the population distribution are known, a “specified” critical value would be used. However, in most cases, the parameters of the distribution are estimated from the same data set for which the goodness-of-fit comparison is made. In this latter situation, the "Unspecified" values should be used. Since this latter case is more common, the “Unspecified” critical values are used in the development of AuvTool. If the critical value of a number n is not listed in the Table 2-2, and when n is less than 30 (“Unspecified”), a linear interpolation is used to calculate the critical value for the number.

The K-S test is a distribution-free; it can be applied to normal, lognormal, beta, gamma, Weibull, uniform, and symmetric triangle distributions. However, the K-S test has several important limitations: (1) it is only valid for continuous distributions; and (2) it tends to be more sensitive near the center of the distribution than at the tails (D’Agostino and Stephens, 1986).

2.8.3 Anderson-Darling Test

The A-D test is used to test if a sample of data is from a population with a specific distribution (Stephens, 1974). It is a modification of the K-S test and gives more weight to the tails than does the K-S test. Unlike K-S test, the A-D test is not a distribution-free test. For different distributions, A-D test statistics and the corresponding critical values are different. For some distributions, relevant information for calculating the A-D test is not available in literature. These distributions include uniform, symmetric triangle and beta distributions. Therefore, the A-D test is only considered for the normal, lognormal, gamma and Weibull distributions.

The A-D test statistic is defined as:

$$A^2 = -n - S \quad (2-12)$$

where,

$$S = \sum_{i=1}^n \frac{(2i-1)}{n} [\ln(F(x_i)) + \ln(1.0 - F(x_{n+1-i}))] \quad (2-13)$$

F is the cumulative distribution function of the specified distribution. X_i is the *ordered* data (Stephens, 1974; D'Agostino and Stephens, 1986).

When parameters of an assumed distribution are not known, and have to be estimated from the sample data, the A-D test statistic must be modified (D'Agostino and Stephens, 1986). For normal and lognormal distribution, the modified statistic is (D'Agostino and Stephens, 1986):

$$A^* = A^2(1.0 + 0.75/n + 2.25/n^2) \quad (2-14)$$

Table 2-3. The Critical Values for Anderson-Darling test for Normal, Lognormal and Weibull distributions

Distribution	$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.01$
Normal, Lognormal	0.631	0.752	0.873	1.035
Weibull	0.637	0.757	0.877	1.038

(D'Agostino and Stephens, 1986, Table 4.7, p=123; Table 4.17, p=146)

Table 2-4. The Critical Values for Anderson-Darling test for the Gamma Distribution

Shape Parameter	Significant Level $\alpha =0.05$
1	0.786
2	0.768
3	0.762
4	0.759
5	0.758
6	0.757
8	0.755
10	0.754
12	0.754
15	0.754
20	0.753
>20	0.752

(D'Agostino and Stephens, 1986, Table 4.21, p=155)

For the Weibull distribution, the modified statistic is (D'Agostino and Stephens, 1986):

$$A^* = A^2(1.0 + 0.2/\sqrt{n}) \quad (2-15)$$

For the gamma distribution, when both the scale and shape parameters are unknown and are estimated from the data, the A-D test statistic does not need to be modified (D'Agostino and Stephens, 1986). However, the critical value at a given significance level for the gamma distribution is dependent on the magnitude of its shape parameter (D'Agostino, Stephens, 1986).

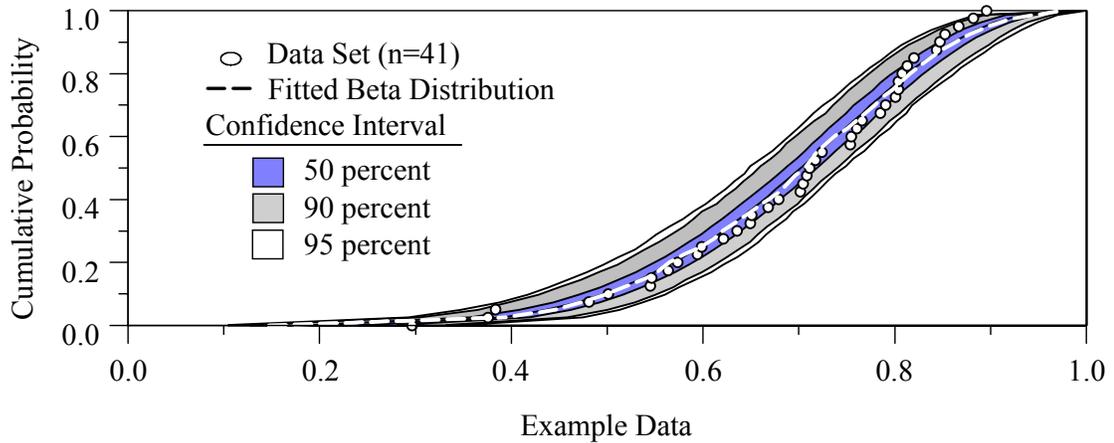


Figure 2-5. An Illustrative Example of Graphical Comparison of Confidence Intervals for CDF of Fitted Distribution to the Data

The critical values of the A-D test for the normal, lognormal, and Weibull distributions are given in Table 2-3 and for the gamma distribution are given in Table 2-4.

The linear interpolation is used to calculate the critical value of the A-D test for any given shape parameter based on the values provided in Table 2-4 for gamma distribution

2.8.4 Graphical Comparison of Confidence Intervals for CDF of Fitted Distribution to the Data

The results from bootstrap simulation can be used to help evaluate the goodness of a fit of a distribution with respect to the original data by graphically comparing confidence intervals for CDF of the fitted distribution to the data. More details on the bootstrap simulation and how the confidence intervals for CDF of the fitted distribution are estimated can be found in Sections 2.10.1, 2.10.3, and 2.10.4.

Figure 2-5 graphically shows a comparison of confidence intervals for the fitted distribution with an example data set. The results are from two-dimensional simulation with the points out of 41 are contained within the 95 percent confidence intervals. Thus,

the fit in this case is a reasonably good one. On average, it is expected that 95 percent of the data will fall inside of a 95 percent confidence interval of CDF of a fitted distribution if the data are a random sample from the assumed population distribution

2.8.5 Summary of Methods for Evaluating Goodness-of-Fit

Several different techniques for evaluating goodness-of-fit of a parametric probability distribution model compared to a data set have been presented. Although it is tempting to base the selection of a parametric probability distribution model solely upon the application of a goodness-of-fit statistical test, this temptation should be strongly resisted. Instead, it is critically important to consider the following questions in making the choice of a parametric distribution:

Is the selected parametric probability distribution model consistent with the data in terms of underlying theory?

Is the selected parametric probability distribution a plausible representation of the data? For example, if the data must be non-negative, does the selected distribution also have this feature?

What characteristics of the distribution are of most concern in your specific assessment, and are these criteria the same as those for the goodness-of-fit test? If so, then the goodness-of-fit test should be treated as a useful consideration in choosing a distribution, but it should not be the only consideration. The latter is especially true if the answers to either of the first two questions are "no".

Are the criteria for the goodness-of-fit test relatively unimportant for a particular assessment? In this case, the user will find it more useful to rely upon a graphical comparison of the fitted distribution with the data, either based upon a comparison of the CDF of the fitted distribution with the data, or based upon a comparison of the confidence intervals of the CDF of the fitted distribution with the data

In fact, both graphical comparison and statistical goodness-of-fit tests involve subjective judgment regarding what constitutes an acceptable fit (Cullen and Frey, 1999).

For example, the K-S and A-D tests involve subjective judgment regarding the choice of

significance levels. Many authors emphasize the subjective nature of statistical tests.

Hann and Shapiro (1967) state this quite well in their excellent book:

“One might conclude.... that a proper procedure for selecting a distribution is to consider a wide variety of possible models, evaluate each by the methods here described, and assume as correct the one that provide the best fit to the data. However, no such approach is being suggested. Where possible, **the selection of the model should be based on an understanding of the underlying physical properties**... The distributional test then provides a useful mechanism for evaluating the adequacy of the physical interpretation. Only as a last resort is the reserves procedure warranted, and then, only with much care, for, although many models might appear appropriate within the range of data, they might well be in error in the range for which predictions are desired,” (pp. 260-261).

2.9 Algorithms for Generating Random Samples from Probability Distributions

Computing efficiency and programming simplicity were used as the criteria for selecting methods for generating random samples from various distributions using Monte Carlo sampling. Monte Carlo simulation methods are based upon the use of a pseudo random number generator (PRNG) that produces a stream of random, independent uniformly distribution numbers. Uniformly distributed random numbers are used as the input to algorithms that generate random numbers from other types of distributions.

The most efficient and simple method for generating random variables from a particular type of probability distribution is the method of inversion (Frey and Rhodes, 1999). This method is always used when the CDF can be inverted. In many cases, however, the inverse CDF cannot be written in a closed form, and an alternative method is used. Some alternative methods are the method of composition, the method of convolution, and the acceptance-rejection method (Law and Kelton, 1991).

In the following subsections, the PRNG and the algorithms used in the accompanying AuvTool developed in this study to generate random variables for step-wise empirical distributions are described. The algorithms for generating random

samples based on the normal, lognormal, Weibull, gamma, beta, uniform, symmetric triangle distributions are presented in the Appendix B.

2.9.1 Pseudo Random Number Generator

The term pseudo-random refers to numbers which appear as if they are uniformly distributed random numbers that actually are generated in a completely deterministic manner (Barry, 1996). Pseudo random numbers are thought to be “good” when they have the following features (Rubinstein, 1981): (1) statistical uniformity, (2) statistical independence, (3) reproducibility, and (4) they can be generated quickly and economically. Another key consideration is the period length, which is the number of random values that are generated before the same sequence begins to be repeated.

There are a variety of methods for generating pseudo-random numbers (Bratley, *et al.*, 1987). The most widely used method is the Linear Congruential Generator (LGC). The advantage of LGC is its speed, simplicity and portability (Barry, 1996). However, a potential problem with a LGC approach is that its period length is easily exhausted (L’Ecuyer, 1996). It is well recognized that, for statistical reasons, the period length of a linear-type generator should be several orders of magnitude of larger than what is actually needed (L’Ecuyer, 1994; 1996).

An approach for increasing the period and improving the structure of the generator is to use combined Multiple Recursive Generators (MRGs) presented by L’Ecuyer (1996). In this method, two or more MRGs are combined. A combined generator with two MRGs is used in this study and is described as:

$$Z_n = (X_n - Y_n) \text{ mod } m_1 \quad (2-16)$$

where the two underlying generators X_n and Y_n are:

$$X_n = (a_1 X_{n-1} + a_2 X_{n-2} + a_3 X_{n-3}) \text{ mod } m_1 \quad (2-17)$$

and

$$Y_n = (b_1 Y_{n-1} + b_2 Y_{n-2} + b_3 Y_{n-3}) \bmod m_2 \quad (2-18)$$

with coefficients

$$a_1 = 0, a_2 = 63308, a_3 = -183326,$$

$$b_1 = 86098, b_2 = 0, b_3 = -539608,$$

and

$$m_1 = 2^{31} - 1 = 2147483647 \text{ and } m_2 = 2145483479.$$

The operator “mod” in Equations (2-16), (2-17) and (2-18) divides two integers and returns the remainder of the division. The period of this PRNG is 2^{205} ; the six initial values for x_0, x_1, x_2 and y_0, y_1, y_2 can be any integers from 1 to $2^{31} - 1 = 2147483647$ (L’Ecuyer,1996).

2.9.2 Empirical Distribution

In an empirical distribution, a data set is described by a step-wise empirical cumulative distribution function, in which the probability of sampling any discrete value within the dataset is $1/n$. A random re-sampled version of the original data set, of size n , is denoted by:

$$\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_n^*) \quad (2-19)$$

The asterisks indicate that \mathbf{X}^* is not actual data set \mathbf{x} , but rather a randomized or resampled version. Since the sampling is done with replacement, it is possible to have repeated values within any given random samples from an empirical distribution.

The algorithm for generating a random sample from an empirical distribution is as follows:

Step 1: Rank an original data set in an ascending order to have an ordered dataset

$$\mathbf{X}^0 \text{ in which } X_m^0 < X_{m+1}^0, \text{ where, } m = 1, 2, \dots, n.$$

Step 2: Generate a random number U from an $U(0,1)$ distribution.

Step 3: Calculate an index using the following formula:

$$i = n \times U \quad (2-20)$$

where,

i is a returned smallest integer that is larger than or equal to $n \times U$ between 1 and n by rounding up the product of $n \times U$

Step 4: Retrieve the data, X_i^o , located at the i^{th} of the ordered dataset \mathbf{X}^o .

2.10 Characterization of Uncertainty in the Distribution for Variability

The primary objective of this section is to introduce relevant methods for characterization of uncertainty in the mean, standard deviation, and parameters of a distribution. Uncertainty in a statistic attributable to random sampling error can be represented by a sampling distribution (Cullen and Frey, 1999). Sampling distributions are used to estimate confidence intervals for the parameters of a distribution. A confidence interval for a statistic is a measure of the lack of knowledge regarding the value of the statistic. There are a variety of methods for characterizing uncertainty in the mean or standard deviation, including analytical solutions and numerical simulations. Analytical solutions are available for cases in which the underlying distribution for a data set is normal or for which the variance is small enough and/or the sample size for a data set is large enough (e.g., >30). If the underlying population distribution is not normal and the sample size for a data set is small, analytical methods based upon normality may lead to significant errors in the estimation of confidence intervals. Therefore, there is a need

for a more flexible approach for estimating sampling distributions and confidence intervals. The numerical simulation method of bootstrap simulation, may be used to estimate confidence intervals for the mean or other statistics (Efron and Tibshirani, 1993).

Bootstrap simulation, introduced by Efron in 1979, is a numerical technique originally developed for the purpose of estimating confidence intervals for statistics based upon random sampling error. This method has an advantage over analytical methods in that it can provide solutions for confidence intervals in situations where exact analytical solutions may be unavailable and in which approximate analytical solutions are inadequate. For example, in estimating uncertainty in the sample mean, bootstrap simulation does not require that the original data set be normally distributed, even for small sample sizes. This advantage over analytical methods that are based on normality assumptions makes bootstrap simulation a more versatile and robust method for estimating uncertainty in a statistic due to sampling error, especially for non-normal data sets (Cullen and Frey, 1999). Bootstrap simulation has been widely used in the prediction of confidence intervals for a variety of statistics.

The method illustrated by Frey and Rhodes (1996;1998) for using bootstrap simulation in the context of an environmental case study is the basis for the simulation technique used in this study. The following subsections introduce the bootstrap method and the two major steps associated with the use of bootstrap method: (1) generating bootstrap samples; and (2) forming bootstrap confidence intervals. In addition, the details of the two-dimensional simulation method presented by Frey and Rhodes (1996; 1998) are described.

2.10.1 Bootstrap Method

The bootstrap method addresses uncertainty due to random sampling error by first assuming that the original data set, \mathbf{x} , of sample size n , is a random sample from the distribution \hat{F} , and then repeatedly asking the question: What if the data set had been a different set of n random values from the same distribution \hat{F} ? This question is answered by repeatedly generating “bootstrap samples.” A bootstrap sample, \mathbf{x}^* , is defined as a random sample of size n taken from the distribution, \hat{F} . Bootstrap samples may be simulated using random Monte Carlo simulation (Rhodes, 1997). A large number, B , of independent bootstrap samples ($\mathbf{x}^{*1}, \mathbf{x}^{*2}, \dots, \mathbf{x}^{*n}$) are selected from the distribution \hat{F} . From each of the B bootstrap samples, a new statistic $\hat{\theta}^*$, is computed such that:

$$\hat{\theta}^{*i} = f(\mathbf{x}^{*i}) \text{ for } i=1, 2, \dots, B \quad (2-21)$$

Each $\hat{\theta}^*$ is referred to as a *bootstrap replicate* of $\hat{\theta}$ (Rhodes, 1997; Frey and Rhodes, 1999).

The bootstrap replications ($\hat{\theta}^{*1}, \hat{\theta}^{*2}, \dots, \hat{\theta}^{*B}$) are each independent realizations of an estimate of the parameter θ . The dispersion of values of the bootstrap replications reflects the uncertainty in the sample estimate of the unknown parameter, θ , attributable to random sampling error. The bootstrap replicate values describe an estimate of the sampling distribution of the statistic. Since a statistic is estimated from randomly drawn values, it is itself a random variable. The number of bootstrap replications necessary to reasonably approximate the true sampling distribution of the statistic depends upon the statistic being estimated. For, example, according to Efron and Tibshirani (1993), to compute the standard error of the mean (the original intent of the bootstrap technique), B

= 200 is generally enough and $B = 25$ is often sufficient. However, for computing confidence intervals or estimating percentiles of sampling distributions, Efron and Tibshirani (1993) suggest $B = 1000$. In examples for computing confidence intervals given in Efron and Tibshirani (1993), the number of bootstrap replications ranges between $B = 1,000$ and $B = 2,000$.

2.10.2 Methods of Generating Bootstrap Samples

In bootstrap simulation, the sample data points, $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ are assumed to be a random sample of size n from some unknown probability distribution F . The parameter of interest, θ , is a characteristic of the distribution of F , $\theta = f(F)$, such as the mean, variance, shape or scale parameter, or any fractile or quantile of the distribution F . An estimate of θ is the statistic $\hat{\theta}$, which is determined from the data set, $\hat{\theta} = f(\mathbf{x})$.

Using the data set, \mathbf{x} , the distribution \hat{F} , is defined to be an estimate of the unknown population distribution F . The distribution may be defined as either an empirical distribution or a parametric distribution. The former is the basis for non-parametric bootstrap, and the latter is the basis for parametric bootstrap (Efron and Tibshirani, 1993). Non-parametric bootstrap is also commonly referred to as "resampling." One of the main shortcomings of resampling of a data set is that the minimum and maximum values obtained in each bootstrap sample are limited to the minimum and maximum values within the data set. When only small data sets are available, this can lead to biases in the representation of a given model input (e.g., failure to consider possible large values that are not present in the limited data set). The use of parametric distributions is one way to allow for the possibility that smaller or higher values than those observed in the data set may occur in the real system being modeled.

The method of generating bootstrap samples based on an empirical distribution for non-parametric bootstrap simulation is discussed in Section 2.9.2. The algorithms for generating bootstrap samples based on parametric distributions for normal, lognormal, beta, gamma, Weibull, uniform, and symmetric triangle distributions are documented in Appendix B.

2.10.3 Methods of Forming Bootstrap Confidence Intervals

The development of good confidence intervals is an important issue in bootstrap simulation. “Good” means that the bootstrap intervals should closely match exact confidence intervals in those special situations where statistical theory yields an exact answer, and the interval should give dependably accurate coverage probabilities in all situations. A method that produces such a good confidence intervals should be both transformation respecting and second-order accurate (Efron and Tibshirani, 1993).

Several bootstrap confidence interval methods have been proposed in the literature (Efron and Tibshirani, 1993; Burr, 1994). These methods include the standard normal, percentile, bootstrap-t, and Efron’s BC_a . The standard normal method requires the imposition of normality assumption on the bootstrap distribution and it is neither transformation respecting nor second-order accurate. Therefore, the standard normal method is not a “good technique” for forming bootstrap confidence interval. The percentile method is possibly the most frequently used in practice. Although it is only first-order accurate, the intervals obtained from this method are the simplest to use and explain (Efron and Tibshirani, 1993). The bootstrap-t and the BC_a intervals are comparable in that both have been demonstrated theoretically to be “second-order correct”, but the bootstrap –t method is not transformation respecting. Burr (1994) suggests that bootstrap-t is unstable. More discussion on these methods can be found in

the Efron and Tibshirani (1993), Burr (1994), and Martin (1990). Though there is no gold standard by which we can make a definitive conclusion which method is the best confidence interval one, but Efron (1993) suggests that the BC_a is recommended for general use, especially for nonparametric problems since it is both transformation respecting and second-order accurate. Therefore, in this dissertation, for a better bootstrap confidence interval, BC_a method will be discussed and be introduced in this study; for simplicity and because it is the most widely used method in practice, the percentile method will also be presented.

2.10.3.1 Percentile Method

Percentile method is the one which forms bootstrap confidence interval based on the percentiles of the bootstrap distribution of a statistic. Percentile method works as the following: Suppose we want to calculate a confidence interval that has a $(1-2\alpha)$ probability of enclosing the true value of a statistic, θ . The upper and lower bounds of this confidence interval are determined by ordering the B bootstrap replicates of $\hat{\theta}^*$, $(\hat{\theta}^{*1}, \hat{\theta}^{*2}, \dots, \hat{\theta}^{*B})$. Given these ordered statistics, the 100α th percentile (the lower bound of the confidence interval) is the $B \cdot \alpha$ th largest value, $\hat{\theta}^{*B \cdot \alpha}$, and the $100(1-\alpha)$ th largest value, $\hat{\theta}^{*B \cdot (1-\alpha)}$. For example, for $B = 1,000$ and $\alpha = 0.05$, the 90 % confidence interval for some parameter, θ , is given by:

$$[\hat{\theta}^{*B \cdot \alpha}, \hat{\theta}^{*B \cdot (1-\alpha)}] = [\hat{\theta}^{*50}, \hat{\theta}^{*950}] \quad (2-22)$$

where, $\hat{\theta}^{*50}$ and $\hat{\theta}^{*950}$ are simply the 50th and 950th values in the ordered set if the bootstrap statistics.

2.10.3.2 BC_a Method

BC_a method is an improved version of the percentile method. It stands for *bias-corrected and accelerated*. The BC_a intervals are substantial improvement over the percentile method in theory and practice, they come close to the criteria of goodness given above, though their coverage accuracy can still be erratic for small sample size (Efron and Tibshirani,1993).

The BC_a interval endpoints are also given by percentiles of the bootstrap distribution, but the percentiles used depend on the *acceleration* \hat{a} and *bias-correction* z_0 values. The BC_a interval of intended coverage $1-2\alpha$, is given by

$$BC_a: (\hat{\theta}_{lo}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}) \quad (2-23)$$

where,

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right) \quad (2-24)$$

$$\alpha_2 = \Phi\left(z_0 + \frac{z_0 + z^{(1-\alpha)}}{1 - a(z_0 + z^{(1-\alpha)})}\right) \quad (2-25)$$

Here $\Phi(\cdot)$ is the standard normal cumulative distribution function and $z^{(\alpha)}$ is the $100\alpha^{\text{th}}$ percentile point of a standard normal distribution. z_0 is the value of bias-correction, and \hat{a} is acceleration constant. When z_0 and \hat{a} equal zero, then BC_a method is the same as the percentile interval method (Efron and Tibshirani,1993).

The z_0 and \hat{a} can be computed by the following formulas:

$$z_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}\right) \quad (2-26)$$

Here $\Phi^{-1}(\cdot)$ is the inverse function of a standard normal cumulative distribution function, B , the number of bootstrap replication.

The \hat{a} is called the acceleration because it refers to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ . It is calculated as the followings (Efron and Tibshirani,1993):

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^3}{6 \left\{ \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right\}^{2/3}} \quad (2-27)$$

where,

$\hat{\theta}$ = The jackknife value of a statistic $s(\mathbf{x})$.

$\hat{\theta}_{(i)}$ = $s(\mathbf{x}_{(i)})$.

$\hat{\theta}_{(.)}$ = $\sum_{i=1}^n \hat{\theta}_{(i)} / n$.

Jackknife is a technique for estimating the bias and standard error of an estimate.

In jackknife, it focuses on the samples that leave out one observation at a time. For example, if we have a sample of $\mathbf{x}=(x_1, x_2, \dots, x_n)$ and an estimator of a statistic $\hat{\theta} = s(\mathbf{x})$,

the i^{th} jackknife sample consists of the data set with the i^{th} observation removed, that is,

$\mathbf{x}_{(i)}=(x_1, x, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$; $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$ will be the i^{th} jackknife replication of $\hat{\theta}$. In

uncertainty analysis, the statistic $\hat{\theta}$ can be mean, standard deviation, or parameters of a probability distribution in the parametric bootstrap simulation (Efron and

Tibshirani,1993). For different statistics, the estimation of *jackknife replication* of $\hat{\theta}$ will

be different. Relevant estimation techniques for the jackknife estimators of the

parameters of parametric distributions can be found in the section 2.7.

BC_a has two important theoretical advantages. First of all, it is transformation respecting. The second, the BC_a method concerns its accuracy and the BC_a interval can be shown to be second-order accurate. The main disadvantage of the BC_a method is the large number of bootstrap replication required; therefore computation load is heavy. For example, in using BC_a method to form a confidence interval for a parameter of a probability distribution model, additional jackknife replications of parameter estimator with the same number of bootstrap replication are required; which will, therefore, lead to the significant increase of computation load.

2.10.4 Two-Dimensional Simulation of Variability and Uncertainty

The two-dimensional approach for simultaneously simulating both variability and uncertainty, presented by Frey and Rhodes (1996, 1998), was used in the accompanying software AuvTool developed in this study. The approach features the use of bootstrap simulation.

As shown in Figure 2-6, bootstrap simulation is used to simulate the uncertainty in the parameters of a frequency distribution, \hat{F} , that has been fitted to a data set of sample size n . A total of B bootstrap samples of sample size n are simulated. For each bootstrap sample, a new distribution is fitted and a bootstrap replication of the distribution parameters is calculated. The bootstrap simulation produces paired parameter estimates. These multivariate sampling distributions of the parameters represent the uncertainty in the distribution parameters. In the two-dimensional simulation, a total of q different frequency distributions are simulated, where $q = B$ in most cases presented here. Each alternative frequency distribution is based upon a different set of bootstrap replicate distribution parameters. For each alternative frequency

distribution, a total of p random samples are simulated to represent one possible realization of variability within the population. For example, suppose $B=500$ and $p = 500$. Thus, a total of 250,000 samples are generated, representing 500 samples from each of 500 alternative frequency distributions. For each realization of uncertainty, the samples are sorted to represent cumulative distribution functions. Thus, there are 500 values for any given statistic (e.g., mean, variance, 95th percentile of variability) which can be used to construct confidence intervals for each statistic. An example graph of probability bands from two-dimensional simulation was shown in Figure 2-5 of Section 2.8.4.

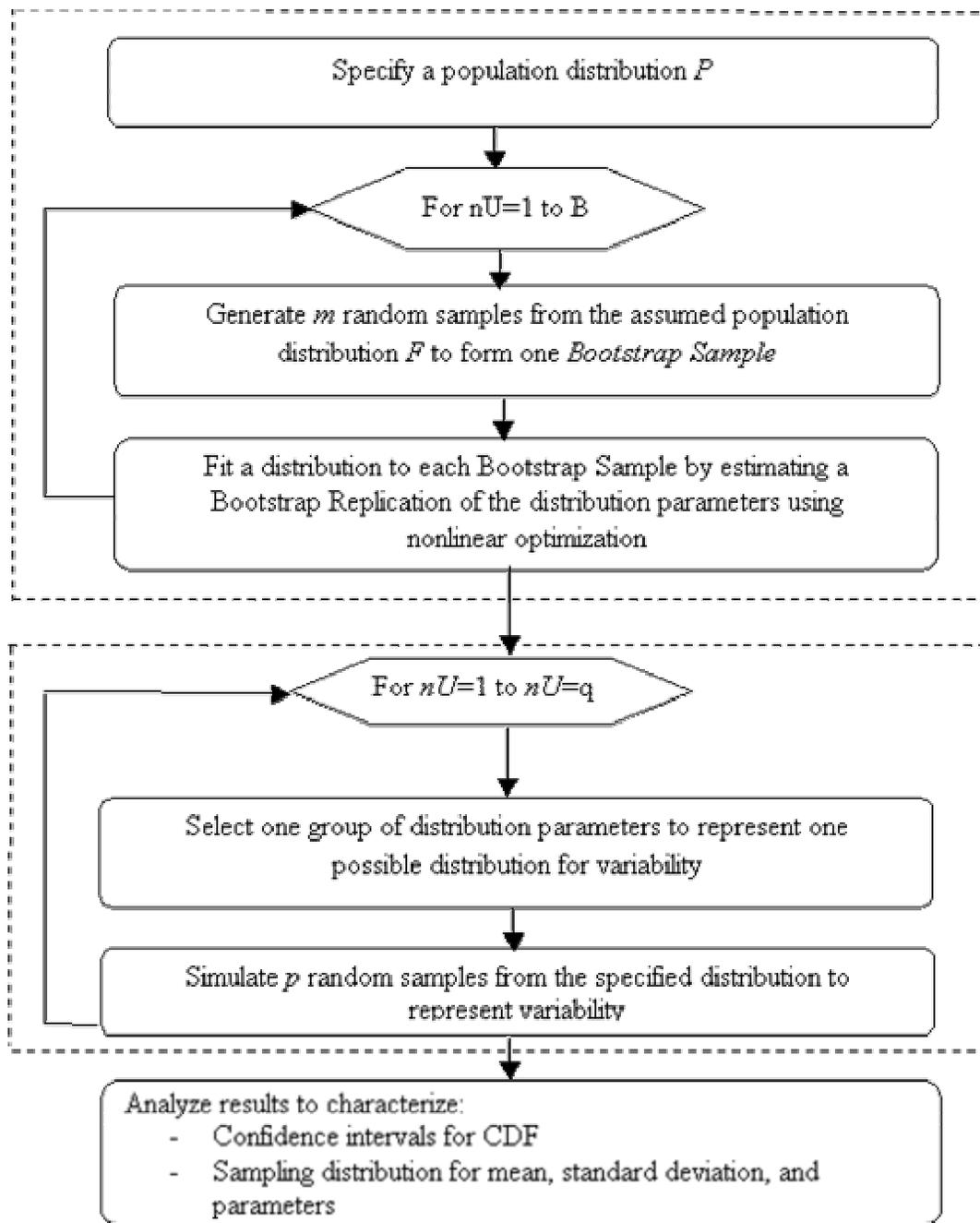


Figure 2-6. Flow Diagram For Bootstrap Simulation and Two-Dimensional Simulation of Variability and Uncertainty. (Where: B=number of Bootstrap Replications, q=Sample Size Used for Uncertainty, p=Sample Size Used of Variability) (Frey and Rhodes, 1998)

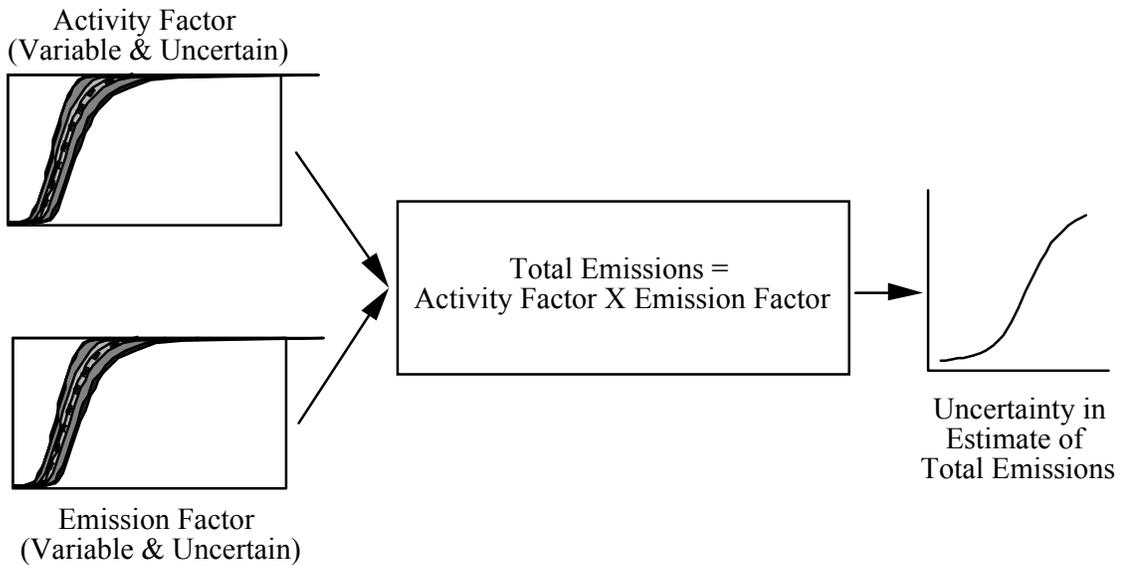


Figure 2-7. Flow Diagram Illustrating the Propagation of Variability and Uncertainty in Emission Inventory Inputs to Quantify the Uncertainty in a specific emission source category

2.11 Probabilistic Approaches for Simulating Variability and Uncertainty in the Emission Inventories

Emission factor and activity factor could also be both variable and uncertain.

Figure 2-7 conceptually illustrates the propagation of variability and uncertainty in the emission factor or activity factor through emission inventory models to quantify the uncertainty in the emission inventory. By using methodologies previously introduced in this part, we have probability distributions describing variability and sampling distributions describing uncertainty for emission and activity factor. The variable and uncertain emission and activity factors are then propagated through the emission inventory model by using numerical sampling technique such as Monte Carlo simulation to simulate the uncertainty in the estimate for the total emissions from a population of emission sources. Based upon the probabilistic simulation, a probabilistic emission inventory from a population of emission sources is developed.

2.12 Identification of Key Sources of Variability and Uncertainty

The identification of the key sources of variability and uncertainty from different model inputs is useful because it can indicate which model input makes most contribution to variability and uncertainty in a selected model output. Such information helps where to target additional research or data collection to reduce uncertainty in a model input, and thereby leading to a reduction in uncertainty in the model output; or will help decision-makers to make correct decisions about which key sources of variability will lead to a significant variation for model output. For example, Identification of key sources of variability and uncertainty in emission inventory has many important implications for decisions, it enables analyst and decision makers to evaluate whether times series trends are statistically significant or not, and to determine the likelihood that an emission budget will be met.

There are many methods to identify the key contributors to variability and uncertainty of model output. For example, summary statistics, scatter plot, correlation coefficient, multivariate linear regression and probabilistic sensitivity analysis etc. (Cullen and Frey, 1999; Morgan and Henrion, 1990). Most of these methods can be used to identify key model input contributor to both variability and uncertainty in a selected model output. More detailed discusses can be found in Cullen and Frey (1999).

Scatter plots and probabilistic sensitivity analysis are probably two general methods to identify the key contributors to the variability in a selected model output. Scatter plot is a direct visual assessment of the significance of the influence of individual model input on a model output. Because each realization in probabilistic simulations such as Monte Carlo simulation generates one value for each input and output. So, simulated pair values can be plotted in two dimensions. Analysts can identify which

model input is a key contributor to model output based on the magnitude that the two dimensions are correlated.

Probabilistic sensitivity analysis is probably the most widely used method to identify the effect of variation of a model input on a model output. In the probabilistic sensitivity analysis, a distribution is assigned to a selected model input, while all other inputs are set to their central values such as mean. By comparing the strength of sensitivity of different model inputs on the effect of the variance of a model output, Analysts can identify which model input contributes most to the variation of the selected model output. Probabilistic sensitivity analysis can provide insight into how the third moment of the inputs may affect the central tendency of output, and is very useful for validating results of a statistical analysis. Another important use of probabilistic sensitivity analysis is that it can assess the relative importance of different sources of variability and uncertainty in two-dimensional analysis (Cullen and Frey,1990).

Multivariate regression analysis and correlation coefficient with rank can be used to identify the key model input contributor to the uncertainty of model output. Multivariate regression analysis considers a least square regression model fitted to estimate the model output as a linear function of the model inputs. The regression coefficients reflect the sensitivity of uncertainty of model output to the corresponding model inputs. In practice, analysts often use standardized regression coefficient to measure uncertainty importance of different model inputs. The standardized regression coefficients of different model inputs can be obtained by multiplying each coefficient by the ratio of the estimated standard deviation of the corresponding model input to model output.

Another important approach of identifying the key contributor to uncertainty of model output is to calculate the correlation coefficient between the model output and a selected model input. In this study and the accompanying software tool AuvTool 1.0, this approach is chosen because it is easy to calculate and program, and its analysis results are easier to be presented and interpreted. The size of correlation coefficients of different model inputs to outputs can measure the degree of association of model input with model outputs. The model input with the largest correlation coefficient is the key contributor to the uncertainty of a selected model output.

For example, an emission source category often consists of multiple subset emission source categories. The overall emission inventory for the emission source category can be characterized by the following equation:

$$EM_{total} = \sum_{i=1}^n EM_i \quad (2-28)$$

where,

EM_{total} = Total emission inventory for the emission source category
(tons/year)

EM_i = The emission at the i^{th} subset source category

n : = The number of the subset source categories

In order to identify which subset source category contribute most to the selected emission source category, *correlation coefficient method* can be used. The sample correlation between a model input, x , and a model output, y , is calculated as follows (Morgan and Henrion, 1990):

$$U_{\rho} = \frac{\sum_{k=1}^m (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^m (x_k - \bar{x})^2 \times \sum_{k=1}^m (y_k - \bar{y})^2}} \quad (2-29)$$

Where,

U_p = Importance of uncertainty from model input y samples.

x_k = Model output samples, in this case, x_k can be considered as the total emission inventory for the specific emission source category.

\bar{x} = The mean of x_k samples.

y_k = Model input samples, y_k can be considered as the subset source category for the emission inventory.

\bar{y} = The mean of y_k samples..

A large magnitude of the uncertainty importance measure, U_p , indicates a stronger linear dependence between the selected model input and model output and the specific subset sources category is a key source of uncertainty in the total emission inventory.

However, the correlation coefficient method does not necessarily provide a good measure of nonlinear monotonic relationship. If the distributions of input or output are far from normal, particularly if they have one or two long tails, they are liable to distortion from the effect of outliers. But using the method of correlation coefficient with rank can avoid this problem. The method ranks the sample values for each input and for the output, and examines rank-order correlations. More information on the method can be found in Cullen and Frey (1999); Morgan and Henrion (1990).

2.13 Summary

This part has described a general approach for developing a probabilistic emission inventory and the steps and methods associated with the development of a probabilistic emission inventory. These steps and methods include the following:

- Development of database used to develop an emission inventory
- Selection or development of emission inventory models
- Numerical sampling techniques
- Plotting of data sets using the Hazen plotting position
- Visualization of the CDF of fitted distributions and graphical comparison of these with the data
- Estimation of parameters for parametric probability distributions using MoMM or MLE approaches
- Presentation of empirical step-wise CDFs
- Generation of random numbers from empirical step-wise CDFs or from parametric probability distribution models
- Calculation of test statistics as an aid in determined goodness-of-fit of a parametric probability distribution to a data set
- Estimation of confidence intervals of the CDF of a parametric probability distribution fitted to a dataset and graphical comparison with the data as an aid in evaluating goodness-of-fit.
- Use of bootstrap simulation to characterize sampling distributions and confidence intervals for key statistics, such as the mean, standard deviation, and parameters of parametric probability distribution models.
- Propagation of uncertainty and variability in model inputs through emission inventory models to estimate uncertainty in category-specific emissions and/or total emissions from a population of emission sources.
- Identification of key sources of variability and uncertainty.

2.14 References

Ang A. H-S, and Tang, W.H., 1984, *Probability Concepts in Engineering Planning and Design, Volume 2*, John Wiley and Sons: New York.

Barry, T., 1996, "Recommendations on the Testing and Use of Pseudo-Random Number Generators Used in Monte Carlo Analysis for Risk Assessment," *Risk Analysis*, 16(1): 93-105.

Beck, L.; Wilson, D. (1997), "EPA's Data Attribute Rating System," In *Emission Inventory: Planning for the Future, The Proceedings of A Specialty Conference, Air & Waste Management Association*, Pittsburgh, PA, pp. 176-189.

Bratley, P., B.L. Fox, and L.E. Schrage, 1987, *A Guide to Simulation*, 2nd ed. Springer-Verlag: New York.

Burmester, D.E., R.H. Harris, 1994, "The Magnitude of Compounding Conservatism in Superfund Risk Assessments", *Risk Analysis*, 13(2):131-143.

Burr, D., 1994, "A Comparison of Certain Confidence Intervals in the Cox model," *J. of the American Statistical Association*, 89, 1290-1302.

Cohen, A.C., B. Whitten, 1988, *Parameter Estimation in Reliability and Life Span Models*, M. Dekker: New York.

Cullen, A.C., H.C. Frey, 1999, *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press: New York.

D'Agostino, R.B., M.A. Stephens, 1986, *Goodness-of-Fit Techniques*, M. Dekker: New York.

Efron, B., 1979, "Bootstrap Method: Another Look at the Jackknife," *The Analysis of Statistics*, 7(1): 1-26.

Efron, B, R.J. Tibshirani, 1993, *An Introduction to the Bootstrap*, Chapman & Hall: London, UK.

EPA, 1995, *Compilation of Air Pollutant Emission Factors* 5th Ed., AP-42 and Supplements, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC.

EPA, 1999a, *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessment*, EPA/630/R-98/004, U.S. Environmental Protection Agency, Washington, DC.

EPA, <http://www.epa.gov/ttn/chief/eiip/>, (accessed 02/20/2002).

ERG, 1997, "Introductions to the Emission Inventory Implementation Program," volume 1, Prepared by Eastern Research Group for Steering Committee, Emission Inventory Implementation Program, Morrisville, NC.

Frey, H.C., R. Bharvirkar, J. Zheng, 1999, "Quantitative Analysis of Variability and Uncertainty in Emissions Estimation," Final Report, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Frey, H.C., D.S. Rhodes, 1996, "Characterizing, Simulating, and Analyzing Variability and Uncertainty: An Illustration of Methods Using an Air Toxics Emissions Example," *Human and Ecological Risk Assessment*, 2(4):762-797.

Frey, H.C., D.S. Rhodes, 1998, "Characterization and simulation of uncertain frequency distributions: Effects of Distribution Choice, Variability, Uncertainty, and Parameter Dependence," *Human and Ecological Risk Assessment*, 4(2):423-468.

Frey, H.C., D.S. Rhodes, 1999, "Quantitative Analysis of Variability and Uncertainty in Environmental Data and Models: Volume 1. Theory and Methodology Based Upon Bootstrap Simulation," Report No. DOE/ER/30250--Vol. 1, Prepared by North Carolina State University for the U.S. Department of Energy, Germantown, MD.

Hahn, G.J., S.S. Shapiro, 1967, *Statistical Models in Engineering*, John Wiley and Sons: New York.

Harter, L.H., 1984, "Another Look at Plotting Positions," *Communications in Statistical-Theoretical Methods*, 13(13): 1613-1633.

Hattis, D., Burmaster, D.E., 1994, "Assessment of Variability and Uncertainty Distributions for Practical Risk Analyses," *Risk Analysis*, 14(5):713:729.

Hazen, A., 1914, "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply," *Transactions of the American Society of Civil Engineers*, 77: 1539-1640.

Holland, D.M., T. Fitz-Simmons, 1982, "Fitting Statistical Distributions to Air Quality Data by Maximum Likelihood Method," *Atmospheric Environment*, 16(5): 1071-1076.

Law, A.M., W.D. Kelton, 1991, *Simulation Modeling and Analysis* 2d ed., McGraw-Hill: New York.

L'Ecuyer, P, 1994, "Uniform Random Number Generation," *Annals of Operation Research*, 53, 77-120.

L'Ecuyer, P., 1996, "Multiple Recursive Random Number Generators," *Operations Research*, 44:816-822.

- Lilliefors, H.W., 1967, "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399-402.
- Martin, M.A., 1990, "On Bootstrap Iteration for Converge Correction in Confidence Intervals," *Journal of the American Statistical Association*, 85, 1105-1108.
- Massey, F.J., 1951, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, 46, 68-78.
- McKay, M.D., R.J. Beckman, W.J. Conover, 1979, "A Comparison of Three Method for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics* 21(2):239-245.
- Morgan, M.G., and M. Henrion, 1990, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York.
- Press, W.H., S.A Teukolsky, W.T. Vetterling, and B.P. Flannery, 1992, *Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed.*, Cambridge University Press: New York.
- Rhodes, D.S., 1997, *Quantitative Analysis of Variability and Uncertainty in Environmental and Risk Assessment*, Masters Thesis, Department of Civil Engineering, North Carolina State University, Raleigh, NC.
- Rubinstein, R. Y., 1981, *Simulation and the Monte Carlo Method*, John Wiley & Sons: New York.
- Seinfeld, J.H., 1986, *Atmospheric Chemistry and Physics of Air Pollution*, John Wiley and Sons: New York.
- Seiler, F.A., J.L. Alvarez, 1996, "On the Selection of Distributions for Stochastic Variables," *Risk Analysis*, 16(1):5-1.
- Small, M.J., 1990. "Probability Distributions and Statistical Estimation," Chapter 5 in *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Morgan, M.G., and Henrion, M., Cambridge University Press: New York.
- Stephens, M.A., 1974, "EDF Statistics for Goodness-of-Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730-737.

PART III

SOFTWARE IMPLEMENTATION

Junyu Zheng

3.0 Software Implementation

The methodology for quantifying variability and uncertainty described in Part 2 was implemented in the accompanying software tools AUVVEE and AuvTool, respectively. AUVVEE was a prototype software tool for calculation of variability and uncertainty in statewide inventories for a selected emission source and pollutants. Its purpose was to demonstrate a general probabilistic approach for developing a probabilistic emission inventories based upon a selected utility power plant NO_x emission source. Because AUVVEE is mainly used for the demonstration, it is restricted to an example case study and not a general tool to do variability and uncertainty analysis. AuvTool was designed as a general tool for quantifying variability and uncertainty in various quantitative analysis fields such as risk or exposure assessment and emission estimation. It implemented many features not included in the AUVVEE. The current AuvTool version has become a module for the EPA SHEDS (Stochastic Human Exposure Dose Simulation) (Zheng and Frey, 2002) model for quantifying variability and uncertainty in SHEDS model inputs.

The design considerations, structure designs, development environment, and the introductions of main function modules for the two accompanying software tools are presented in the following sections.

3.1 Software Implementation of AuvTool

In this section, the design considerations, development environment and tools, structure design, and the main function modules and associated main features of AuvTool are presented.

3.1.1 AuvTool Software Design Considerations

An objective of AuvTool is to make it generally applicable for quantifying variability and uncertainty in various quantitative analysis fields such as risk assessment and emission estimation. Thus, AuvTool was designed as a stand-alone program.

Another goal of AuvTool is to provide a user-friendly preprocessor module for the EPA SHEDS model which incorporates appropriate algorithms for fitting distributions to model inputs and for quantifying variability and uncertainty in each input. Therefore, one design concern in the development of AuvTool system is to make the output of AuvTool appropriate for use as input to the SHEDS model. Because the SHEDS model involves a large number of model inputs, and because variability and uncertainty must be quantified for such inputs, a batch analysis feature was included in AuvTool.

The future objective for AuvTool is planned to have capabilities which allow users to specify their own models, and to propagate the variability and uncertainty from model inputs to model outputs, therefore, the extensibility and expansion of the AuvTool was another main design concern. Based on these considerations, an object-oriented programming technique was used in the development of AuvTool system to promote modularity, extensibility, and reusability of the source code.

3.1.2 Development Environment and Tools

The Windows 98/ME platform was chosen as the AuvTool development environment. This choice was made because the Windows is a widely used operating system; another reason is to ensure compatibility of AuvTool with the SHEDS model. The software development tools used included Microsoft Visual C++, Graphic Server and Spread Active X controls. The reason for choosing Visual C++ lies in that it not only

provides an object-oriented programming environment, which makes the software more extensible and expandable, but it also facilitates the development of a user-friendly graphic interface. The Graphic Server and Spread tools help visualize the simulation results and organize the data input and result outputs.

3.1.3 Structure Design of the AuvTool System

Figure 3-1 shows the conceptual design and the dependency between modules AuvTool system modules. AuvTool can currently be divided into seven groups. Table 3-1 summarizes the composition of the groups and their main functions. As shown in Figure 3-1 and Table 3-1, the Data Import/Export group provides data for the Variability and Uncertainty Analysis group. The analysis results from the Variability and Uncertainty Analysis group are reported to the Variability and Uncertainty Resulting Reporting group, and to the Further Analysis group for further analysis of the sampling distribution data for the statistics of interest (e.g., mean, standard deviation, and distribution parameters). The results from the Further Analysis group are reported to the Variability and Uncertainty Resulting Reporting group for summarization. The modifications of the *Random Seed Setting* module in the Random Sampling group are passed to other analysis modules.

3.1.4 AuvTool Main Modules

As shown in Figure 3-1, AuvTool is composed of different function modules. The following subsections briefly describe the main functions modules and the associated features that the function modules provide.

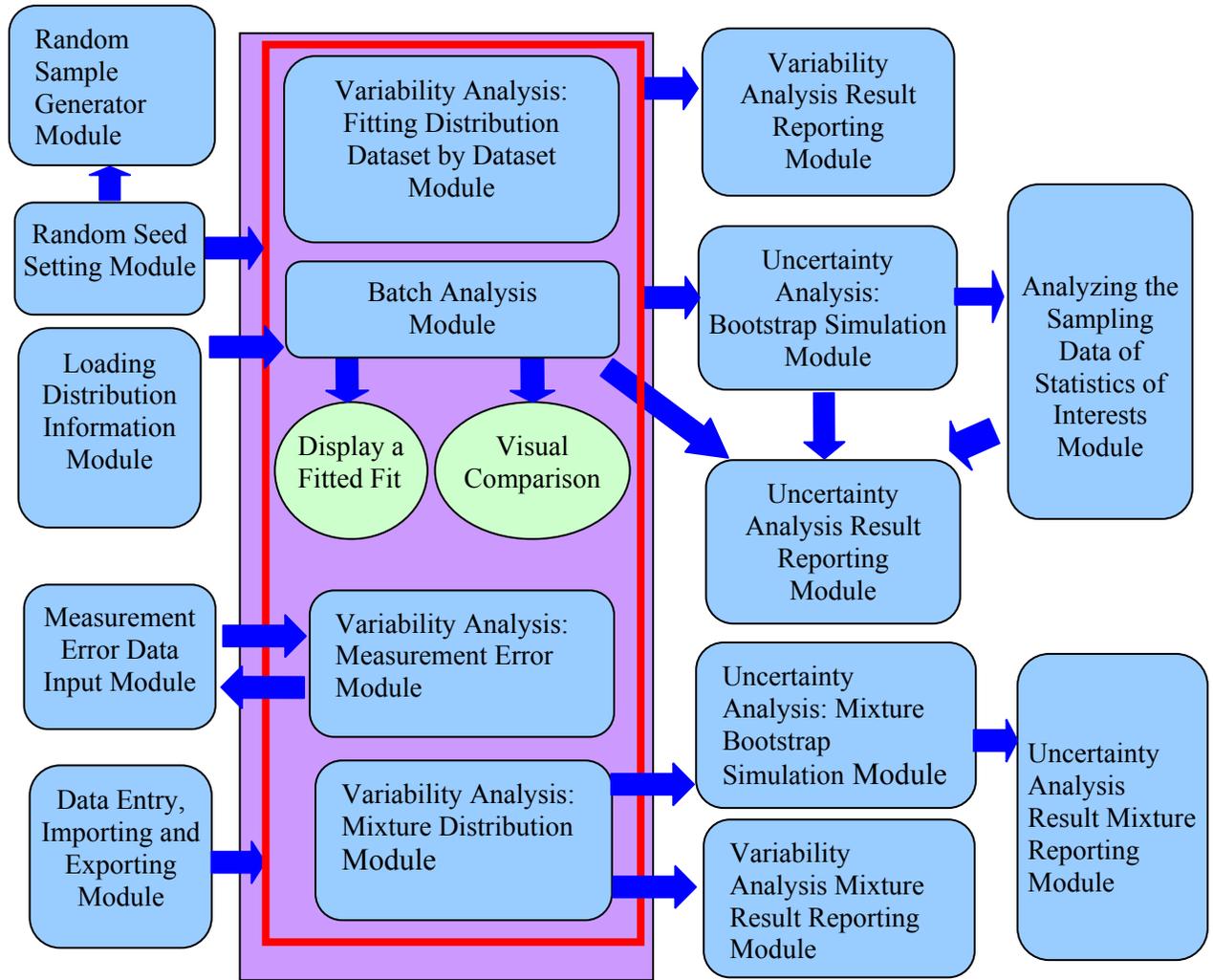


Figure 3-1. The Conceptual Structure Design and Context Diagram of AuvTool System

3.1.4.1 Data Entry, Importing and Exporting Module

The *Data Entry, Importing and Exporting* module provides a data sheet similar to a spreadsheet for users to input or output data. In this module, users can enter data from the keyboard, load an existing AuvTool format data file, and import a Microsoft Excel 97 data file or tab-delimited text files from other application programs into the main data sheet. In the data sheet, AuvTool specifies that each column represents one data set, and users can have multiple data sets by using multiple columns in the input format. Users

Table 3-1. AuvTool Function Module Summarization Table

Group Name	Modules	Main Functions
Data Import/Export	<i>Data Entry, Importing and Exporting</i> module and <i>Loading Distribution Information</i> module	Provides the required data for variability and uncertainty analysis, and exports the input data for future analysis and other applications
Random Sampling	<i>Random Seed Setting</i> module and <i>Random Sample Generator</i> module	Sets the random seeds and generates random samples
Variability and Uncertainty Analysis	<i>Variability Analysis-Fitting Distribution Dataset by Dataset</i> module, <i>Batch Analysis</i> module and <i>Uncertainty Analysis</i> module	Implements all simulations and calculations related to variability and uncertainty analysis
Further Analysis	<i>Analyzing the Sampling Data of Statistics of Interest</i> Module	Does further analysis of the sampling data of interests of statistics from bootstrap simulation
Variability and Uncertainty Result Reporting	<i>Variability Analysis Result Reporting</i> module and <i>Uncertainty Analysis Result Reporting</i> module	Provides summarization tables for user's variability and uncertainty analysis cases
Variability and Uncertainty Analysis with Measurement Error	<i>Variability Analysis: Measurement Error</i> module, <i>measurement Error Data Input Module</i> , <i>Uncertainty Analysis Module</i>	Implements variability and uncertainty analysis when known measurement error is available
Variability and Uncertainty Analysis Using Mixture Distributions	<i>Variability Analysis: Mixture Module</i> , <i>Uncertainty Analysis: Mixture Bootstrap Simulation Module</i> , <i>Variability and Uncertainty Analysis Result Reporting Modules</i>	Implements variability and uncertainty analysis when mixture distribution is used, and provides summarization tables for mixture distribution variability and uncertainty analysis results.

can name each data set. The module automatically counts the number of data points in a data set and logically checks the users' inputs. For example, if there are some invalid numerical value inputs, AuvTool will prompt the user to correct their inputs before they can do variability and uncertainty analysis. This module allows the user to save their data into an AuvTool file format or to export their data to an Excel file or tab-delimited text file. The data in the module will be used in the other analysis modules as a basis of variability and uncertainty analysis.

3.1.4.2 Loading Distribution Information Module

It often happens that users can obtain distribution information for some variables from some other sources such as technical reports, while no original data for those variables are available. However, in this situation, it is still possible for users to do uncertainty analysis by using bootstrap simulation if they have sufficient information about the distribution describing the variable. This information includes the type of parametric distribution, the parameter values, and the sample size. The implementation of the *Loading Distribution Information* module enables users to complete uncertainty analysis for this situation. This module allows users to provide the distribution information from the keyboard, from an existing AuvTool disk file, or from other file formats such as Excel. The information is passed to the batch analysis module for uncertainty analysis. Currently, the module allows users to provide common single component parametric distributions. The distribution models include normal, lognormal, gamma, beta, Weibull, uniform, and symmetric triangle distributions.

3.1.4.3 Random Seed Setting and Random Sample Generator Modules

By default, any analysis modules will use the default random seed provided by AuvTool. However, in some situations, users may want to change the random seed for their needs. For example, they want to check the repeatability of simulation results for different random sample series. Keeping the same seed will help users to duplicate the simulation results. The *Random seed setting* module implemented in AuvTool provides options for users to keep or modify the default random seed. The choice of random setting in this module is passed to all other modules. AuvTool also provides a *random sample generator* module, in which users can generate random samples by specifying the corresponding distribution information and the number of random samples they want to

generate. This module can generate random samples based on an empirical distribution. The results generated in the module can be easily copied or exported to other application programs such as Excel or Notepad.

3.1.4.4 Variability Analysis-Fitting Distribution Dataset by Dataset Module

The *variability analysis-fitting distribution dataset by dataset* module automatically lists the data sets needing to be analyzed based on the data that users provide in the *data entry, importing and exporting* module. In the module, users are able to perform variability analysis data set by data set. This module provides seven distribution types which include normal, lognormal, beta, gamma, Weibull, uniform and symmetric triangle distributions that can be fit to a data set, and (in most cases) a choice of two parameter estimation methods, including method of matching moments (MoMM) and maximum likelihood estimation (MLE).

The user can choose the K-S and A-D statistical goodness-of-fit tests, where applicable, to help in choosing a best fitting parametric distribution for a particular dataset. When users select a data set to analyze, the module allows users to choose the parameter estimation method and the preferred distribution type. The data set and fitted distribution will be graphically and instantly visualized, which will help users to judge if the distribution they chose is a good representation of the data set or not. The K-S test and A-D statistical test results are presented on the right side of the user interface as shown in the Figure 3-2, which shows the value of the calculated test statistic; the critical value of the test statistic and whether or not the test was passed. If the users find that no parametric distribution offers a good enough fit to represent a data set, they can choose an

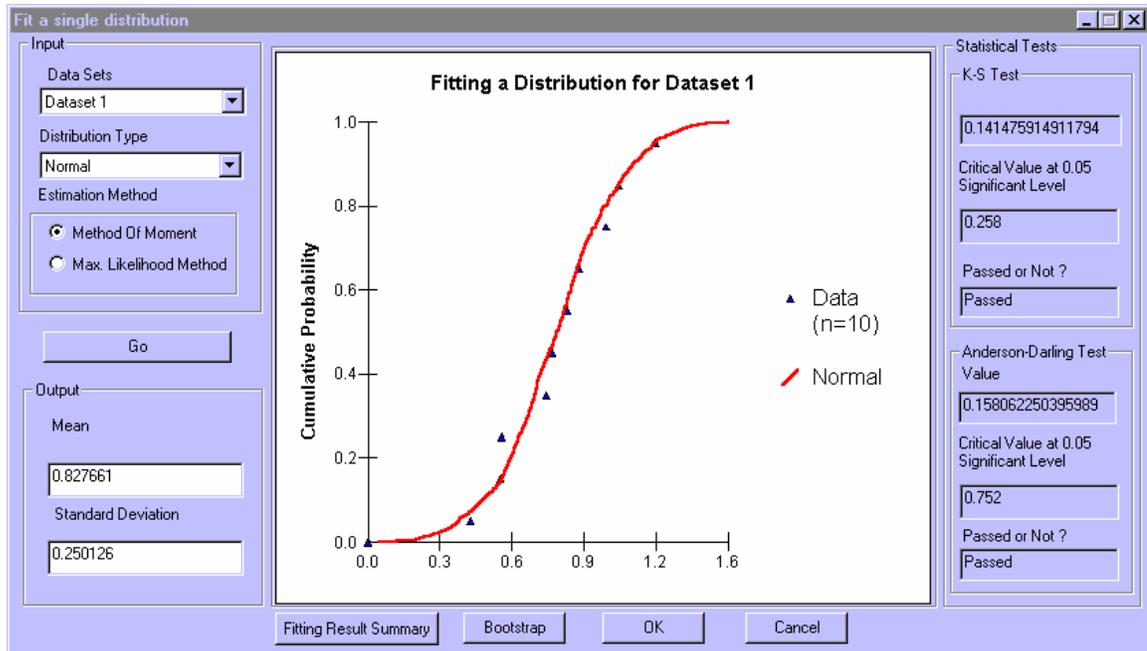


Figure 3-2. Variability Analysis-Fitting Distribution Dataset by Dataset Module

empirical distribution. The decisions made via the module provide a basis for uncertainty analysis as described in Section 3.1.4.6. The variability analysis results in the module are reported to the *variability analysis reporting* module.

3.1.4.5 Batch Analysis Module

The *batch analysis* module is a core one in the AuvTool. Based on data provided in the *data entry, importing and exporting* module and the distribution information in the *loading distribution information* module, the *batch analysis* module automatically generates the control options for each data set or variable being analyzed. In the sheet inside the module, as shown in the Figure 3-3a and 3-3b, each row represents a data set or a variable; any choices and actions made on the selected row will only be effective for the data set or variable on the row.

For any data sets or variables with original data, the program will set “Auto” as the default option in the column of Distribution Choice. The user can modify the default

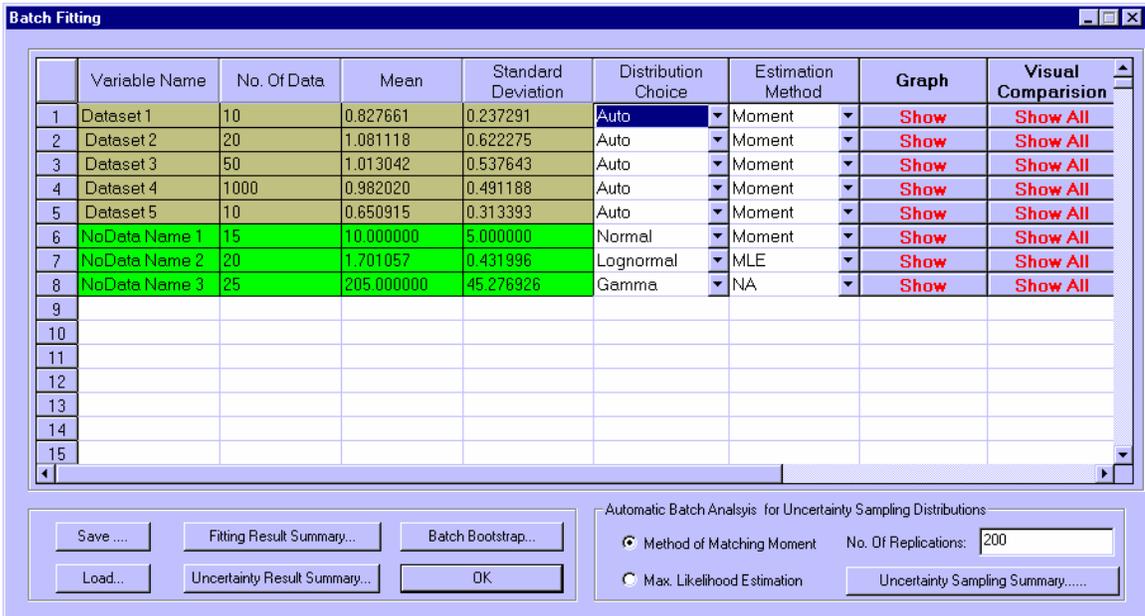


Figure 3-3a. Batch Analysis Module (1)

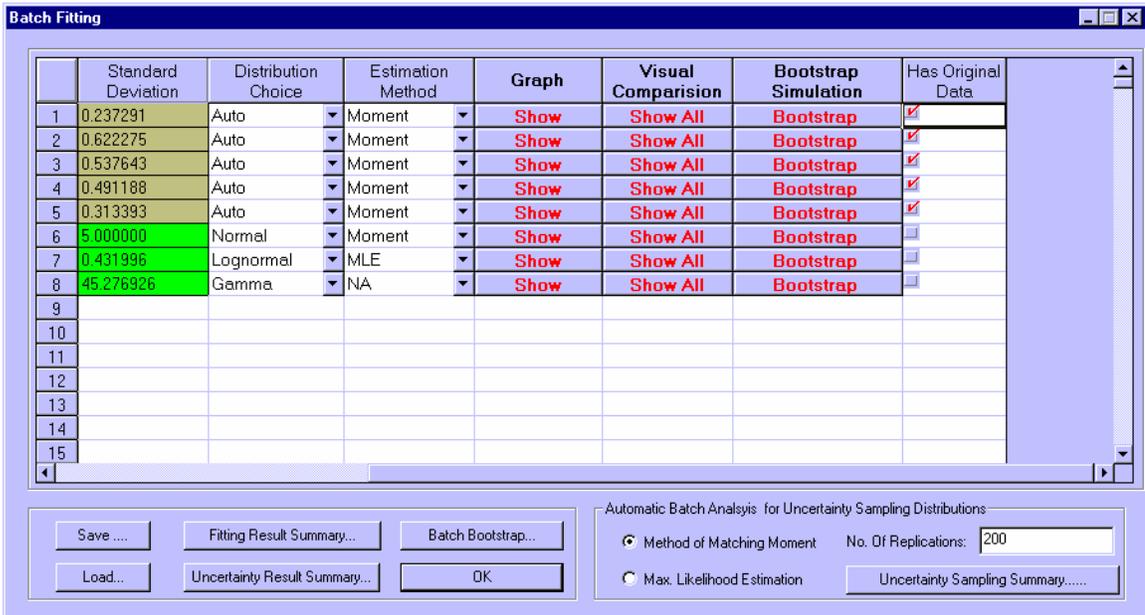


Figure 3-3b. Batch Analysis Module (2)

option to one of the specific distribution types listed in the Distribution Choice combo box. “Auto” is not a distribution type, but an option, in which the user lets the program automatically choose a good fit for the selected data set.

For those cases that do not have original data, there is no “Auto” option available, and distribution information is from the data provided in the *loading distribution information* module. Users cannot modify the distribution type in these cases.

The module also allows users to choose parameter estimation methods. By default, the program will choose MoMM for cases with original data. For those cases without original data, and if no information is available for the parameter estimation methods, the program will mark “NA” on the row of the dataset. However, in uncertainty analysis, the program will by default assign MoMM to these cases. The module provides a feature to graphically display the fitted distribution and the data set. Another main feature of the module is that it allows users to visually compare different distributions fitted to a data set by graphically showing all reasonable fitted distributions in the same window, which will help users to choose a good fit.

The main advantage of this module is that not only it covers all features implemented in the *variability analysis-fitting distribution dataset by dataset* module, but also it provides features of automatic batch variability and uncertainty analysis, visual comparisons of different distribution types fitted to a data set, and uncertainty analysis for the variables without original data. In the module, if users prefer to use the default settings for all data sets analyzed, they do not need to make any choice or to go to any other analysis modules, but they still can complete their variability and uncertainty analyses.

The program automatically helps users choose best fits and performs uncertainty analysis. This feature is very helpful when users have a large number of data sets to be analyzed simultaneously. It must be pointed out that automatically choosing a best fit is

based on a specified criterion. The criterion used in the AuvTool 1.0 is the minimum K-S test value. However, it must also be mentioned that a best fit in terms of the minimum K-S test statistic value does not mean that the fit is the most reasonable one. In fact, users are cautioned that blind application of the K-S criterion to choosing a best fit may lead to selections of parametric distributions that are less than ideal fits in ways not captured by the K-S statistic or that may not have the most relevant theoretical underpinnings.

As mentioned above, the *batch analysis* module allows users to do uncertainty analysis based on the users' own judgments or selections. Any choices made via the module will be passed to the *uncertainty analysis-bootstrap simulation* module to do bootstrap simulations, and will be reported to the *variability analysis-reporting* module.

3.1.4.6 Uncertainty Analysis-Bootstrap Simulation Module

The *uncertainty analysis-bootstrap simulation* module features the use of bootstrap simulation and two-dimensional Monte Carlo simulation for simultaneously quantifying variability and uncertainty. The simulations are based on the choices of distribution types and parameter estimation results from the *variability analysis-fitting distribution dataset by dataset* module or *batch analysis* module.

The module allows users to modify the parameters for bootstrap simulations. For example, users can specify the number of bootstrap replications, and the sample size for variability. The program will, by default, show the probability band graph for the selected variable or data set when the bootstrap simulation is done. An example of band graph is shown in Figure 2-5 in the Part 2. The probability band depicts a plausible range which may enclose the “true” but unknown distribution. For example, the 95 percent probability band may be thought of as a 95 confidence interval. This interval has a 95 percent probability of enclosing the true but unknown distribution. The probability bands

tend to be wider with very small datasets and/or in situations with large variation within the available sample of data. From the probability bands, users can obtain a confidence interval for any percentile of the distribution. This module graphically displays the sampling distributions of the statistics of interest for the selected variable. The sampling distributions are the basis for constructing confidence intervals for the statistics. These statistics include the mean, standard deviation and distribution parameters. Because there are no parameters for an empirical distribution, the statistics for which sampling distributions are reported include only the mean and standard deviation. The module also provides a data sheet to hold the simulation data for the current variable in the data page of the module where users can export the results to other application programs. The simulation results from the module are passed to the *analyzing the sampling data of statistics of interest* module.

3.1.4.7 Analyzing the Sampling Data of Statistics of Interest Module

The sampling distribution data from bootstrap simulation, which describe uncertainty for the selected statistics, are often described using an empirical distribution. The advantage of using empirical distributions is that they do not need any parametric distribution assumptions. However, a potential problem is that there is a large data storage requirement to save all of the replicate values of each statistic. A parametric probability distribution can also be used to represent the sampling distribution for the statistics in a more compact form. For example, in classical statistical theory, the confidence interval for the mean is often described using a normal distribution if the sample data are from a normal distribution or if the sample size is large enough. The use of bootstrap simulation makes the sampling data for statistics available for all other parametric population distribution and eliminates the often restrictive or incorrect

normality assumption imposed upon the sampling distribution of the mean in case with small sample size and skewed data. Therefore, it is often the case that other parametric distributions besides the normal distribution should be used to represent the sampling distribution data for statistics such as the mean. The role of the *analyzing the sampling data of statistics of interest* module in the AuvTool is to implement the further analysis of the sampling data from the bootstrap simulations feature. The batch analysis feature and the further analysis feature in the module embody the advantage of the AuvTool over the other commercial software packages.

This module is very similar to the *variability analysis-fitting distribution module*. The main difference is that the former analyzes the sampling data of statistics from bootstrap simulation for a chosen variable, and uses a parametric distribution model to represent the uncertainty for a statistic, while the later focus on characterizing the variability of a variable based on an original data set using a distribution model. Another difference is that this module also has a feature that can automatically help users to choose a best fit to the sampling distribution data of a statistic; while the *variability analysis-fitting distribution* module does not. Like the *variability analysis-fitting distribution* module, the module also allows users to choose different distribution types and different parameter estimation methods when they analyze a statistic for a selected variable or data set. The choices made via the module will be used to construct the uncertainty analysis summary table in *the uncertainty analysis result- reporting* module.

3.1.4.8 Variability and Uncertainty Reporting Analysis Modules

The purposes of the *Variability and Uncertainty Reporting Analysis* modules are to report the variability and uncertainty analysis results in a tabular form and to facilitate export of the results to other application programs such as Microsoft Excel. The

variability analysis result-reporting module summarizes the variability analysis results from the *variability analysis-fitting distribution* module or *batch analysis* modules.

These results include the summarization of the variable or data set names analyzed, the number of data points for each variable or data set, the distribution types representing variability, the corresponding distribution parameters, the parameter estimation methods, and the K-S and A-D test results. For the beta, uniform and symmetric triangle distributions, the A-D test is not available, and the corresponding cells will be marked “NA”.

The *uncertainty analysis result-reporting* module summarizes the 95 percent confidence intervals for the mean, standard error, and the variable or data set names analyzed, the number of bootstrap replication for each variable or data set, the distribution types fitted to the sampling distributions of the statistics of mean, standard error and distribution parameters, and the K-S and A-D statistical test results for those distributions. The module also reports all pair-wise sampling data combinations of all possible statistics and the correlation coefficients between all statistics.

3.1.4.9 Variability and Uncertainty Analysis with Measurement Error Modules

The modules provide necessary interfaces to implement variability and uncertainty analysis for datasets with known measurement errors. *Variability Analysis-Measurement Error* module automatically lists the data sets needing to be analyzed based on the data that users provide in the *data entry, importing and exporting* module. In this module, before users are able to perform variability analysis for data set with measurement errors, the *measurement error data input* module must be first invoked to obtain measurement error data used to do variability and uncertainty analysis for each

data set. Then, by eliminating measurement error from the observed dataset, error free data set was constructed for each observed dataset and a distribution is fitted to each error free dataset. The choice made in this module will provide a basis to do two dimensional variability and uncertainty analysis by using bootstrap pair technique in the *Uncertainty analysis: bootstrap simulation* module.

3.1.4.10 Variability and Uncertainty Analysis Using Mixture Distribution Modules

These modules implement the variability and uncertainty analysis when mixture distributions are used. *Variability analysis-mixture distribution* module automatically lists the data sets needing to be analyzed based on the data that users provide in the *data entry, importing and exporting* module. In this module, before users try to fit a mixture distribution to a dataset, the program first requests users to construct an empirical distribution for a dataset. This procedure provides users with a visual impression of the distribution shape so that a better initial value can be set to improve the stability of nonlinear optimal results. At present, two component mixture distributions are provided, which include mixture lognormal and normal distributions. The fitting results from the module are passed to the *Uncertainty analysis: mixture bootstrap simulation* module to do uncertainty analysis. The module implements similar features to the *Uncertainty analysis: bootstrap simulation* module. However, the bootstrap simulation is done based on mixture distribution in the module. The variability and uncertainty analysis results from the two modules will be reported to *Variability and Uncertainty Analysis Mixture Result Reporting Modules*. The two modules provide the same summarization forms as the ones that are done in the result reporting modules for single component distribution.

More details on the use of AuvTool and the algorithms used in the AuvTool can be found in Zheng and Frey (2002) and Frey and Zheng *et al.* (2002). The methods for mixture distribution and measurement errors are documented in Part 4 and Part 5. AuvTool has been validated by a team of testing who didn't involve in its development, and has been verified to be correct in quantifying variability and uncertainty.

3.2 Software Implementation of AUVVEE

In this section, the structure and functional design of AUVVEE, the main modules and databases, the relationships among the modules and databases, and development tools are introduced below.

3.2.1 General Structure of the AUVVEE Prototype Software

In AUVVEE, the user sets up a project. The project contains information on the choice of an internal emission factor and activity factors database, project name, project comments, and user data regarding the number of power plant units included in the inventory, the boiler and emissions control technology for each unit, and the capacity of each unit.

Figure 3-4 shows the conceptual design of AUVVEE. AUVVEE is composed of three databases, which include an internal database, a user input database and an interim database. In addition, AUVVEE includes four main modules: (1) fitting distributions; (2) characterizing uncertainty; (3) calculating emission inventories; and (4) user data input. AUVVEE features an interactive Graphical User Interface (GUI).

3.2.2 Databases in the AUVVEE Prototype Software

The internal database for AUVVEE includes emission and activity factors obtained from CEMS data. The development of the internal database was described in detail in Frey and Zheng (2001). The user may select either a 6-month average or a 12-month

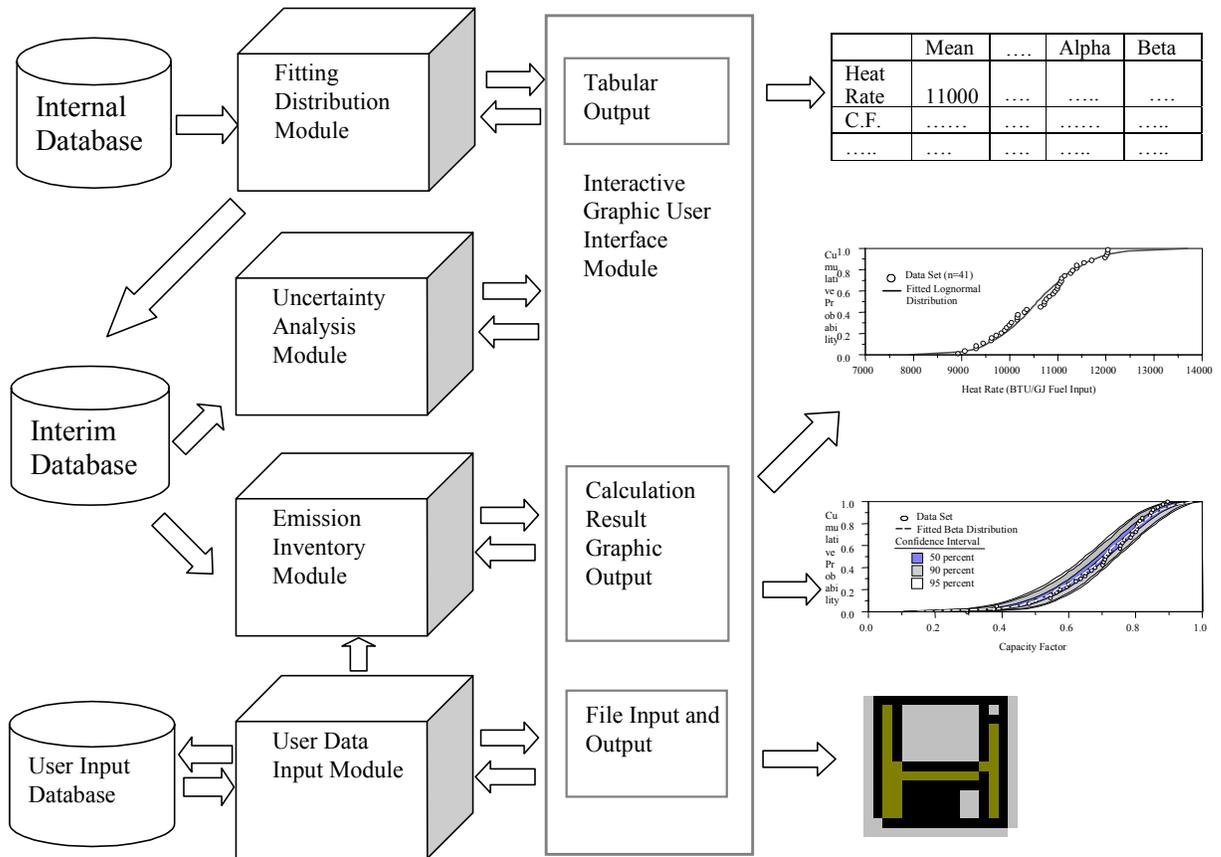


Figure 3-4. Conceptual Structure Design of the Analysis of Uncertainty and Variability in Emissions Estimation (AUVVE) Prototype Software.

average database as the basis for developing either a 6-month or 12-month emission inventory, respectively. The internal database cannot be modified by the user in the prototype version of the software.

The user input database stores data that the user provides regarding the number of power plant units in the emission inventory that the user wants to calculate, the boiler and emission control technology for each unit, and the capacity of each unit. This database can be edited by the user via the user data input module shown in Figure 3-4.

The interim database in AUVVE is used to store the results from the fitting distribution module and to store project information. The interim database provides fitted

distribution information needed by the uncertainty analysis and emission inventory modules shown in Figure 3-4. A default interim database is provided so that the user can proceed to calculate emission inventory results even without making a new selection of parametric distributions to represent each input to the emission inventory. The advantage of the interim database is that it can be used to store default assumptions and can be modified by the user to save project-specific assumptions. The interim database also allows for data to flow between modules of the software.

3.2.3 Modules in the AUVÉE Prototype Software

In this section, each of the four modules indicated in Figure 3- 4 are described. In addition, the GUI is also briefly described.

3.2.3.1 Fitting Distribution Model

The fitting distribution module implements all calculations for fitting parametric distributions to emission factor and activity factor data. This module provides graphs comparing fitted distributions to the data, allowing the user to evaluate the goodness of fit of parametric distributions fitted to datasets from the internal database. The user has the option, via a pull-down menu, to select alternative parametric distributions for fit to the data. When the user exits the fitting distribution model, the current set of fitted distributions are saved to the interim database for use by other modules in the program.

3.2.3.2 Characterizing Uncertainty Module

The characterizing uncertainty module implements the function of characterizing uncertainty in emission factors or activity factors based upon the internal database and based upon the number of units of each technology group that are in the internal database. The characterizing uncertainty module uses data from the interim database to get

distribution information including distribution type and the parameters of the fitted distributions for emission and activity factors. Uncertainty estimates of the mean emission and activity factors, and other statistics, are calculated using the numerical method of bootstrap simulation. The results of the uncertainty analysis are displayed in the GUI. Because this module uses data from the internal database, which may contain a relatively large number of power plant units compared to an individual state emission inventory, the estimates of uncertainty in the mean and in other statistics are typically a lower bound on the range of uncertainty in the same statistic applicable to an emission inventory that includes a smaller number of power plant units.

3.2.3.3 Emission Inventory Module

The emission inventory module has the following functions: (1) it allows the user to visit the user database and append, modify or delete user input data; (2) it characterizes the uncertainty in emission factors and activity factors based on user project data; (3) it calculates uncertainty in the emission inventory; and (4) it calculates the key sources of uncertainty from among the different technology groups. It is via the emission inventory module that the user has access to the user data input module. The estimates of uncertainty in the emission inventory module are based upon the number of power plant units of each technology group specified by the user. For example, although there may be 36 power plant units of a given type in the internal database, the user may have only 10 units of that type in the emission inventory of interest. The uncertainty in the emission and activity factors for that technology group will be estimated based upon a sample size of 10, not 36.

3.2.3.4 User Data Input Module

The user data input module is packaged with the emission inventory module. The user data input module is the portion of the software that enables the user to add, modify, or delete information in the user database.

3.2.3.5 Graphical User Interface (GUI)

The GUI is actually a general control module in AUVÉE, and it makes all of the independent modules, platforms and databases work together. In addition, the GUI is a bridge which links user input to internal implementation within AUVÉE, and provides model output to the user. Through the GUI, the user can build or open a project, enter a database of emission sources, implement user's choice of parametric distributions, view or save all calculation results, and manage the message passing between the different modules.

3.2.4 Software Development Tools

The development of AUVÉE is based on the Windows 95/98 platform. According to different functional requirements and considering convenience of implementation, different software development tools were used for different aspects of the software system. The roles of the different software tools used to develop the AUVÉE prototype software are as follows:

- Visual Fortran 6.0, a product of Digital Equipment Corporation (now Compaq) was used as the programming language for the algorithms that implement the probabilistic simulation capabilities.
- Microsoft Access, a product of Microsoft Corporation, was used to develop the internal and user databases.

- Visual C++ 6.0, a product of Microsoft Corporation, was used to develop the GUI.
- Graphic Sever 5.1, a product of Bits Per Second Ltd., was used to produce charts for visualization of data, fitted distributions, and bootstrap simulation results. These charts are contained within the GUI.

More details regarding the prototype AUVÉE software and algorithms used are available in the User's Manual (Frey and Zheng, 2000) and Technical Documentation (Frey and Zheng, 2001).

3.3 References

1. Zheng, J., H.C. Frey, "AuvTool 1.0 User's Guide," Prepared by North Carolina State University for Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, February, 2002.
2. Frey, H.C., J. Zheng *et al.*, "Technical Documentation of the AuvTool Software Tool for Analysis of Variability and Uncertainty," Prepared by North Carolina State University for Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, February, 2002.
3. Frey, H.C., Zheng, J., "Methods and Example Case Study for Analysis of Variability and Uncertainty in Emission Estimation (AUVVE)," Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, Feb. 2001.
4. Frey, H.C., Zheng, J., "User's Guide for Analysis of Variability and Uncertainty in Emission Estimation," Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S Environmental Protection Agency, Research Triangle Park, NC, Sep., 2000.

PART IV

**QUANTIFICATION OF VARIABILITY AND UNCERTAINTY
USING MIXTURE DISTRIBUTION: EVALUATION OF SAMPLE
SIZE, MIXING WEIGHTS AND SEPARATION BETWEEN
COMPONENTS**

Junyu Zheng and H. Christopher Frey

Prepared to submit to:

Risk Analysis

Quantification of Variability and Uncertainty Using Mixture Distribution: Evaluation of Sample Size, Mixing Weights and Separation between Components

Junyu Zheng and H. Christopher Frey
Department of Civil Engineering, North Carolina State University,
Campus Box 7908, Raleigh, NC 27695-7908, U.S.A.

ABSTRACT

Variability is the heterogeneity of values with respect to different times, locations, and uncertainty refers to lack of knowledge regarding the true value of a quantity. Mixture distributions are potentially useful in the quantification of variability and uncertainty because they can improve the goodness of fit to datasets that cannot be adequately described by a single parametric distribution. In this paper an approach is developed for quantification of the variability and uncertainty based on mixture distributions using bootstrap simulation. 108 synthetic datasets generated from the selected population mixture lognormal distributions are investigated, and properties of quantification of variability and uncertainty based on those mixture distributions are evaluated with respect to variation in sample size, mixing weight and separation between components. Furthermore, mixture distributions are compared with single distributions. Findings include: (1) mixing weight will influence the stability and accuracy of variability and uncertainty estimates; (2) bootstrap simulation results tend to be more stable normally for larger sample size; (3) when two components are well separated, the stability and accuracy of quantification of variability and uncertainty are improved, however, a high uncertainty arises regarding percentile of mixture distributions coinciding with the separated region; (4) when two components are not well separated, single distribution may often be a better choice because it has fewer parameters and better

numerical stability. (5) Dependencies may exist in sampling distributions of parameters of mixtures and are influenced by the amount of separation between the components. An emission factor case study based upon NO_x emissions from coal-fired tangential boilers with low NO_x burners and overfire air is used to illustrate the use of the approach. Results from the use of single parametric distributions are compared with ones from the use of a mixture distribution

KEY WORDS: Variability; uncertainty; mixture distributions; parameter estimation; Bootstrap simulation

1.0 INTRODUCTION

There are a number of distinct sources of variability and uncertainty in risk analysis. Variability is the heterogeneity of values with respect to different times, locations, or members of a population. Uncertainty refers to as fundamental or epistemic uncertainty, arises due to lack of knowledge regarding the true value of a quantity. ^(1, 2, 3, 4) Both variability and uncertainty may be quantified using probability distributions. The interpretation of the distributions differs in two cases. Kaplan and Garrick ⁽⁵⁾ suggest that uncertainty regarding variability may be viewed in terms of probability regarding frequencies. The International Atomic Energy Agency (IAEA)⁽²⁾ interprets distributions for variable quantities as representing the relative frequency of values from a specified interval, and distribution for uncertain quantities as representing the degree of belief, or subjective probability, that a known value is within a specified interval. Morgan and Henrion ⁽⁶⁾ and Frey ⁽⁷⁾ suggest that variability is described by frequency distribution, and that uncertainty in general, including sampling error and measurement error, and estimates based upon judgment, is described by probability distributions.

There are increasing demands for the use of quantitative probabilistic analysis in exposure or risk assessment. For example, Whitmore⁽⁸⁾ and Rish and Marnico⁽⁹⁾ provide background on probabilistic methods and their application in the quantitative analysis of human exposure and risk assessment. There is a growing track record of the demonstrated use of quantitative methods for characterizing variability and uncertainty applied to emission inventories. For example, There have been a number of efforts aimed at probabilistic analysis of various other emission sources, including power plants, non-road mobile sources, and natural gas-fired engines.^(10, 11, 12, 13)

A widely accepted method for uncertainty analysis is to identify inputs to a model or calculation which are known to have uncertainties, and to quantify the uncertainties in each such input using a probability distribution model.⁽¹⁴⁾ The commonly used probability distribution models include the empirical distribution and parametric distributions. A parametric distribution is described by a specific type of distribution, represented by a mathematical equation, and the parameters of the distribution. The parameters are estimated based upon a random sample of data, and the goodness-of-fit of the distribution may be evaluated using a variety of techniques, ranging from visualization methods to statistical tests.⁽¹⁵⁾

While it is possible to use empirical representations of the distribution of available data, rather than parametric distributions, there are some shortcomings to empirical distributions. An empirical distribution may be thought of as a step-function in which each data point is assigned equal probability. No probability is assigned to any interpolated values between observed data, nor is any probability assigned to values below the minimum data point or above the maximum data point. Therefore, analyses

based upon empirical distributions are constrained to the range of observed data, even though it is typically the case that, with more measurements, values lower than the minimum data point or higher than the maximum data point would likely be obtained. The use of parametric distributions allows for interpolation within the range of observed data and for extrapolations beyond the range of observed data to represent the tails of the distribution. As concluded in an expert workshop convened by the U.S. EPA in 1997, the choice of empirical versus parametric distributions is not inherently a matter of right or wrong, but more a matter of preference of the analyst.⁽¹⁶⁾ In practice, a parametric distribution is more often used since it has a compact means for representing either variability or uncertainty in a quantity.

The specification of a probability distribution model for a model input is an essential step to quantitatively characterizing variability and uncertainty. In previous studies, single component distribution models such as the normal or lognormal distribution are often used to describe variability in a quantity. However, in some cases, some single component distributions often cannot well describe the variation in a quantity or are not good fits to a dataset. Because the accuracy of quantifying variability and uncertainty in part depends on the goodness of fit of the distributions with respect to the available data, the use of single distributions that are poor fits to data will lead to bias in the quantification of variability and uncertainty. In these situations, an alternative is to use a finite mixture of distributions. A mixture distribution is comprised of two or one component distributions that are each weighted. Typically, a mixture distribution will produce a better fit to a data set than a single component distribution, because there are more parameters in the mixture distribution than the single component case. With an

improved fit, in most cases there will also be an improvement in the characterization of both variability and uncertainty.⁽¹⁴⁾

Mixture distributions have been extensively used as models in a wide variety of important practical situations because it can provide a powerful way to extend common parametric families of distribution to fit dataset not adequately fitted by single common parametric distribution and it was less restrictive than the usual distributional assumptions.⁽¹⁷⁾ Mixture models have been used in the physical, chemical, social science and biological fields etc., for example, Hariris⁽¹⁸⁾ applied mixtures of geometric and negative binomial distributions to modeling crime and justice data, Kanji⁽¹⁹⁾ described wind shear data. M. Wedel *et al.*⁽²⁰⁾ utilized a finite mixture of Poisson distribution to model the data on customer purchases of books offered through direct mail. David E. Burmaster *et al.*⁽²¹⁾ used mixture lognormal models to re-analyze data set collected by the U.S. EPA for the concentration of Radon²²² in drinking water supplied from ground water, and found that the mixture model yielded a high-fidelity fit to the data not achievable with any single parameter distributions.

Frey and Rhodes^(10, 22) presented a two-dimensional probabilistic approach for simultaneously quantifying variability and uncertainty based on single distributions by featuring the use of bootstrap simulation; and evaluated the property of quantification of variability and uncertainty for several commonly used parametric probability distributions, including normal, lognormal, gamma, Weibull and beta distributions. However, these studies are focused on the use of single component distributions to represent variability in a quantity. The general framework for quantification of variability and uncertainty as presented by Frey and Rhodes^(10, 22) is adopted here.

However, the approach is extended to include mixture distributions as one method for representing variability in a quantity.

Because a mixture distribution has a more complicated mathematical form and more parameters than a single component distribution, the processes of parameter estimation and quantification of variability and uncertainty are more challenging. For example, (1) there are often no analytic parameter estimators available for any mixture distributions; how are the parameters estimated? (2) how will the sample size, mixing weight and degree of separation between components possibly affect on convergence of parameter estimation and stability of bootstrap simulation? (3) there are no random sampling formulas and cumulative probability functions available for any finite mixture distributions, how is a bootstrap sample drawn from a mixture distribution? And (4) how is a confidence interval formed from a given bootstrap sample? There are some important differences in the quantification of variability and uncertainty when comparing the use of a single distribution and mixture distributions. However, there is little study on quantification of uncertainty in statistics, such as the mean, based on mixture distributions

This paper has five main objectives: (1) to develop a demonstrative approach for quantification of variability and uncertainty based on mixture distributions; it include the discussion of parameter estimation of mixture distribution models; of the methodology for the quantification of variability using mixture distributions; and of the methodology for quantification of uncertainty in statistics inferred from the mixture distribution, such as the mean; (2) to evaluate properties of quantification of variability and uncertainty with respect to variation in sample size, mixing weight and separation between

components; (3) to compare the results from two component mixture distributions with the ones from single distributions and hence to figure out under which situations a single distribution might be a better choice; (4) to evaluate the dependencies among sampling distributions of parameters of mixtures; and (5) to illustrate the use of the approach demonstrated in this study. This paper will focus on the discussion of mixture lognormal distribution with two components, but methods introduced here can be easily extended to any other mixture distributions with more than two components.

2.0 METHODOLOGY

In this section, methods for fitting mixture distributions to data are presented. In addition, methods for quantifying uncertainty in statistics estimated based upon the mixture distribution are discussed

2.1 Mixture Distribution

According to the definition from Titterington *et al.* ⁽²³⁾, a mixture model for a random variable or vector, X , takes values in a sample space and can be represented by a probability density function of the form:

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \dots + w_k f_k(x) \quad (1)$$

With

$$w_j > 0 \quad \text{for } j=1, \dots, k$$

And

$$w_1 + w_2 + \dots + w_k = 1$$

Where,

$f(x)$: Probability density function for the mixture model

$f_k(x)$: Probability density function (PDF) for a component of the mixture.

w_k : The mixing weight

Thus, Equation (1) describes a mixture distribution which has a total of k components. Each component is itself a probability distribution. Each component of the distribution has a weight of greater than zero and less than one. For example, a mixture distribution might be comprised of three component distributions, in which the first component has a weight of 0.2, the second has a weight of 0.3, and third has a weight of 0.5. This implies that, for a large number of samples, approximately 20 percent of the samples would be obtained from the first component, 30 percent would be obtained from the second component, and 50 percent would be obtained from the third component.

In most situations, the components of the mixture, $f_j(x)$, have specified *parametric* forms:

$$f(x) = w_1 f_1(x|\theta_1) + w_2 f_2(x|\theta_2) + \dots + w_k f_k(x|\theta_k) \quad (2)$$

Where θ_j denotes the vector of parameters in the probability density function $f_j(x)$. For example, the normal distribution is a parametric distribution, with parameters of mean and standard deviation. Therefore, the vector of parameters in this case would be the mean and the standard deviation. For the gamma distribution, there is a scale parameter and a shape parameter, which comprise the vector of parameters.

For a mixture model with two components, it can be expressed in the form of the following:

$$f(x) = w f_a(x) + (1 - w) f_2(x) \quad (3)$$

with $0 \leq w \leq 1$. In this paper, we will focus on discussion of mixture lognormal distribution with two components. Thus, $f_1(x)$ has the following form:

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(\ln(x) - \mu_i)^2}{2\sigma_i^2}\right] \quad (4)$$

Where,

μ_i = The mean of $\ln(x)$ in the i th component of a mixture model

σ_i = The standard deviation of $\ln(x)$ in the i th component of a mixture model

2.2 Parameter Estimation of Mixture Distributions

A key problem in developing the approach for quantification of variability and uncertainty using mixture distribution is parameter estimation of mixture distributions. Many methods have been devised and used for estimating the parameters of a mixture distributions including, among others, Pearson's method of matching moments, formal maximum likelihood estimation (MLE) approaches, and informal graphical techniques. Pearson first presented the method of moments in 1896 to estimate the parameters in a mixture of two univariate normal densities. From the time of the appearance of Pearson's paper until the use of computers became widespread in the 1960's, only fairly simple mixture density estimation problems were studied. Although some work has been done extending Pearson's method of moments to more general mixtures of normal densities and to mixtures of other continuous densities, the method of matching moments has long been disfavored because of its statistical inefficiency relative to the method of MLE. An efficient statistical estimation method is one that yields a relatively narrow confidence interval for the estimated statistics. In addition, an important problem in the method of matching moments is based on the assumption that the components in a mixture model, or at least some useful statistics associated with them, are known when estimating the

parameters in a mixture models. ^(24, 25) However, the assumption cannot be met in most cases.

With the advent of high-speed computers, more interests are turned to likelihood estimation of the parameters in a mixture distribution. The general idea behind MLE is to choose values of the parameters of the fitted mixture distribution so that the likelihood that the observed data is a sample from the fitted distribution is maximized. For example, Hasselblad treated maximum-likelihood estimation for mixtures of any number of univariate normal densities; his major results were later obtained independently by Behboodian, Day and John used MLE to address mixtures of two multivariate normal densities with a common unknown covariance matrix. ^(24, 25)

In addition to the method of moments and the method of maximum-likelihood, a variety of other methods have been proposed for estimating parameters in mixture densities. Some of these methods are general purpose methods. Others are (or were at the time of their derivation) intended for special mixture problems. For example, Cassie ⁽²⁶⁾ suggested graphical procedures employing probability paper as an alternative to moment estimates. These graphical procedures work best on mixture populations which are well separated in the sense that each component has an associated region in which the presence of the other components can be ignored ^(24, 25). More recently, Diebolt ⁽¹⁷⁾ discussed the estimation of finite mixture distributions through Bayesian sampling by giving a proper prior value for estimators.

In this paper, MLE will be considered as the preferred method for estimating parameters in a mixture distribution due to its relative efficiency and its generality. The likelihood is calculated by evaluating the probability density function for each observed

data point and multiplying the results. Alternatively, and more commonly, the log-transformed version of the likelihood function is used, which is based upon the sum of the natural log of the probability density evaluated for each data point. The general idea is to choose the estimators of the parameters of the distribution so as to make the probability of the sample a maximum. The MLE parameter estimators can be obtained by finding the maximum of a log-likelihood function.

The log-likelihood function of a univariate (describing one data set) mixture distribution is given by:

$$L = \sum_{i=1}^n \ln f(x_i | w, \mu, \sigma) = \sum_{i=1}^n \ln \left\{ \sum_{j=1}^c w_j f_j(x_i | \mu_j, \sigma_j) \right\} \quad (5)$$

Where,

$$\sum_{j=1}^c w_j = 1$$

n , the number of data points

c , the number of components in a mixture distribution

L , Log-likelihood function

μ_j, σ_j , the parameters in the j^{th} component in a mixture distribution

There are three approaches that can be used to find the maximum of Equation (5) and, hence, obtain the parameter estimates of a mixture distribution. One is the application of the Expectation-Maximization (EM) algorithm, suggested by Dempster *et al.*⁽²⁷⁾ The EM algorithm has the advantage of reliable global convergence, low cost per iteration, economy of storage and ease of programming; however, its convergence can be very slow in simple problems which are often encountered in practice,⁽²⁴⁾ and its results are strongly depend upon the initial guesses assumed for the parameters.⁽²⁵⁾ The second

approach is the Newton-Raphson iterative scheme. This scheme requires a calculation of the inversion of the matrix of second derivatives of the log-likelihood function, which is complicated and must be done separately for each combination of parametric distributions assumed in a mixture (e.g., normal, lognormal, gamma, Weibull) thereby limiting general applicability.^(24,25) The third approach is to use nonlinear optimization methods to directly maximize the log-likelihood function by finding optimal values of the parameters. In this paper, nonlinear optimization was chosen to estimate the parameters of a mixture distribution due to its efficiency and wide use.

For a mixture of two lognormal distributions, the following optimization problem is formulated for parameter estimation:

$$\begin{aligned}
 \text{Maximize} \quad & L = \sum_{i=1}^n \ln[w f_1(x_i | \mu_1, \sigma_1) + (1-w)f_2(x_i | \mu_2, \sigma_2)] & (6) \\
 \text{Subject to} \quad & 0 \leq w \leq 1 \\
 & \mu_1, \mu_2 > 0 \\
 & \sigma_1, \sigma_2 > 0 & \text{for mixture log normal}
 \end{aligned}$$

Where,

n= the number of samples

The optimization problem here is a multidimensional constrained one. A variety of methods are available to solve such problems. These include: the downhill simplex method; the direction-set method, of which Powell's method is the prototype;⁽²⁸⁾ the penalty function method; and others. In this paper, Powell's method is employed. This method is relatively easy to program and provides good results

2.3 Quantification of Variability and Uncertainty Using Mixture Distribution

Bootstrap simulation, introduced by Efron in 1979, is a numerical technique originally developed for the purpose of estimating confidence intervals for statistics

based upon random sampling error.⁽²⁹⁾ The confidence interval for a statistic is a measure of the lack of knowledge regarding the value of the statistic: the larger (wider) the confidence interval, the greater the uncertainty. Bootstrap has been widely used in the prediction of confidence intervals for a variety of statistics. For example, Angus⁽³⁰⁾ developed a bootstrap procedure to calculate the upper and lower confidence bounds for the mean of a lognormal distribution based on complete samples. Freedman and Peters⁽³¹⁾ presented empirical evidence that the bootstrap method provides good estimates of standard errors of estimates in a multi-equation linear dynamic model.

In quantifying variability and uncertainty using bootstrap simulation, there are two major aspects. The first aspect is a procedure for generating random samples from an assumed population distribution, and the second aspect is the method of forming confidence intervals for statistics estimated from the random samples⁽³⁰⁾. The notion behind bootstrap simulation is to repeatedly simulate a synthetic data set of the same sample size as the observed data. The observed data are used either to specify an empirical probability distribution or as a basis for fitting a parametric probability distribution. In either case, the distribution developed based upon the observed data is assumed to be the best estimate of the true but unknown population distribution from which the data are but a finite sample. Numerical methods may be used to generate random samples from the assumed population distribution.⁽¹⁵⁾ In order to simulate random sampling error, a synthetic data set of the same sample size as the observed data is simulated, and statistics such as the mean, standard deviation, distribution parameters, percentiles, and others may be calculated. The process of simulating synthetic data sets of the same sample size as the observed data is repeated perhaps 500 to 2,000 times.

Each time, new values of the statistics are estimated. Each synthetic data set is referred to as a "bootstrap sample" and represents one possible realization of observed values from the assumed population distribution. Each statistic estimated based upon a single bootstrap sample is referred to as a "bootstrap replicate" of the statistic. The set of 500 to 2,000 bootstrap replicates of a statistic represent a "sampling distribution". A sampling distribution is a probability distribution for a statistic. From a sampling distribution, a confidence interval can be inferred.

While there are standard numerical methods for drawing random samples from single component parametric distributions (e.g., see Cullen and Frey ⁽¹⁵⁾ for an overview), the methods for drawing random samples from mixture distributions are more complicated in the context of bootstrap simulation. Although it is possible to obtain a single random sample from a mixture distribution by sampling from a weighted proportion of single component distributions, one of the objectives in bootstrap simulation is to develop confidence intervals for all statistics, including the component weights. Therefore, it is necessary to develop an estimate of the assumed population distribution in a manner that allows for the weight to vary randomly from one bootstrap sample to the next. For this purpose, an empirical distribution is used to represent the assumed population distribution for the mixture.

As shown in Figure 1, the first step in developing the assumed population distribution is to generate a large number of random samples using standard simulation methods. For example, suppose there is a mixture of two lognormal components, one with a weight of 40 percent and the other with a weight of 60 percent. In order to develop a stable estimate of the cumulative distribution function of this mixture, one may

choose to simulate 2,000 or more random values. Thus, on average 800 values would be simulated from the first component, and on average 1,200 values would be simulated from the second component. These values would be rank-ordered in order to describe the cumulative distribution function. The cumulative distribution function of the assumed population may be represented by an empirical distribution of these 2,000 values.

Once an empirical representation of the assumed population mixture distribution is available, it is then possible to randomly sample from it to generate bootstrap samples, as indicated in Figure 1. From each bootstrap sample, the bootstrap replicates of the component parameter values and of the weight may be estimated. For each bootstrap replication of the distribution parameters, the mean and other statistics may be simulated. The sampling distributions of these statistics are the basis for estimating confidence intervals for these statistics.

There are several variations on bootstrap simulation. Methods commonly studied in the literature include the percentile, hybrid, bootstrap-t, and Efron's BCa⁽³¹⁾. The percentile method is possibly the most frequently used in practice though the theoretical justification for this method is the weakest,⁽³²⁾ however, the intervals obtained from this method are the simplest to use and explain. The Hybrid method is justified by asymptotic results for the bootstrap in complicated models. The bootstrap-t and the BCa intervals are comparable in that both have been demonstrated theoretically to be "second-order correct" for one-sided intervals in some relatively simple situations.⁽³²⁾ However, the process for estimating BCa confidence interval is very complicated and the computation burden is heavy. In this paper, for simplicity and because it is the most widely used method in practice, the percentile method is used as a main one to construct bootstrap

confidence intervals. As a comparison, BCa method is provided to construct the confidence interval for the mean statistic.

3.0 INTRODUCTION TO STUDY DESIGN

In order to illustrate the use of the approach for quantification of variability and uncertainty using mixture distribution and to evaluate the behavior of confidence interval of the cumulative density functions (CDFs) due to variation in sample size, mixing weight and magnitude of component separation, the synthetic datasets with different sample size, mixing weights and magnitude of separation between two component were generated from the assumed population mixture lognormal distributions with two components. The assumed population mixture lognormal distributions are described in the Table 1.

There are 12 groups of population mixture distributions listed in Table 1. The difference in mean and standard deviation between components reflects the variation in component separation. For example, when $\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=1.5$, $\sigma_2=0.5$, two components are said not to be well separated; however, when $\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=6.0$, $\sigma_2=0.5$, two components are said to be well separated.

To investigate the effect of variation in mixing weight and sample size on quantification of variability and uncertainty, for each group, different mixing weights and sample sizes were considered, the weights with 0.1, 0.3 and 0.5, and sample sizes with 25,50 and 100 are studied, separately. Therefore, there are 9 synthetic datasets with different mixing weights and sample size generated from each group of population distribution, in total, 108 synthetic datasets which cover the variation in mixing weight, sample size and separation were studied and analyzed in this study.

4.0 RESULTS AND DISCUSSION

The results of selected case studies are analyzed in the section. The following subsections present the results about the properties of confidence intervals of CDFs of the fitted mixture distributions, comparisons between single distribution and mixture distributions and dependencies among sampling distributions of parameters of mixture distributions.

4.1 Properties of Confidence Intervals of Cumulative Distributions

Parts of selected case study results of variability and uncertainty using the approach introduced in the Section 2 are presented in this paper. Figure 2 shows 95 percent confidence intervals of the cumulative distribution of two component lognormal distributions fitted to a mixture population distribution with $\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=1.5$, $\sigma_2=0.5$ when sample sizes are 25, 50 and 100, mixing weights are 0.1, 0.3 and 0.5, respectively. In Figure 2, the thick black line represents a population mixture distribution; the thin lines represents the 95 percent confidence interval of fitted mixture distribution. The results shown in the Figure 2 suggest that reasonable simulation results can be obtained for all cases when mixing weight is 0.5. However, when weight is 0.3, the case of $n=25$ failed; and when weight is 0.1, all cases failed. In this paper, there are two criteria to judge if or not a case fails, one is if a complete bootstrap simulation can be finished or not, another is if or not simulation results are reasonable or correct. Figure 2 represents the situation at which two components are not separated well, the shape of cumulative distributions of the population mixture distributions all look like a single component distribution.

Figure 3 displays results of confidence intervals of cumulative distribution of two component lognormal distributions fitted to a mixture population distribution with

$\mu_1=1.0, \sigma_1=0.5, \mu_2=2.0, \sigma_2=0.5$. The cases shown in the Figure 3 have higher separation when compared to the cases shown in the Figure 2. Almost all cases can successfully gain reasonable predication results of confidence interval except the two cases of $n=25$ and $n=50$ when weight is 0.1.

Figure 4 and Figure 5 show the simulation results for a mixture lognormal with $\mu_1=1.0, \sigma_1=0.5, \mu_2=3.0, \sigma_2=0.5$, and a mixture lognormal population distribution with $\mu_1=1.0, \sigma_1=0.5, \mu_2=6.0, \sigma_2=0.5$, separately. Figure 4 represents the situation in which two components are moderately separated, and Figure 5 represents the situations in which two components are highly separated. From both Figure 4 and Figure 5, it is found that there are wider confidence intervals at the location where two components are separated. In all cases shown in the Figure 4 and Figure 5, only the cases that sample size are 25 and 50 for weight =0.1 failed.

The simulation results in Figure 2 through 5 display some common characteristics: (1) when weight is 0.5, all cases with different sample size can succeed in calculating confidence intervals, but when weight is 0.1, the case with $n=25$ and 50 failed. This is probably because there is relatively more uniform sampling behavior for weight of 0.5. However, when weight is 0.1, due to the severe unbalance of proportions that two components account for in a mixture distribution, any poor sampling behavior in bootstrap simulation will possibly lead to the failure of parameter estimation for the bootstrap sample or inaccurate estimates of parameters, especially when sample size is small. It indicates that degree of balance in mixing weights of components will affect the stability and accuracy of variability and uncertainty estimates; (2) with the increase of sample size, the stability of bootstrap simulation also increase and the range of

confidence intervals become narrower. This is because the increase of sample size improves the sampling behavior and reduces the variability in the bootstrap sample, which leads to the improvement of stability and uncertainty estimates. The result shows that large sample size will favor the quantification of variability and uncertainty based on mixture distribution; and (3) Figure 2 through Figure 5 also display the variation from the slight separation to high separation between two components. The results suggest that good separation improves the stability of simulations and the accuracy of estimate results. However, with the increase of separation, there appears to be a "bulge" in the confidence interval in the region of cumulative probability representing the inflection point between one component of the mixture and the other component of the mixture. The reason that the phenomenon arises is because the weighting parameter is a source of uncertainty as reflected in the confidence interval for the mixture distribution. Because the weight parameter is itself a random variable, there is uncertainty regarding where the inflection point between the components should be, leading to a widening of the confidence interval a wider confidence interval or a higher uncertainty. These results indicate that although stability and accuracy of quantification of variability and uncertainty are improved when two components are well separated, higher uncertainty arises in the separated region.

The results not shown in this paper for case studies with $\sigma=0.1$ and $\sigma=1.0$, also display similar characteristics to the results with $\sigma_1=0.5$ shown here.

4.2 Comparisons between Single Distribution and Mixture Distributions

As a comparison, a single lognormal distribution is used to fit the datasets generated from the specified mixture population distribution listed in the Table 1. Figure 6 and Figure 7 show 95 confidence intervals of cumulative distribution of single lognormal distributions fitted to mixture population distributions with $\mu_1=1.0$, $\sigma_1=0.5$,

$\mu_2=1.5, \sigma_2=0.5$ and $\mu_1=1.0, \sigma_1=0.5, \mu_2=2.0, \sigma_2=0.5$, respectively. Table 2 and Table 3 summarize the results of 95% confidence intervals of selected statistics of single and two component mixture lognormal distributions fitted to mixture populations with varying component separation and standard deviation for $n=100$ when weights are 0.3 and 0.5, separately.

These results shown here suggest that (1) when two components are not well separated, especially when there is one standard deviation difference between two components, a single lognormal distribution fitted to the bootstrap samples from a mixture population distribution also can accurately quantify the variability and uncertainty, its estimates of 95 confidence interval of selected statistics are almost the same as the ones using a mixture lognormal distribution. The finding is very useful in practice if a single distribution rather than mixture distribution can be used to represent a mixture population distribution since bootstrap simulation based on mixture distribution is often unstable, especially when two components are not well separated. The results suggest that an additional component is not necessary when two components are not well separated even if the underlying distribution is a mixture one. (2) With the increase of the magnitude of separation, using single distributions is obviously unreasonable and their estimate results are not accurate. Around the location where two components are separated, the 95 confidence intervals can even not enclose the population value. It suggests that goodness of fit will have an impact on the accuracy of variability and uncertainty estimates. In such situation, it is worthwhile to use a mixture distribution to improve the goodness of a fit, even though more efforts in parameter estimations and bootstrap simulation are needed.

However, even though more components may improve the fit to a particular data set, complications may arise if a larger number of parameters are used. For example, while a two-component mixture of two-parameter distributions has a total of five parameters, a three-component mixture would have a total of eight. If the number of parameters becomes large with respect to the number of available data points, improvements in fit may arise spuriously because of over-fitting. In addition, it is possible that numerical simulation problems will arise in attempting non-linear optimization with a large number of parameters as discussed above. While it is clear that a two-parameter mixture can offer substantial benefits compared to a single component distribution, in terms of improved fit, it is likely that the marginal improvement in fit will diminish as more and more components are added to the mixture. Thus, there is a clear trade-off between an improved fit and the width of the confidence interval of the fitted distribution. As Leoroux²⁹ points out, the elimination of unnecessary components in a mixture might lead to more precise estimates of the parameters, and, by extension, of other statistics.

4.3 Dependencies among Sampling Distributions of Parameters of Mixture Distributions

The dependencies among estimated parameters of a fitted mixture distribution with different separation magnitude are investigated. Figure 8~ Figure 10 display scatter plots of bootstrap simulation results for parameters of two component lognormal distributions with $\mu_1=1.0, \sigma_1=0.5, \mu_2=2.0, \sigma_2=0.5$; $\mu_1=1.0, \sigma_1=0.5, \mu_2=3.0, \sigma_2=0.5$; $\mu_1=1.0, \sigma_1=0.5, \mu_2=6.0, \sigma_2=0.5$ for $n=100$, and weight of 0.5, respectively. Figure 8 ~ Figure 10 actually represents three situations from slight separation to high separation. The results shown in the figures suggest that there exists obvious linear or dependent

relationship between parameters when two components are slightly separated. However, with the increase of the magnitude of the separation, the dependent or correlated linear relationships become weaker and weaker. These results are reasonable. When two components are not well separated, there is a small difference between the means of samples from two components, random variation in mixing weights will lead to an obvious decrease or increase of mean and standard deviation in one component, while obvious increase or decrease of mean and standard deviation in another component. The parameters in a distribution are associated with the mean and standard deviation of samples, therefore, an obvious linear relationship between parameters in such a situation is found. However, when there is a higher separation between two components, it implies that there is a bigger difference between means of two components, the variation in the mixing weight will not lead to a distinct increase or decrease of mean in a component, hence the correlated relationship is weaker. It is possible two components are uncorrelated or independent if they are well separated. These results indicate that the correlation among parameters in a mixture distribution is associated with the magnitude of separation between two components. When two components are well separated, there is less correlation among the parameters; however, when there is no obvious separation between components, an obvious linear relationship among parameters is found. It must be pointed out that the conclusion should be true for mixture lognormal or normal distributions, for other kinds of mixture distributions, it need to be further confirmed.

5.0 AN ILLUSTRATIVE CASE STUDY: NO_x EMISSION FACTOR FOR A COAL-FIRED POWER PLANT

The methodology for simulating variability and uncertainty based upon mixture distributions is demonstrated via a case study of an emission factor for a type of coal-

fired power plant. The case study is based upon a six month average NO_x emission factor for a tangential-fired, coal-fired boiler with low NO_x burners and overfire air. The specific emission control technology is referred to as "LNC1". The dataset is derived from a 1998 US EPA database based on a six month average.⁽³³⁾ This scenario was chosen because: (1) this dataset can not be fit well by any single distribution (Zheng and Frey, 2001); (2) NO_x is one of the most important primary pollutants from power plants; and (3) the number of data points is relatively small ($n=41$).

5.1 Parameter Estimation for the Fitted Distribution

A mixture distribution with two lognormal components was fit to the case study dataset. The parameter estimation results for the mixture of two lognormal distributions are:

Mixing weight=0.291

1st component: Mean of $\ln(x)=6.071$, Standard deviation of $\ln(x)=0.368$

2nd component: Mean of $\ln(x)=6.249$, Standard deviation of $\ln(x)=0.0898$

The fitted distributions are shown in Figures 11 for the two-component lognormal mixture distribution.

5.2 Variability and Uncertainty in the NO_x Emission Factor

The results of the two-dimensional bootstrap simulation of the fitted distributions for the NO_x emission factor are shown in Figures 12 for the two-component lognormal mixture distribution. The dark-gray areas represent the 50 percent probability band of the results, the light gray areas depict the 90 percent probability band, and the white areas show the 95 percent probability band. The empirical distribution of the dataset is plotted with open circles, and the fitted distribution is drawn with a line.

On average, we expect that five percent of the data, or two of the 41 observed data points, will fall outside of a 95 percent confidence interval if the data are a random

sample from the assumed population distribution. In the case of the mixture distribution shown in Figure 6, none of the data are outside of the 95 percent confidence interval, and only approximately 12 percent of the data are outside of the 50 percent confidence interval. This indicates that the data are highly consistent with the assumed mixture distribution and that the mixture lognormal distribution is a good assumption regarding the unknown population distribution, in that they are consistent with the observed data. Figure 12 also display that there is a good agreement in the tails of the mixture distribution with the observed data. The agreement is important in that if errors in the tails of fitted distributions are largest, they will confound comparisons between distributions and bring bias in the quantification of variability.

It is typically the case that the confidence interval for a positively skewed fitted single component distribution is widest at the upper percentiles of the distribution. However, in the case of the fitted mixture distribution, there is also a widening of the confidence interval at a cumulative probability between approximately 0.05 and 0.40. Table 4 shows estimates of uncertainty in the parameters of the fitted mixture distribution. The 95 confidence interval of weight parameter is from 0.045 to 0.553. The range of uncertainty in the weight parameter causes the ‘bulge’ in the confidence interval of the fitted mixture distribution.

5.3 Uncertainty in the Mean NO_x Emission Factor

Compared to the results for single component distributions done in Zheng and Frey⁽¹⁴⁾, the 95 percent confidence interval for the mean by using percentile method is narrowest for the case of the mixture distribution, with a range from -6.5 percent to +6.4 percent of the estimated mean value, or from 471 gram/GJ fuel input to 536 gram/GJ fuel input; and the 95 percent confidence interval for the mean by using BC_a method ranges

from 483 gram/GJ fuel input to 534 gram/GJ fuel input. In contrast, the width of the estimated confidence interval is as much as 30 percent wider based upon the single component distributions, as in the case of the Weibull distribution.

6.0 CONCLUSION

Mixture distribution has the potential to be very useful in the quantification of variability and uncertainty because it can improve the goodness of fit to dataset not adequately described by a single parametric distribution. In this paper an approach for quantifying the variability and uncertainty based on mixture lognormal distribution with two components was demonstrated. The approach can be easily extended to other kinds of mixture distribution with more than one or two component.

Properties of quantification of variability and uncertainty using mixture distribution are evaluated and investigated with respect to variation in sample size, mixing weight and component separation. The findings include: (1) mixing weight will influence the stability and accuracy of variability and uncertainty estimates; (2) bootstrap simulation results tend to be more stable normally for larger sample size; (3) when two components are well separated, the stability and accuracy of quantification of variability and uncertainty are improved, however, a high uncertainty arises regarding percentile of mixture distributions coinciding with the separated region; (4) when two components are not well separated, single distribution may often be a better choice because it has fewer parameters and better numerical stability. (5) Dependencies may exist in sampling distributions of parameters of mixtures and are influenced by the amount of separation between the components.

The use of mixture distributions is a promising method for improving the fit of distributions to data and for obtaining improved estimates of uncertainty in statistics

estimated from the fitted distribution. The use of mixture distributions should be considered and evaluated in situations in which single component distributions are unable to provide acceptable fits to the data, or in situations in which it is known that the data arise from a mixture of distributions. In the illustrative case study, we have successfully demonstrated a method for fitting mixture distributions to data and for making inferences regarding uncertainty.

The characterization of uncertainty in emission factors and other components of emission inventories, as well as in environmental modeling in general, is an important means for conveying to analysts and decision-makers quantitative information regarding the comparative strengths and limitations of inputs to an analysis. Those inputs which contribute most to uncertainty in environmental decisions, and which are amenable to additional study, should be identified and targeted for additional data collection or research to reduce uncertainty. The methods presented in this paper, therefore, are intended to support a rational approach to environmental decision-making.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) grants program via grants R826766 and R826790.

REFERENCE

1. Bogen, K.T. and Spear R.C., "Integrating uncertainty and interindividual variability in environmental risk assessment," *Risk Analysis* **7**, 427-436 (1987).
2. IAEA (International Atomic Energy Agency), "Evaluating the reliability of predictions made using environmental transfer models", *Safety Series*, No.100, Vienna, Australia, International Atomic Energy Agency (1989).
3. Hattis, D. and Burmaster, D.E., "Assessment of variability and uncertainty distributions for piratical risk analyses," *Risk Analysis* **14**, 713-729 (1994)
4. Haimes, Y.Y, Barry, T., and Lambert, J.H., Eds., " Workshop proceedings: When and how can you specify a probability distribution when you don't know too much? " *Risk Analysis* **14**, 661-706 (1994).
5. Kaplan, S. and Garrick, B.J., "On the quantitative definition of risk," *Risk Analysis* **1**, 11- 27 (1981).
6. Morgan, M.G., and M. Henrion, *Uncertainty: A Guide to dealing with uncertainty in quantitative Risk and Policy Analysis* (Cambridge University Press, New York, 1990).
7. Frey, H.C., "Variability and Uncertainty in Highway Vehicle Emission Factors," Emission Inventory: Planning for the Future (held October 28-30 in Research Triangle Park, NC), Air and Waste Management Association, Pittsburgh, Pennsylvania, pp. 208-219, (1997).
8. Whitmore, R.W., "Methodology for Characterization of Uncertainty in Exposure Assessments," EPA/600/8-85/009, Prepared by Research Triangle Institute for U.S. Environmental Protection Agency, Washington, DC (1985)
9. Rish, W.R. and Marnicio, R.J., "Review of Studies Related to Uncertainty in Risk Analysis," ORNL/TM-10776, Prepared by Research Triangle Institute for U.S. Environmental Protection Agency, Washington, DC (1988)
10. Frey, H.C. and Rhodes, D.S., " Characterizing, simulating and analyzing variability and uncertainty: An illustration of methods using an air toxics emissions example," *Human and Ecological Risk Assessment* **2**, 762-797 (1996).
11. Frey, H.C., R. Bharvirkar, J. Zheng; "Quantitative Analysis of Variability and Uncertainty in Emissions Estimation"; Final Report, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, pp 1-10, (1999).
12. Frey, H.C., and S. Bammi; "Quantification Of Variability and Uncertainty in Lawn And Garden Equipment NO_x and Total Hydrocarbon Emission Factors,"

- Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, (2001).
13. Frey, H.C., and S. Li; "Quantification of Variability and Uncertainty in Natural Gas-fueled Internal Combustion Engine NO_x and Total Organic Compounds Emission Factors," Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, (2001)
 14. Zheng, J., H.C. Frey, "Quantitative Analysis of Variability and Uncertainty in Emission Estimation: An Illustration of Methods Using Mixture Distributions," Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, (2001)
 15. Cullen, A.C., and Frey, H.C.; *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, (Plenum Press: New York, 1999).
 16. U.S. EPA, Guiding Principles for Monte Carlo Analysis, EPA/630/R-97/001, U.S. Environmental Protection Agency, Washington, DC, (1997).
 17. Diebolt, J., Robert, C.P., "Estimation of finite, mixture distributions through Bayesian sampling," *J. of the Royal Statistical Society. Series B* **56**, 363-375 (1994).
 18. Harris, C.M. , "On finite mixtures of geometric and negative binomial distributions," *Commun, Statist.-Ther. Meth.* **12**, 987-1007 (1983).
 19. Kanji, G.K., "A mixture model for wind shear data, *J.Appl. Statist.* **12**, 49-58 (1985).
 20. Wedel, M., Desarbo, W. S., Bult, J. R, Ramaswamy, V. .1993. A Latent Class Poisson Regression Model for Heterogeneous Count Data, *Journal of Applied Econometrics*, Vol. 8, No. 4. , 397-411.
 21. Burmaster, D.E. and Wilson A.M., "Fitted second-order finite mixture models to data with many censored values using maximum likelihood estimation," *Risk Analysis* **20**, 235-255 (2000).
 22. Frey, H.C. and Rhodes, D.S., "Characterization and simulation of uncertain frequency distributions: Effects of distribution choice, variability, uncertainty, and Parameter dependence," *Human and Ecological Risk Assessment* **4**, 423-468 (1998).
 23. Titterington, D.M, A.F.M. Smith, and U.E.Makov, *Statistical Analysis of Finite Mixture Distributions*, (John Wiley & Sons, New York, NY, 1985).
 24. Redner, R.A., Walker, H.F., "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, **26**, 195-239 (1984).

25. Everitt, B.S. and D. J. Hand, *Finite Mixture Distributions*, (Chapman & Hall, London, UK,1981)
26. Cassie, R.M.,” Some uses of probability paper in the analysis of size frequency distributions,” *Austral. J. Marine and Freshwater Res.*, 5, pp. 513-523 (1954).
27. Dempster, A.P., Laird, N.M. and Rubin, D.B., “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statist. Soc., Series B*, **39**,1-38 (1977).
28. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery B.P., *Numerical Recipes in FORTRAN*, (Cambridge University Press, New York, NY,1992).
29. Efron,B. and Tibshirani,R.J., *an Introduction to the Bootstrap*, (Chapman & Hall, London, UK,1993).
30. Angus, J.E., “Bootstrap one-sided confidence intervals for the log-normal mean,”*Statistician*, **43**, 395-401(1994).
31. Freedman,D.A., and Peters, S.C. , “ Bootstrapping a regression equation: some empirical results,” *J. of the American Statistical Association*, **79**, 97-106 (1984).
32. Thombs, L.A., Schucany,W.R., “Bootstrap predication intervals for autoregression,” *J. of the American Statistical Association* **85**, 486-492 (1990).
33. Frey, H.C., and J. Zheng; "Quantification of Variability and Uncertainty in Emission Inventories: A Prototype Software Tool with Application to Utility NOx Emissions," Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, (2001)

Table 1. Selected Population Mixture Lognormal Distributions with Two Components

Group No.	μ_1	μ_2	$\sigma_2=\sigma_2$	Group No.	μ_1	μ_2	$\sigma_2=\sigma_2$	Group No.	μ_1	μ_2	$\sigma_2=\sigma_2$
1	1.0	1.1	0.1	5	1.0	1.5	0.5	9	1.0	2.0	1.0
2	1.0	1.2	0.1	6	1.0	2.0	0.5	10	1.0	3.0	1.0
3	1.0	1.4	0.1	7	1.0	3.0	0.5	11	1.0	5.0	1.0
4	1.0	2.0	0.1	8	1.0	6.0	0.5	12	1.0	11.0	1.0

Table 4. Uncertainty of estimated parameters of the fitted mixture lognormal Distribution

Parameter	2.5 th Percentile	Mean	97.5 th Percentile
Weight	0.045	0.236	0.553
μ_1	5.569	5.922	6.233
σ_1	0.006	0.233	0.474
μ_2	6.218	6.259	6.309
σ_2	0.047	0.097	0.190

Table 2. Comparison of 95% Confidence Intervals of Selected Statistics of Single and Two Component Mixture Lognormal Distributions Fitted to Mixture Populations with Varying Component Separation and Standard Deviation for n=100 and w=0.3

Population Parameters ^a				Fitted Dist. ^b	2.5 Percentile PV (CI) ^c	30 Percentile PV (CI) ^c	50 Percentile PV (CI) ^c	75 Percentile PV (CI) ^c	97.5 Percentile PV (CI) ^c	Mean PV (CI) ^c	Standard Deviation PV (CI) ^c		
μ_1	σ_1	μ_2	σ_2										
1.0	0.1	1.1	0.1	Mixture	0.87 (0.82-0.91)	1.01 (0.98-1.04)	1.06 (1.04-1.10)	1.14 (1.11-1.17)	1.28 (1.27-1.34)	1.06 (1.02-1.07)	0.11 (0.10-0.13)		
				Single	0.87 (0.84-0.91)	1.01 (0.98-1.04)	1.06 (1.04-1.09)	1.14 (1.11-1.17)	1.28 (1.24-1.35)	1.06 (1.02-1.07)	0.11 (0.09-1.13)		
		1.2	0.1	Mixture	0.85 (0.82-0.94)	1.07 (1.03-1.11)	1.15 (1.11-1.18)	1.24 (1.20-1.27)	1.39 (1.33-1.44)	1.14 (1.11-1.16)	0.14 (0.12-0.15)		
				Single	0.85 (0.85-0.94)	1.07 (1.03-1.09)	1.15 (1.10-1.16)	1.24 (1.19-1.27)	1.39 (1.36-1.50)	1.14 (1.11-1.17)	0.14 (0.12-0.16)		
		1.4	0.1	Mixture	0.87 (0.81-0.93)	1.07 (1.05-1.30)	1.34 (1.29-1.38)	1.43 (1.40-1.47)	1.58 (1.53-1.64)	1.27 (1.24-1.32)	0.21 (0.18-0.23)		
				Single	0.87 (0.83-0.96)	1.07 (1.10-1.20)	1.34 (1.21-1.31)	1.43 (1.35-1.48)	1.58 (1.64-1.88)	1.27 (1.23-1.32)	0.21 (0.19-0.26)		
		2.0	0.1	Mixture	0.87 (0.82-0.92)	1.77 (1.05-1.88)	1.94 (1.88-1.97)	2.04 (1.99-2.06)	2.18 (2.12-2.22)	1.71 (1.60-1.780)	0.46 (0.42-0.50)		
				Single	0.87 (0.76-0.99)	1.77 (1.25-1.48)	1.94 (1.51-1.75)	2.04 (1.97-2.20)	2.18 (2.68-3.45)	1.71 (1.60-1.82)	0.46 (0.47-0.67)		
		1.0	0.5	1.5	0.5	Mixture	0.44 (0.35-0.640)	1.01 (0.87-1.14)	1.27 (1.13-1.41)	1.63 (1.50-1.82)	2.62 (2.21-3.03)	1.34 (1.23-1.45)	0.55 (0.47-0.64)
						Single	0.44 (0.44-0.64)	1.01 (0.87-1.09)	1.27 (1.11-1.35)	1.63 (1.48-1.83)	2.62 (2.33-3.36)	1.34 (1.16-1.39)	0.55 (0.48-0.77)
2.0	0.5			Mixture	0.43 (0.35-0.62)	1.34 (1.12-1.57)	1.69 (1.57-1.88)	2.13 (1.98-2.31)	3.01 (2.66-3.43)	1.67 (1.58-1.84)	0.68 (0.59-0.78)		
				Single	0.43 (0.46-0.72)	1.34 (1.02-1.31)	1.69 (1.34-1.67)	2.13 (1.84-2.34)	3.01 (3.10-4.70)	1.67 (1.53-1.87)	0.68 (0.68-1.12)		
3.0	0.5			Mixture	0.46 (0.35-0.63)	1.75 (1.12-2.44)	2.67 (2.47-2.85)	3.13 (2.98-3.29)	4.02 (3.65-4.31)	2.35 (2.15-2.58)	1.06 (0.94-1.15)		
				Single	0.46 (0.45-0.82)	1.75 (1.21-1.72)	2.67 (1.76-2.33)	3.13 (2.62-3.63)	4.02 (5.67-8.08)	2.35 (2.14-2.82)	1.06 (1.25-2.22)		
6.0	0.5			Mixture	0.45 (0.36-0.65)	1.88 (1.19-5.45)	5.64 (5.48-5.88)	6.13 (6.03-6.34)	6.93 (6.70-7.24)	4.40 (4.08-4.97)	2.37 (2.09-2.51)		
				Single	0.45 (0.34-0.97)	1.88 (1.57-2.78)	5.64 (2.68-4.25)	6.13 (4.89-7.57)	6.93 (12.3-27.3)	4.40 (4.03-6.22)	2.37 (3.57-8.71)		
1.0	1.0			2.0	1.0	Mixture	0.20 (0.15-0.40)	1.04 (0.84-1.29)	1.48 (1.30-1.76)	2.25 (1.93-2.60)	4.24 (3.42-5.23)	1.68 (1.51-1.91)	1.07 (0.85-1.26)
						Single	0.20 (0.21-0.46)	1.04 (0.74-1.13)	1.48 (1.12-1.60)	2.25 (1.85-2.65)	4.24 (3.99-7.99)	1.68 (1.47-2.09)	1.07 (1.05-2.12)
		3.0	1.0	Mixture	0.20 (0.14-0.37)	1.77 (1.05-2.08)	2.46 (2.07-2.68)	3.26 (2.90-3.55)	5.06 (4.39-6.09)	2.41 (2.12-2.65)	1.33 (1.14-1.53)		
				Single	0.20 (0.22-0.60)	1.77 (0.94-1.57)	2.46 (1.54-2.32)	3.26 (2.65-4.01)	5.06 (6.16-12.7)	2.41 (2.16-3.19)	1.33 (1.70-3.77)		
		5.0	1.0	Mixture	0.25 (0.12-0.33)	3.35 (1.01-3.97)	4.39 (3.95-4.72)	5.26 (4.95-5.62)	7.00 (6.39-7.72)	3.83 (3.37-4.22)	2.04 (1.88-2.35)		
				Single	0.25 (0.23-0.82)	3.25 (1.24-2.37)	4.39 (2.22-3.78)	5.26 (4.25-6.93)	7.00 (10.8-26.5)	3.83 (3.51-5.57)	2.04 (3.12-9.16)		
		11.0	1.0	Mixture	0.23 (0.14-0.37)	8.87 (1.05-9.87)	10.5 (9.85-10.8)	11.4 (11.0-11.7)	12.9 (12.4-13.5)	8.11 (6.94-8.77)	4.63 (4.26-5.06)		
				Single	0.23 (0.25-1.17)	8.87 (1.88-4.74)	10.5 (4.01-8.29)	11.4 (9.14-17.6)	12.9 (30.9-96.7)	8.11 (8.05-17.0)	4.63 (10.5-43.5)		

^a: Footnote to define parameters, refers to the Equation (5) in text.

^b: Fitted mixture is a two component lognormal, single distribution is lognormal.

^c: PV=Population value, CI =95 % confidence interval. Shading indicates that confidence interval does not enclose population value.

Table 3. Comparison of 95% Confidence Intervals of Selected Statistics of Single and Two Component Mixture Lognormal Distributions Fitted to Mixture Populations with Varying Component Separation and Standard Deviation for n=100 and w=0.5

Population Parameters ^a				Fitted Dist. ^b	2.5 Percentile PV (CI) ^c	30 Percentile PV (CI) ^c	50 Percentile PV (CI) ^c	75 Percentile PV (CI) ^c	97.5 Percentile PV (CI) ^c	Mean PV (CI) ^c	Standard Deviation PV (CI) ^c		
μ_1	σ_1	μ_2	σ_2										
1.0	0.1	1.1	0.1	Mixture	0.84 (0.81-0.89)	0.97 (0.94-1.00)	1.04 (1.01-1.07)	1.12 (1.09-1.16)	1.26 (1.22-1.32)	1.04 (1.02-1.07)	0.11(0.10 -0.13)		
				Single	0.84 (0.81-0.88)	0.97 (0.94-0.99)	1.04 (1.01-1.07)	1.12 (1.09-1.15)	1.26 (1.22-1.33)	1.04 (1.02-1.07)	0.11(0.09 -0.13)		
		1.2	0.1	Mixture	0.84 (0.80-0.89)	0.99 (0.95-1.03)	1.10 (1.05-1.14)	1.20 (1.16-1.24)	1.37 (1.31-1.42)	1.10 (1.07-1.12)	0.14(0.13 - 0.16)		
				Single	0.84 (0.80-0.89)	0.99 (0.96-1.03)	1.10 (1.05-1.12)	1.20 (1.15-1.23)	1.37 (1.33-1.47)	1.10 (1.07-1.13)	0.14 (0.12 -0.16)		
		1.4	0.1	Mixture	0.84 (0.80-0.89)	1.00 (0.96-1.04)	1.22 (1.08-1.32)	1.40 (1.34-1.43)	1.56 (1.50-1.60)	1.20 (1.15-1.23)	0.22 (0.20 -0.24)		
				Single	0.84 (0.76-0.87)	1.00 (0.99-1.09)	1.22 (1.13-1.23)	1.40 (1.28-1.41)	1.56 (1.58-1.83)	1.20 (1.16-1.24)	0.22 (0.20 -0.26)		
		2.0	0.1	Mixture	0.85 (0.79-0.88)	0.99 (0.94-1.03)	1.20 (1.07-1.90)	2.00 (1.93-2.03)	2.18 (2.10-2.21)	1.49 (1.38-1.58)	0.51 (0.49-0.53)		
				Single	0.85 (0.61-0.82)	0.99 (1.00-1.21)	1.20 (1.30-1.51)	2.00 (1.63-1.96)	2.18 (2.38-3.19)	1.49 (1.39-1.61)	0.51 (0.44-0.66)		
		1.0	0.5	1.5	0.5	Mixture	0.41 (0.33-0.53)	0.83 (0.73-0.99)	1.19 (1.06-1.35)	1.58 (1.44-1.79)	2.58 (2.18-3.12)	1.26 (1.16-1.39)	0.57 (0.48-0.69)
						Single	0.41 (0.36-0.57)	0.83 (0.73-0.94)	1.19 (1.02-1.27)	1.58 (1.38-1.75)	2.58 (2.24-3.33)	1.26 (1.16-1.39)	0.59 (0.49-0.77)
2.0	0.5			Mixture	0.39 (0.34-0.53)	0.88 (0.72-1.12)	1.48 (1.25-1.66)	1.99 (1.75-2.19)	2.96 (2.55-3.44)	1.49 (1.37-1.62)	0.71 (0.61-0.81)		
				Single	0.39 (0.36-0.58)	0.88 (0.78-1.07)	1.48 (1.16-1.50)	1.99 (1.64-2.17)	2.96 (2.88-4.69)	1.49 (1.36-1.69)	0.71 (0.67-1.13)		
3.0	0.5			Mixture	0.41 (0.31-0.52)	0.87 (0.72-1.11)	2.06 (1.24-2.63)	2.95 (2.69-3.15)	3.89 (3.53-4.23)	1.97 (1.74-2.34)	1.13 (0.99-1.23)		
				Single	0.41 (0.30-0.59)	0.87 (0.82-1.84)	2.06 (1.37-1.91)	2.95 (2.15-2.98)	3.89 (4.45-7.99)	1.97 (1.74-2.35)	1.13 (1.15-2.16)		
6.0	0.5			Mixture	0.37 (0.31-0.52)	0.88 (0.73-1.12)	2.30 (1.31-5.58)	5.99 (5.75-6.19)	6.79 (6.57-7.16)	3.44 (3.05-4.00)	2.55 (2.44-2.64)		
				Single	0.17 (0.16-0.62)	0.88 (0.79-1.69)	2.30 (1.75-3.10)	5.99 (3.50-5.90)	6.79 (9.54-23.6)	3.44 (2.91-4.91)	2.55 (2.81-5.43)		
1.0	1.0			2.0	1.0	Mixture	0.18 (0.12-0.28)	0.71 (0.51-0.93)	1.31 (1.07-1.54)	2.08 (1.74-2.35)	4.32 (3.19-5.28)	1.54 (1.3-1.72)	1.11 (0.85-1.41)
						Single	0.18 (0.15-0.37)	0.71 (0.50-0.87)	1.31 (0.95-1.46)	2.08 (1.64-2.49)	4.32 (3.75-7.95)	1.54 (1.33-1.96)	1.11 (0.97-2.32)
		3.0	1.0	Mixture	0.18 (0.12-0.30)	0.71 (0.50-1.09)	1.89 (1.37-2.37)	2.93 (2.49-3.30)	4.86 (4.12-5.91)	1.97 (1.75-2.27)	1.37 (1.18-1.61)		
				Single	0.18 (0.14-0.42)	0.71 (0.57-1.05)	1.89 (1.13-1.84)	2.93 (2.10-3.36)	4.86 (5.27-11.8)	1.97 (1.75-2.77)	1.37 (1.54-3.84)		
		5.0	1.0	Mixture	0.16 (0.12-0.27)	0.71 (0.49-1.02)	3.55 (1.36-4.16)	5.01 (4.42-5.28)	6.78 (6.01-7.64)	3.06 (2.56-3.44)	2.26 (2.05-2.46)		
				Single	0.16 (0.12-0.49)	0.71 (0.63-1.47)	3.55 (1.44-2.80)	5.01 (3.09-5.80)	6.78 (9.46-26.3)	3.06 (2.66-4.98)	2.26(2.99-10.3)		
		11.0	1.0	Mixture	0.16 (0.11-0.27)	0.69 (0.48-1.06)	8.83 (1.38-10.2)	10.9 (10.5-11.4)	12.7 (12.3-13.5)	6.02 (5.20-7.22)	5.09 (4.88-5.32)		
				Single	0.16 (0.12-0.63)	0.69 (0.90-2.48)	8.83 (2.42-5.51)	10.9 (5.72-13.3)	12.7 (21.1-81.9)	6.02 (5.22-13.2)	5.09 (7.65-43.3)		

^a: Footnote to define parameters, refers to the Equation (5) in text.

^b: Fitted mixture is a two component lognormal, single distribution is lognormal.

^c: PV=Population value, CI =95 % confidence interval. Shading indicates that confidence interval does not enclose population value.

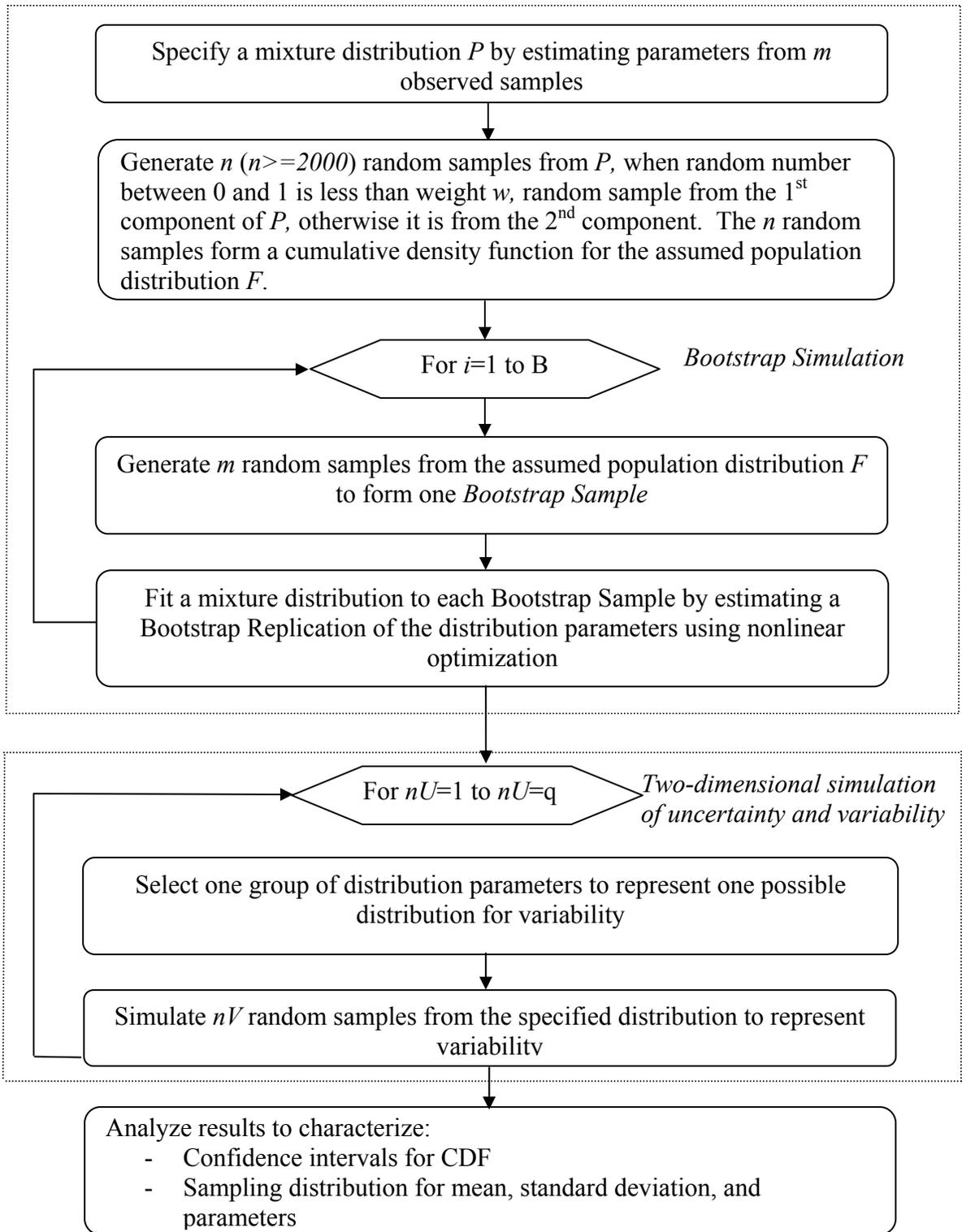


Figure 1. Simplified Flow Diagram for Quantification of Variability and Uncertainty Using Bootstrap Simulation based upon Mixture Distributions

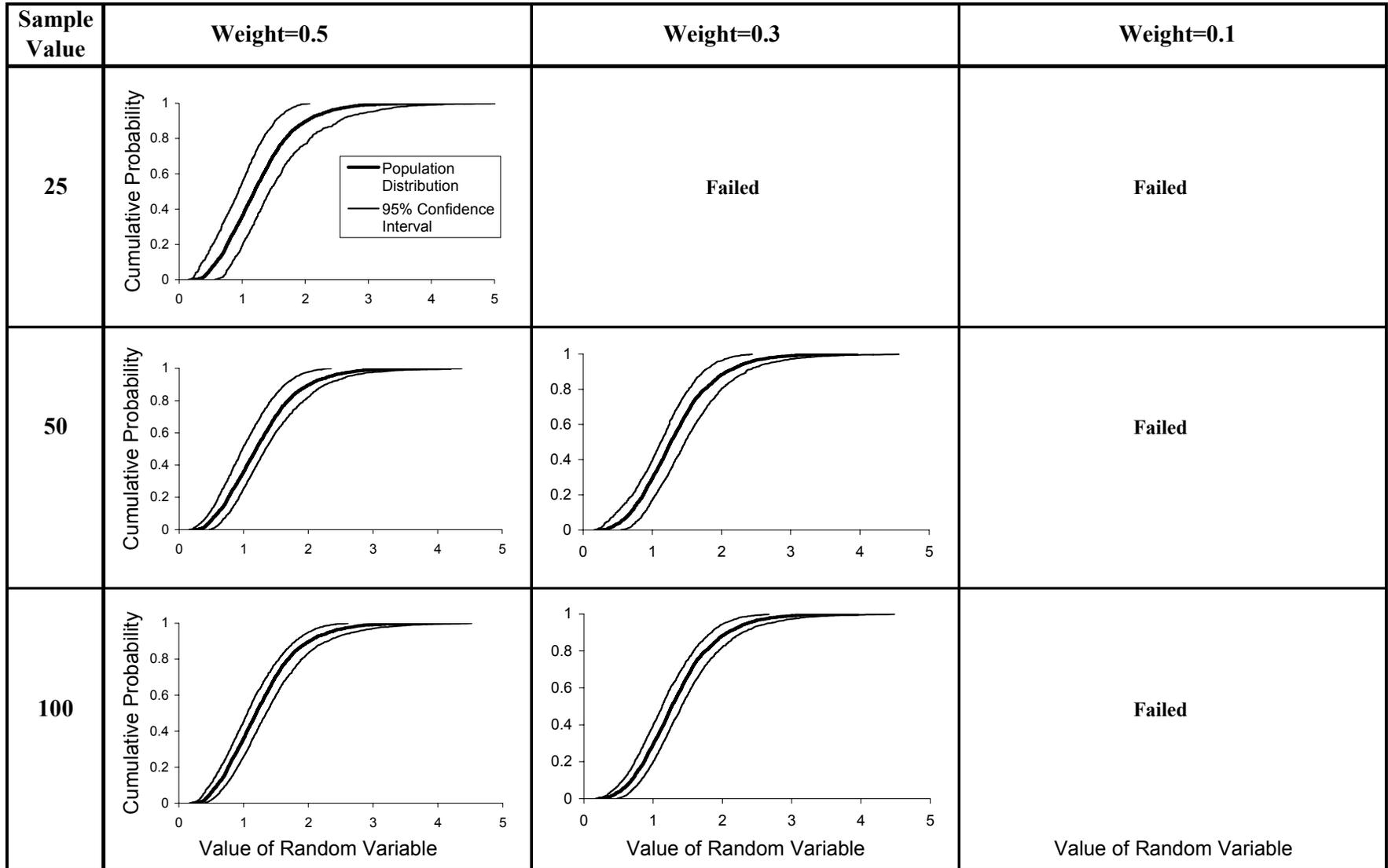


Figure 2. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=1.5$, $\sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$)

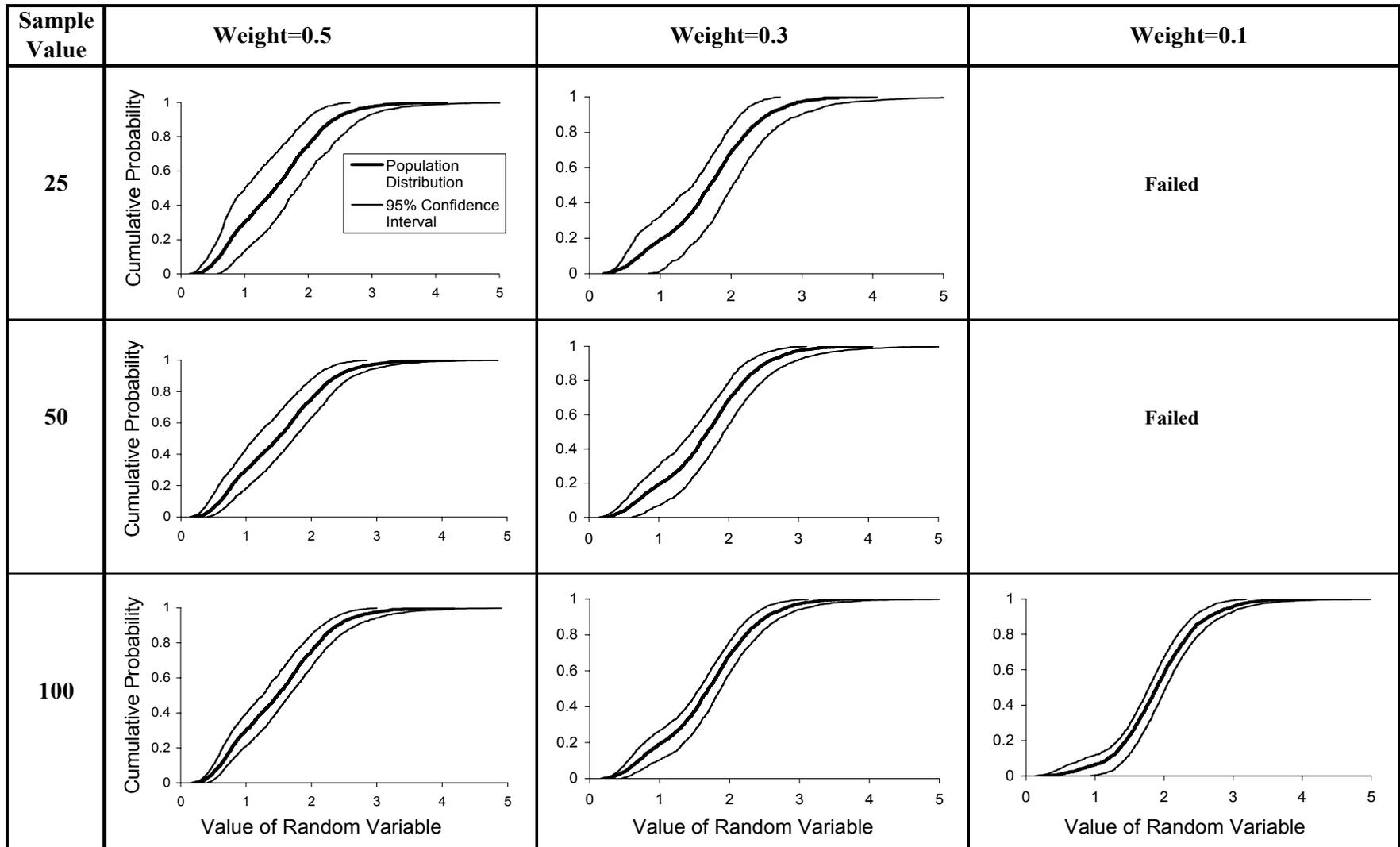


Figure 3. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=2.0$, $\sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$)

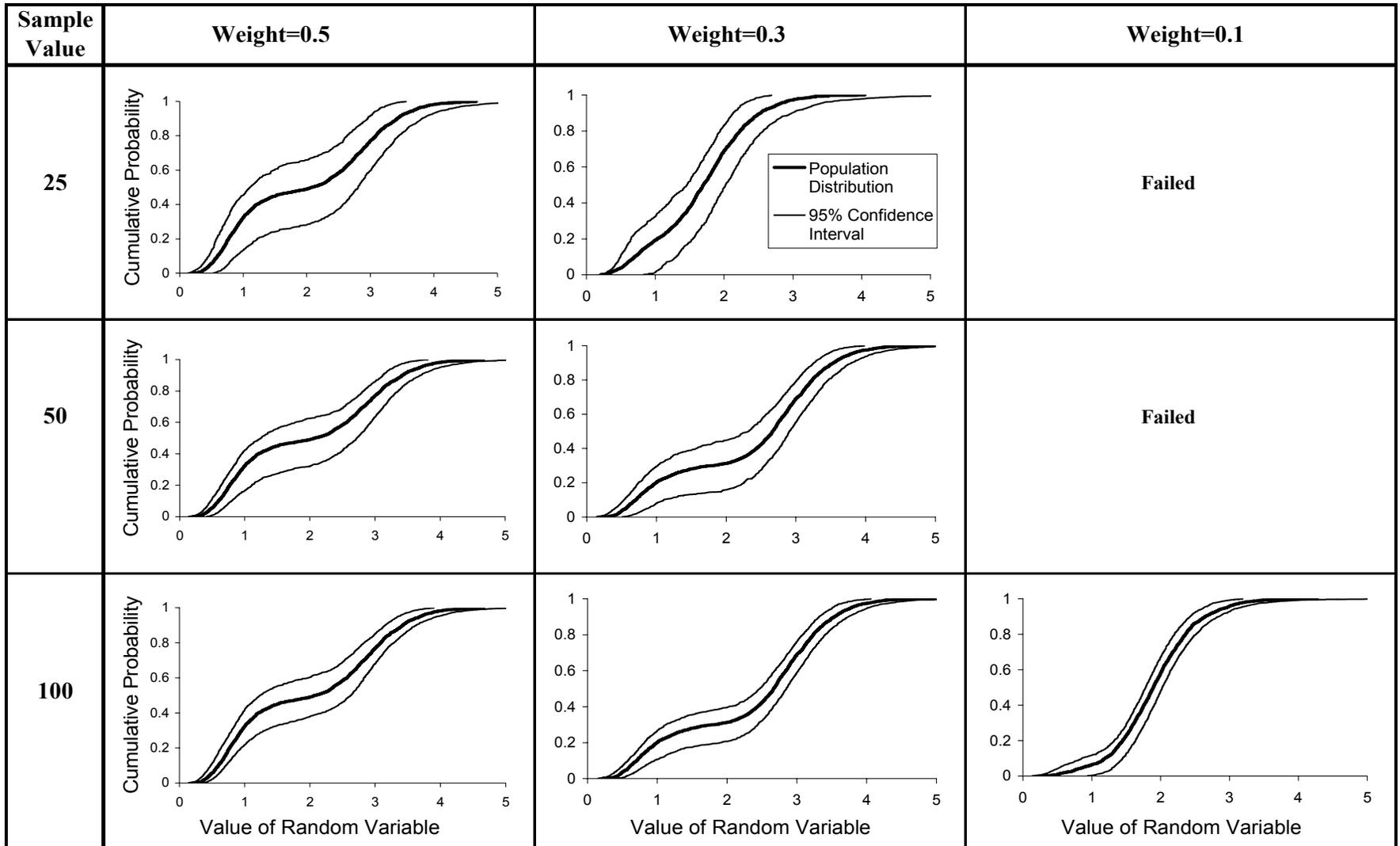


Figure 4. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=3.0$, $\sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$)

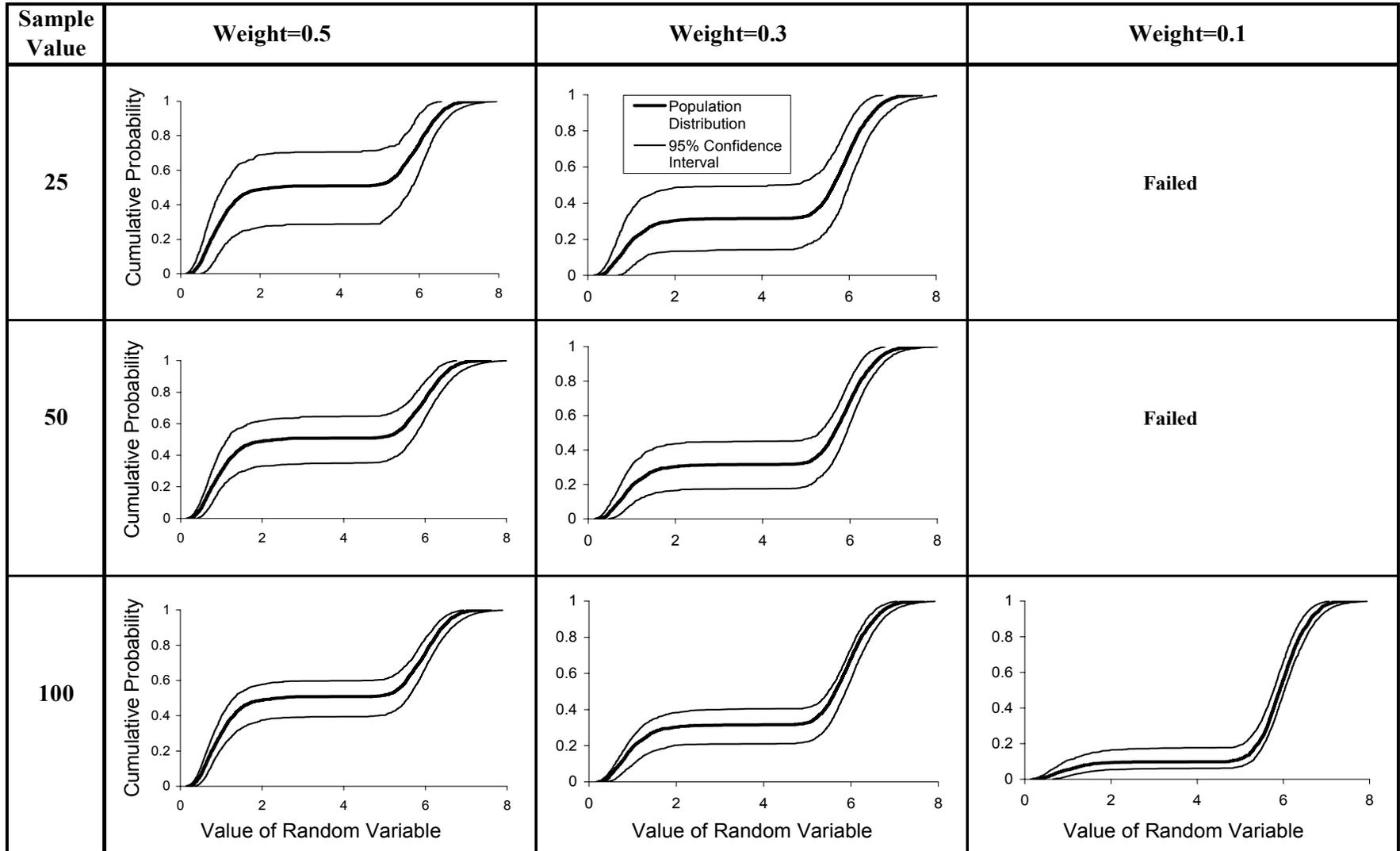


Figure 5. 95 Percent Confidence Intervals of Cumulative Distribution of Two Component Lognormal Distributions Fitted to a Mixture Population Distribution ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=6.0$, $\sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$)

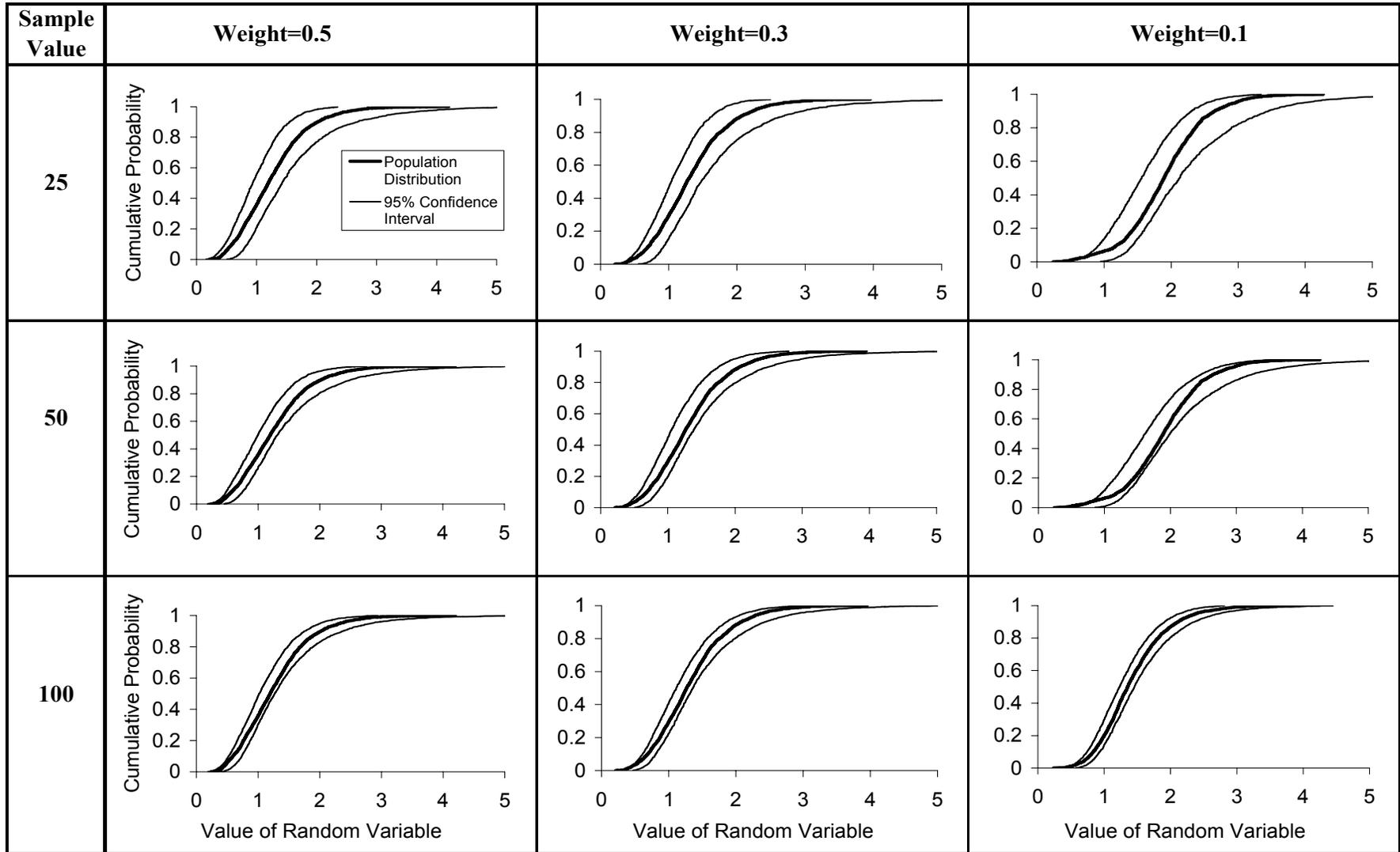


Figure 6. 95 Percent Confidence Intervals of Cumulative Distributions of a Single Lognormal Distribution Fitted to a Mixture Population Distribution ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=1.5$, $\sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$)

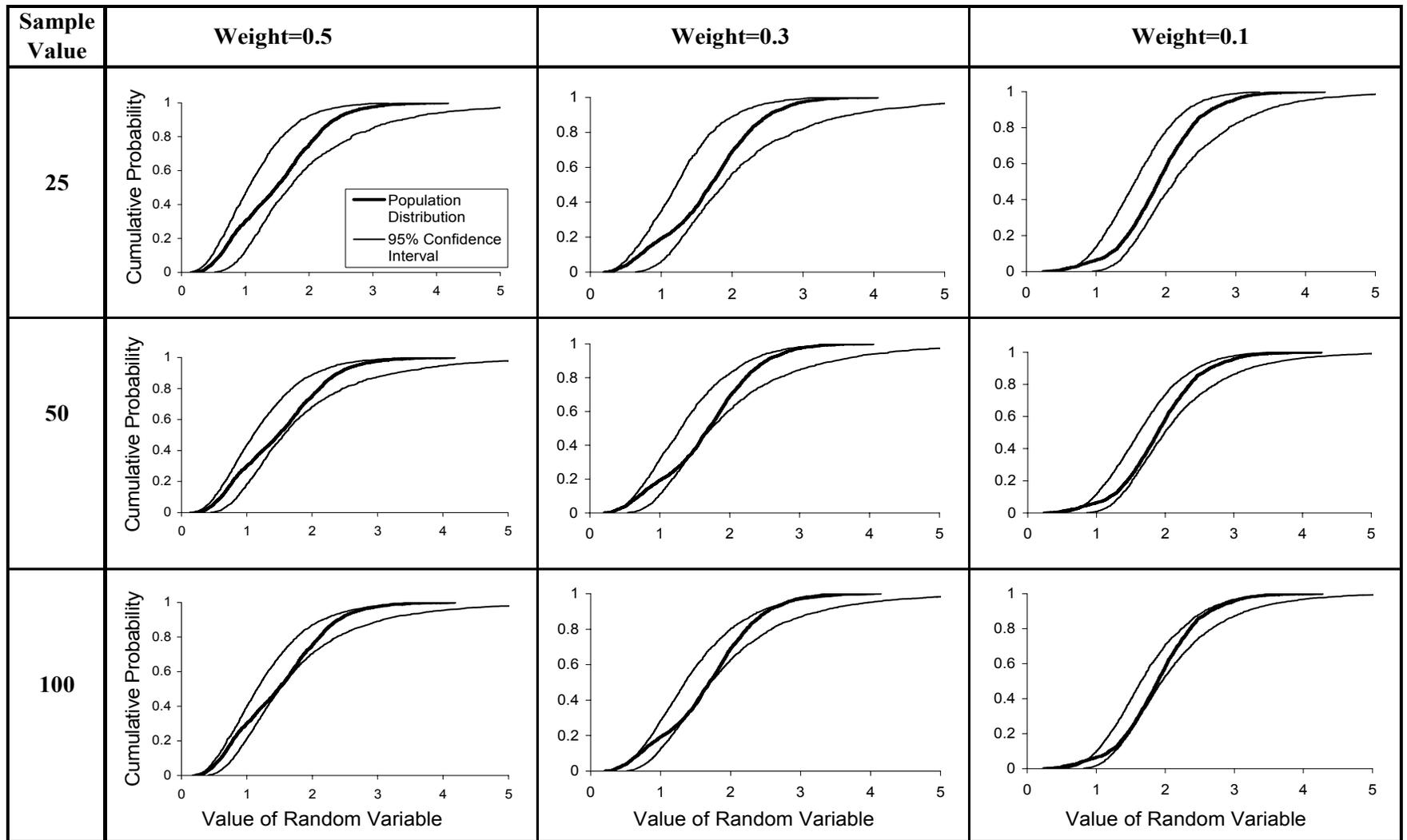


Figure 7. 95 Percent Confidence Intervals of Cumulative Distributions of a Single Lognormal Distribution Fitted to a Mixture Population Distribution ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=2.0$, $\sigma_2=0.5$) for $n=25,50$ and 100 , for $w=0.1,0.3$ and 0.5 Based on Bootstrap Simulation ($B=500$)

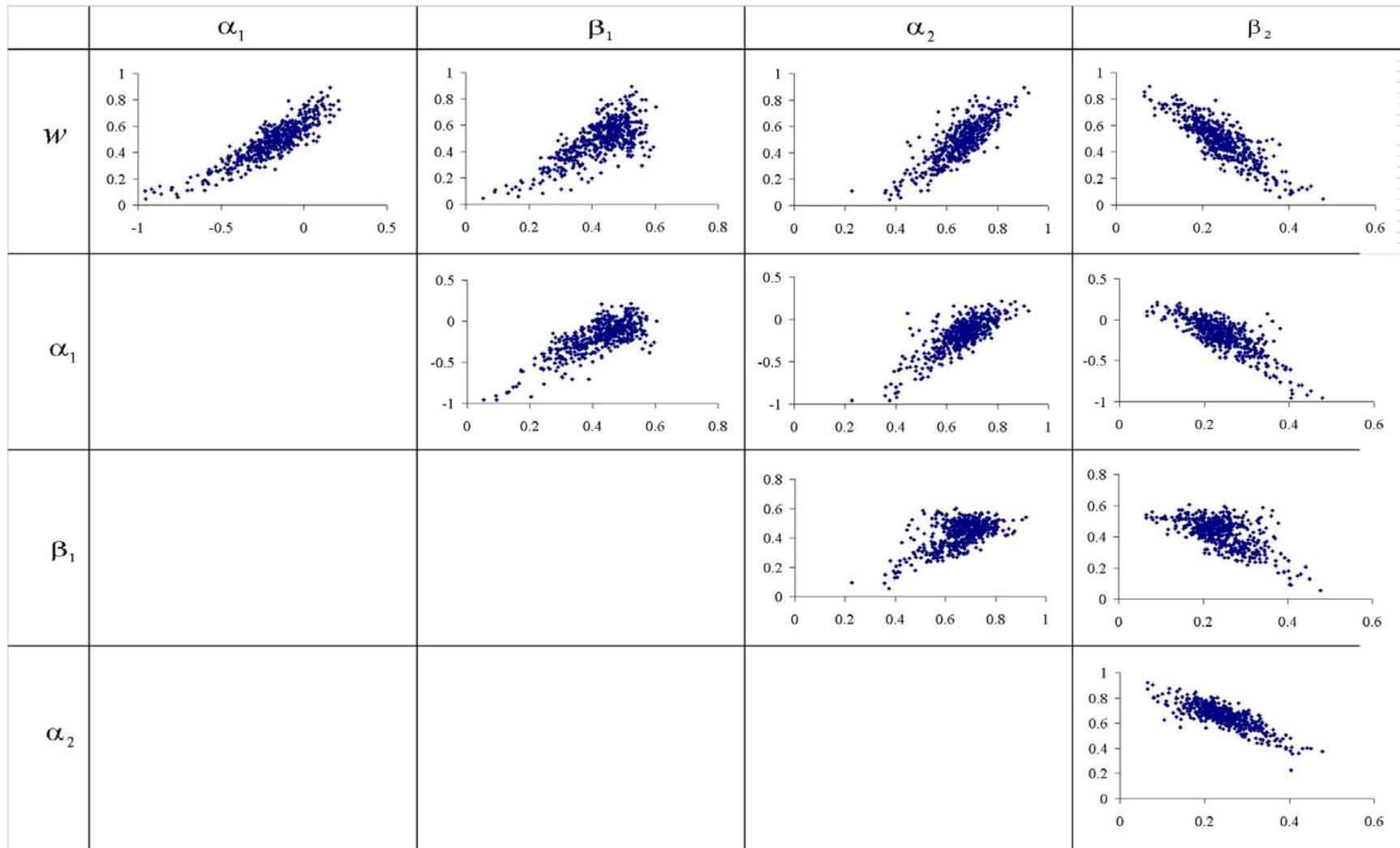


Figure 8. Scatter Plots of Bootstrap Simulation ($B=500$) Results for Parameters of Two Component Lognormal Mixture Distributions for $n=100$, $w=0.5$ with Slightly Separated Components ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=2.0$, $\sigma_2=0.5$).

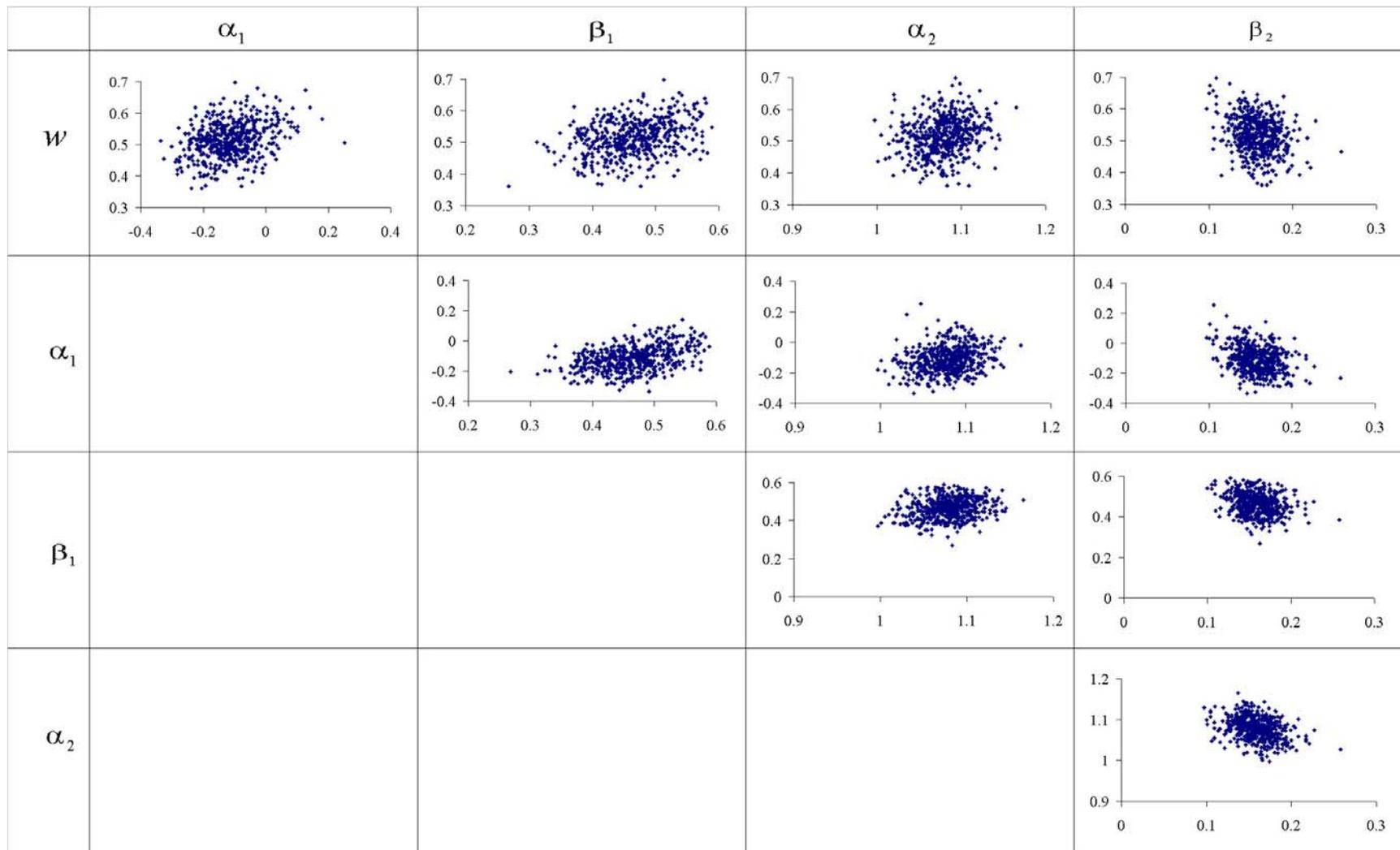


Figure 9. Scatter Plots of Bootstrap Simulation ($B=500$) Results for Parameters of Two Component Lognormal Mixture Distributions for $n=100$, $w=0.5$ with Moderately Separated Components ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=3.0$, $\sigma_2=0.5$).

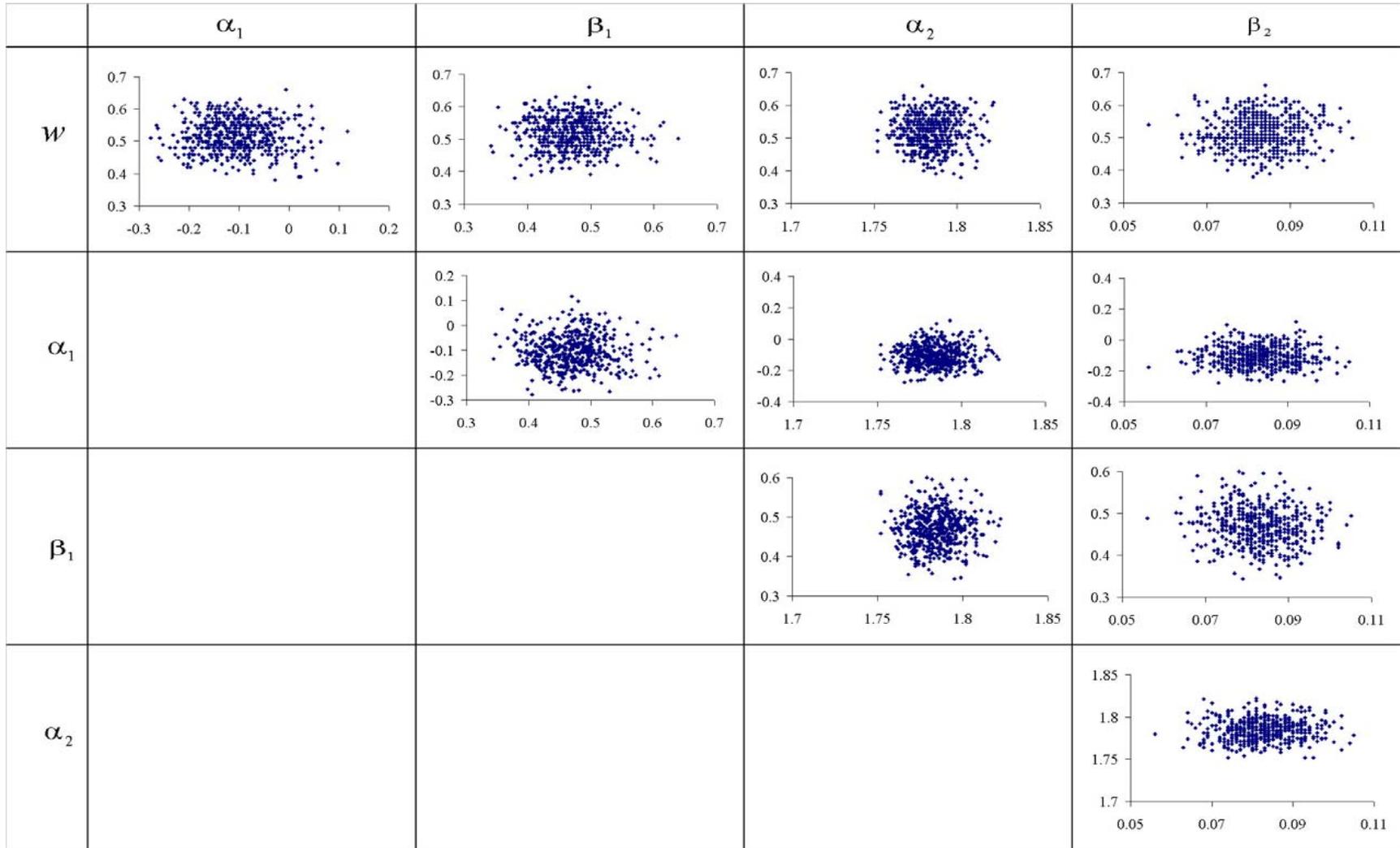


Figure 10. Scatter Plots of Bootstrap Simulation ($B=500$) Results for Parameters of Two Component Lognormal Mixture Distributions for $n=100$, $w=0.5$ with highly Separated Components ($\mu_1=1.0$, $\sigma_1=0.5$, $\mu_2=6.0$, $\sigma_2=0.5$).

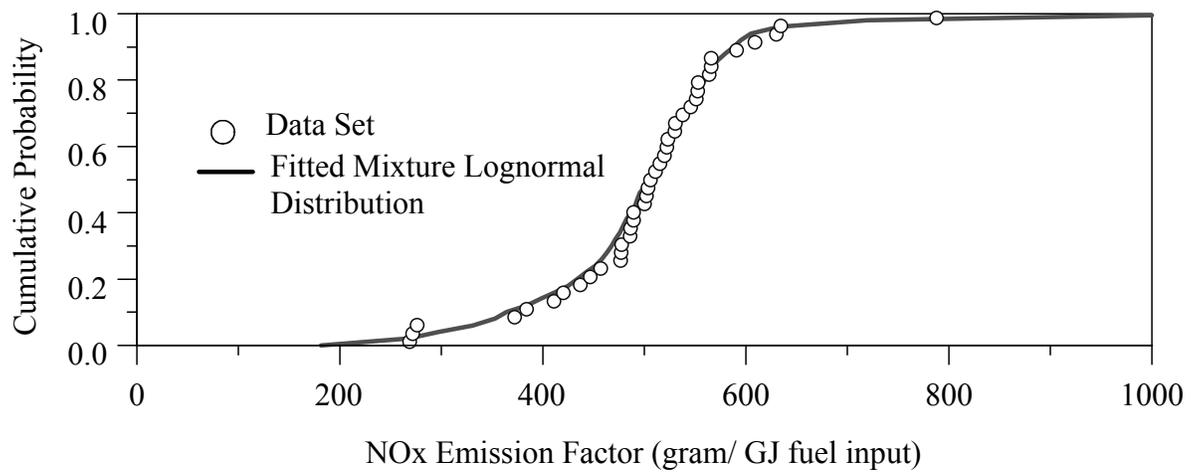


Figure 11. Mixture lognormal distribution fitted to six-month average NO_x emission factor data for T/LNC1 technology group (n=41).

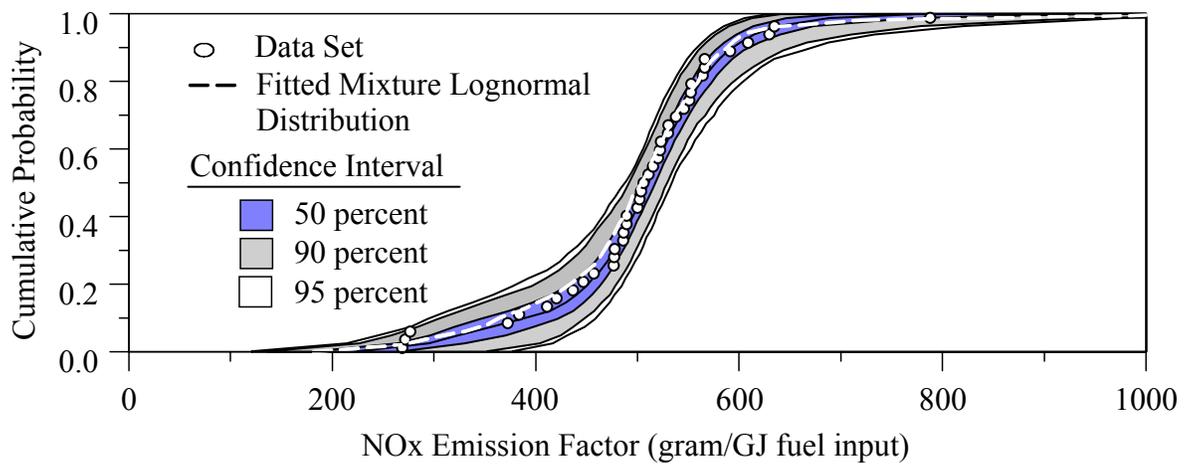


Figure 12. Probability band for fitted mixture lognormal distribution.

PART V

**QUANTIFICATION OF VARIABILITY AND UNCERTAINTY
WITH MEASUREMENT ERROR**

Junyu Zheng and H. Christopher Frey

Prepared to submit to

Risk Analysis

Quantification of Variability and Uncertainty with Measurement Error

Junyu Zheng and H. Christopher Frey
*Department of Civil Engineering, North Carolina State University,
Campus Box 7908, Raleigh, NC 27695-7908, U.S.A.*

ABSTRACT

Variability is the heterogeneity of a quantity over time, space or members of a population. Uncertainty refers to a lack of knowledge about the true value of a quantity; it may arise from measurement error because of biases in the measuring apparatus and the experimental procedures and random sampling error. The appearance of measurement error potentially affects any statistical analysis because the distribution representing the observed data set possibly deviates from the distribution which would have generated an error free data set. A methodology for improving the characterization of variability and uncertainty with known measurement errors in environmental data is demonstrated in this paper. The method for constructing an error free data set based on the observed data set, known measurement error and measurement error models is developed. Numerical methods based upon bootstrap pair techniques were applied to characterize uncertainty in statistics for the measurement error problem. The effect of measurement error on quantification of variability and uncertainty was evaluated and investigated with respect to the size of measurement error. The results indicate that when measurement error is a main source of uncertainty, substantial difference between the distribution representing variability of the observed dataset and the distribution representing variability of the error free data set will appear and variability will be underestimated; and that the shape and

range of the probability band based upon the observed data set is largely different from the ones for the error free data set. It suggests that ignorance of uncertainty from measurement error will underestimate total uncertainty and failing to separately characterize contributions from random sampling error and measurement error will bring bias in the variability and uncertainty estimates. An approach presented in this paper, in which the contribution from measurement error and random sampling error to uncertainty are separately characterized, is suggested for use in quantification of variability and uncertainty of a quantity if there are known measurement errors.

KEY WORDS: Measurement error; Bootstrap pair; Variability; Uncertainty; Observed data; Error free data.

1.0 INTRODUCTION

In exposure or risk assessment and emission estimation, deterministic analysis method incorporating the use of point estimates such as average or conservative upper bound assumptions for model inputs has often been used. The problem with this approach is that it provides no indication of the extent to which the conservative assumptions overestimate actual risks or emissions and of where to focus research to reduce uncertainty in the estimates. ⁽¹⁾ Limitations of the deterministic approach are detailed elsewhere. ^(2, 3, 4, 5, 6)

In recent years there has been attention to and use of probabilistic approaches, in which variability and uncertainty in model inputs are explicitly represented and are propagated through the model. With quantitative information regarding both the variability and uncertainty of exposures or risk and emissions, decision-makers can assess

whether a particular decision is likely to be robust to variability and and/or incomplete knowledge. ⁽¹⁾

1.1 Variability and Uncertainty

Variability refers to the heterogeneity of values with respect to time, space, or a population. For example, there exists variability in emissions across individual units in the utility emission source category. ⁽⁷⁾ Variability is an inherent property of a system under study and therefore is irreducible, even in principle for a given population. ⁽²⁾ Variability can be represented by a frequency distribution showing the variation in a characteristic of interest over time, space. ⁽⁸⁾

Uncertainty arises due to lack of knowledge regarding the true value of a quantity. For example, there may be uncertainty regarding the true average emission rate for a source category in emission estimation. Uncertainty can be attributed to random errors, human error and measurement error because of biases in the measuring apparatus and the experimental procedures. Uncertainty can be quantified as a probability distribution representing the likelihood that the unknown quantity falls within a given range of values. ⁽⁸⁾

1.2 Limitations of Current Studies in Variability and Uncertainty Analysis

The use of quantitative methods for dealing with variability and uncertainty is becoming more widely recognized and recommended for environmental modeling and assessment applications. For example, the National Research Council has recommended that the distinction between variability and uncertainty should be maintained rigorously at the level of individual components of a risk assessment as well as at the level of an integrated risk assessment. ⁽⁹⁾ There have been a number of efforts aimed at probabilistic analysis of various other emission sources, including power plants, non-road mobile

sources, and natural gas-fired engines.^(10, 11, 12, 13, 14) However, in these studies and applications, more attention is focused on the quantification of variability and uncertainty arising from random sampling error, while measurement error, another source of uncertainty, and its effect on variability and uncertainty analysis was not characterized. Chesher pointed out that measurement error potentially affects all statistical analysis, both formal and informal because it causes the probability distribution that generates the observed data to deviate from that which generates unobservable, error free data that bear directly on the model with which an analyst works.⁽¹⁵⁾ This implies that ignorance of the effect of measurement error may bring bias in the estimates of variability and uncertainty.

1.3 Measurement Error and Uncertainty

Measurement is the process of finding the value of a physical quantity experimentally with the help of measuring instruments. The true value of a measurable quantity is the value of the measured physical quantity, which being unknown, would ideally reflect, both qualitatively and quantitatively, the corresponding property of an object.^(16, 17) Because it is inevitable that there exists error from measurement instruments and procedures, any observed measurement results are actually composed of the true value of the measurable quantity and measurement error caused by the measurement instruments and measurement procedures. Therefore, measurement error can be defined as the deviation of the result of measurement from the true value of the measurable quantity.⁽¹⁷⁾ Quantitatively the measurement inaccuracy can be characterized by the notion of either limits of error or uncertainty. Uncertainty of measurement is an interval within which a true value of a measurement lies within a given probability.⁽¹⁶⁾

1.4 Classification of Measurement Errors

Based on the cause of errors, measurement error can be classified as methodological error, instrument error and personal error.⁽¹⁶⁾ The methodological errors can arise as a result of an inadequate theory of the phenomena on which the measurement is based and inaccuracy of the relations that are employed to find an estimate of the measurable quantity. Instrumental measurement errors are caused by the imperfection of measurement instruments. It is often referred to as the intrinsic error of measuring instruments. Personal error often depends on the individual characteristics of the person performing the measurement instrument. It can include the errors owing to the incorrect reading, recording. Personal errors are usually insignificant.⁽¹⁶⁾ Personal error can be minimized by appropriate Quality Assurance (QA) / Quality Checking (QC).

Another important classification of measurement errors is based on their properties. If the component of the inaccuracy of measurement remains constant or varies in the course of a number of measurements of the same measurand,⁽¹⁸⁾ or if the inaccuracy arises from consistent and repeatable sources (like an offset in calibration),⁽¹⁹⁾ the measurement errors are said to be "systematic" or "bias" errors. The observed and estimated systematic error is often eliminated from measurements by introducing corrections.⁽¹⁶⁾

If errors arise from random fluctuations in the measurement⁽¹⁹⁾ or if there are differences between the results of single measurements, and these differences cannot be predicted individually and any regularity inherent to them are manifested only in a significant number of results,⁽¹⁸⁾ the error owing to this scatter of the results is called "random" or, occasionally and incorrectly, "statistical" errors.⁽¹⁹⁾ Random errors are

reduced when an experiment is repeated many times and the results averaged together.⁽¹⁹⁾ Random error is often assumed as normal distribution with mean of 0.⁽¹⁶⁾

Systematic and random error must be treated differently in variability and uncertainty analysis due to their different properties. In this study, we assume that systematic error relevant to instrument error and personal error has been corrected during the process of data collection. Therefore, the component of measurement error addressed in this study is random error, which can be expressed in absolute form and can be assumed as a normal distribution with mean of 0.

1.5 Purpose of This Study

The purpose of this study includes: (1) introducing a measurement error model used to describe the measurement error problem; (2) developing methods for constructing an error free data set based upon the observed dataset and known measurement errors; (3) developing methods for quantifying variability and uncertainty in a data set with the inclusion of measurement errors; (4) evaluating the effect of measurement error size on the results of variability and uncertainty analysis; and (5) presenting suggestions and recommendations for variability and uncertainty analysis if there are known measurement errors in a dataset.

2.0 METHODOLOGY

A fundamental prerequisite for analyzing a measurement error problem is the specification of a model describing measurement error. Therefore, this section will begin with the discussion of a measurement error model often used for measurement error problems. The method for constructing error free dataset and the method for

quantifying variability and uncertainty with the consideration of measurement error will be presented in the subsequent subsections.

2.1 Measurement Error Models

There are two kinds of often-used measurement error models, which are an additive error model and a multiplicative model, respectively. An additive error model, which is known as a classical error model, is often used to address the measurement error problem in which measurement error is expressed in the form of range of absolute errors. A multiplicative model, which is a variation of the additive error model, can deal with the situation where measurement error is described by the relative error form.^(20, 21) This paper focuses on the use of an additive error model to describe the measurement error problem. An additive error model can be expressed as:

$$z_i = X_i + e_i \quad (1)$$

Where,

z_i = Error contaminated data or observed data

X_i = Error free data or true value (which is not known)

e_i = Measurement error, which is often assumed as a normal distributed model with mean of 0, and is independent of z_i and X_i .

Based upon these assumptions, we have $E(e_i) = 0$, therefore

$$E(z_i) = E(X_i) \quad (2)$$

$$\text{Var}(z_i) = \text{Var}(X_i) + \text{Var}(e_i) \quad (3)$$

Where,

$E(\bullet)$ = Mathematical expectation of a variable

$\text{Var}(\bullet)$ = Variance of a variable

2.2 Error Free Data Construction

The purpose of the error free data construction is to find the density function of the error free data through a classical additive error model when the density functions of the observed data and measurement error are known. It is suggested that the density function f_z of the observed data is the convolution of the density functions of error free data and measurement error, ⁽²⁰⁾ which can be expressed as the following form:

$$f_z(z) = \int f_x(x) f_e(z-x) dx \quad (4)$$

The problem of finding $f_x(x)$ in the absence of parametric assumptions is a deconvolution problem. $f_e(e)$ is often assumed as a normal distribution, and it is typically possible to find a density function for $f_z(z)$ by using common distribution models to fit the observed data. Thus, $f_x(x)$ can be recovered by Fourier inversion. Letting $\phi_a(t)$ denote the characteristic function of the random variable A, we have:

$$f_x(x) = \frac{1}{2\pi} \int e^{-itz} f_x(x) dx = \frac{1}{2\pi} \int e^{-itz} \frac{\phi_z(t)}{\phi_e(t)} dt \quad (5)$$

For example:

$$\phi_z(t) = \int e^{itz} f_z(z) dz \quad (6)$$

Because the work of finding $f_x(x)$ involves complicated mathematical inferences and computations, many researchers have sought simplified solutions for the deconvolution problem ^(22, 23, 24) Ideally, an analytic density function $f_x(x)$ may be a more accurate description of the variation of a measurable quantity; however, there is a high possibility that the density function $f_x(x)$ will not be found to be a common probability distribution model when using the deconvolution approach, or the analytic density function may not be available. If the unknown density function for error free data is not

one of the common probability models (e.g., normal, lognormal, gamma or Weibull distribution), the deconvolution approach may possibly bring more difficulties in characterizing variability and uncertainty. For example, the sampling algorithm for the unknown density function may not be available and may need to be developed, which may lead to more work.

If measurement error is known and the observed dataset (or error contaminated dataset) is available, an alternative method can be used to construct an error free dataset. In the alternative approach, it is assumed that an estimate of a true value (error free data) of a quantity is a linear combination of the sample mean of the error-contaminated dataset and an observed data point. ⁽²⁰⁾

$$\hat{X}_i = cz_i + (1 - c)\bar{z} \quad (7)$$

Where,

\hat{X}_i = Estimated error free data

\bar{z} = Sample mean of error contaminated or observed data

c = Variance adjusting constant, $0 \leq c \leq 1$

If $c=1$, no measurement error exists. In this case, an estimate of the true value of each measured data point is its corresponding observed data. If $c=0$, the measurement error is so large that a true value cannot be estimated based on the corresponding observed data. In this situation, a best estimate of the true value is the sample mean of the observed data.

Based on the Quasilikelihood and Variance Function (QVF) models described by Carrol *et al.* ⁽²⁰⁾ and the statistical assumption that the sample variance of error-

contaminated data is the sum of the sample variance of error free data and the variance of measurement error as shown in Equation (3), the constant c can be found as follows:

$$c = \sqrt{\frac{S_z^2 - \sigma_e^2}{S_z^2}} \quad (S_z^2 > \sigma_e^2) \quad (8)$$

Where:

S_z^2 = The sampling variance for the observed data set

σ_e^2 = The variance for the measurement error

Once an error free dataset is constructed, the methodologies previously developed for quantification of variability and uncertainty in the dataset without measurement error can be applied to the measurement error problems. ^(10, 11, 25)

The advantage of using this approach to construct error free data set is that it is easy to implement. However, it will fail if the sample variance of the observed data is less than that of the measurement error.

2.3 Quantification of Variability and Uncertainty with Measurement Error

Bootstrap simulation, presented by Efron in 1979, ⁽²⁶⁾ is a numerical technique originally developed for the purpose of estimating confidence interval for statistics. This method has an advantage over analytical methods in that it can provide solutions for confidence intervals in situations where exact analytical solutions may be unavailable and in which approximate analytical solutions are inadequate. Bootstrap simulation has been widely used in the uncertainty analysis in emission estimation and risk assessment. ^(10, 11, 27) In these applications, bootstrap simulation is based on a one-sample model in which a random sample is generated from a distribution representing a variable. A one-sample model is the simplest and often used one in bootstrap simulation. However, the

probability model for the error free data is a convolution of probability distributions of error contaminated data and measurement error; thus, this is a two-sample problem. Efron suggested that “Bootstrap pairs” could be used to deal with a two-sample problem. “Bootstrap pairs” may also be based on either nonparametric or parametric approaches or a combination of the two ⁽²⁸⁾.

For a measurement error problem, a two-sample problem can be constructed as:

$$\mathbf{Z}=(\mathbf{x}, \mathbf{e}) \quad (9)$$

Let \mathbf{x} denote a probability distribution model for error free data, \mathbf{z} denote a distribution for the observed data with known measurement error, and \mathbf{e} indicate the probability model for measurement error. \mathbf{x} is often described using an empirical distribution or a common parametric probability distribution model based on the error free data constructed by using the approach presented in Section 2, and \mathbf{e} is a normal distribution with a mean of 0 based on measurement error.

Each bootstrap sample \mathbf{z}^* can be computed based upon the classical additive model introduced in the Section 2.1:

$$\mathbf{Z}_i^*=(\mathbf{x}_i^*, \mathbf{e}_i^*) = x_{i,j} + e_{i,j} \quad (i=1, 2, \dots B), (j=1, 2, \dots n) \quad (10)$$

Where,

i = Subscript for bootstrap samples

j = Subscript for data point

B = The number of bootstrap samples

n = The sample size of a dataset

$x_{i,j}$ = A random sample from a distribution describing error free data

$e_{i,j}$ = A random sample from a distribution describing measurement error

The bootstrap replications of θ^* for the mean can be computed as: ⁽²⁸⁾

$$\hat{\theta}_{i(\hat{z})}^* = \bar{x}_i^* + \bar{e}_i^* \quad i = 1, 2, \dots, B \quad (11)$$

Where,

$\hat{\theta}_{i(\hat{z})}^*$ = The estimated bootstrap mean for the i^{th} bootstrap sample

For the standard deviation: ⁽²⁸⁾

$$SE(\hat{\theta}_{i(\hat{z})}^*) = \sqrt{\text{Var}(\bar{x}_i^* + \bar{e}_i^*)} = \sqrt{\text{Var}(\bar{x}_i^*) + \text{Var}(\bar{e}_i^*)} \quad i = 1, 2, \dots, B \quad (12)$$

Where,

$SE(\hat{\theta}_{i(\hat{z})}^*)$ = The standard deviation of the estimated bootstrap means

$$\text{Var}(\bar{x}_i^*) = \frac{\sigma_{x_i}^2}{n}$$

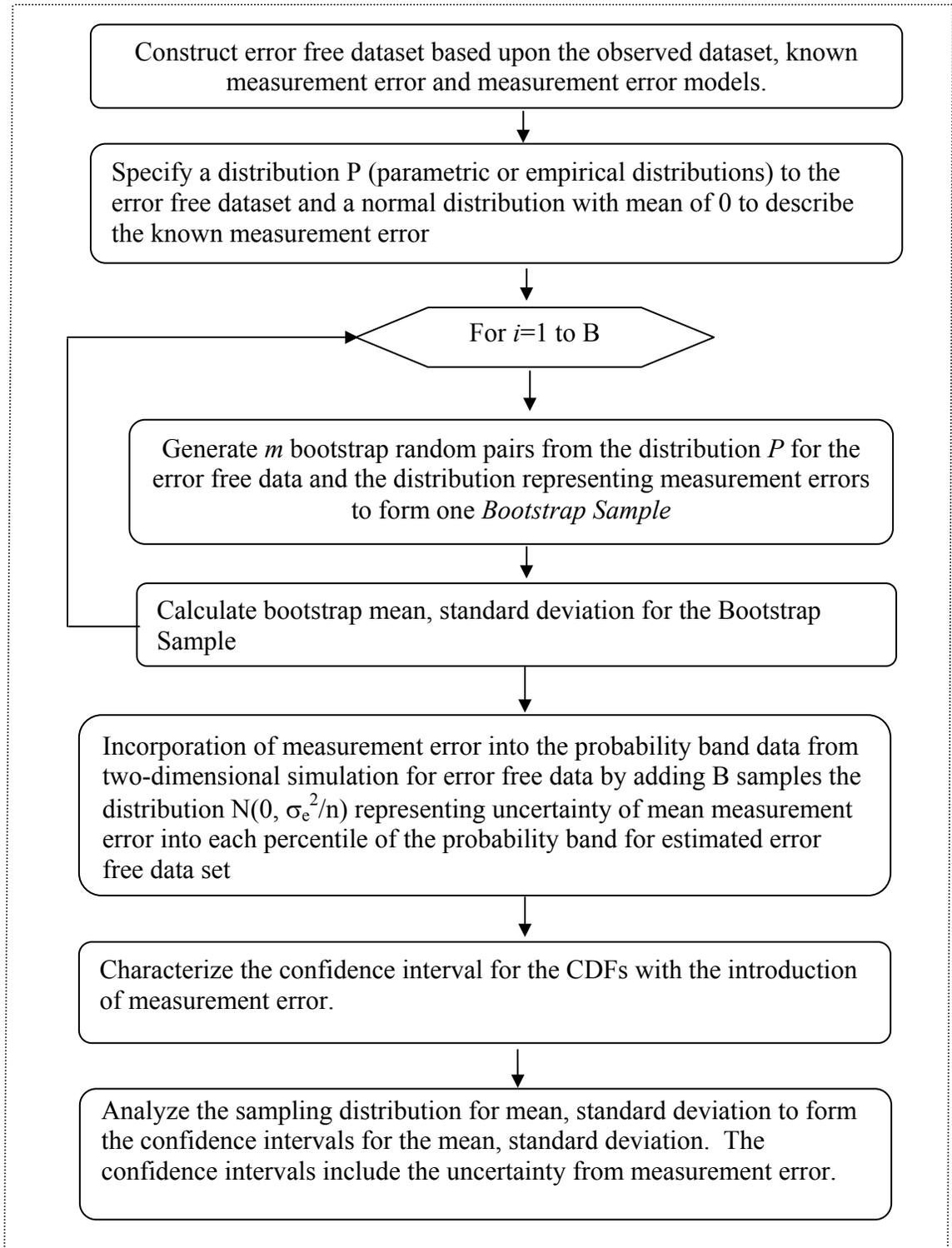
$$\text{Var}(\bar{e}_i^*) = \frac{\sigma_{e_i}^2}{n}$$

The variance of \bar{x}_i^* and \bar{e}_i^* are calculated based upon estimators reported by Casella and Berger. ⁽²⁹⁾

A two-dimensional probabilistic modeling framework for simultaneously quantifying variability and uncertainty in the context of the use of a single component distribution featuring the use of bootstrap simulation, presented by Frey and Rhodes, ^(10, 11) can be extended to deal with a measurement error problem using bootstrap pairs. A flow diagram for characterizing variability and uncertainty with measurement error featuring the use of bootstrap pairs is shown in Figure 1.

As shown in Figure 1, the first step is to construct an error free dataset based upon the observed dataset, known measurement error and measurement error model. A

Figure 1. Flow Diagram for Characterizing Variability and Uncertainty with Measurement Error by Using the Bootstrap Pair Technique



distribution P is specified for the error free dataset and a normal distribution with a mean of 0 is specified for measurement error. The distribution P reflects the best estimate of the true variability in the measured quantity.

The bootstrap pair technique as shown in Equations (10) through (12) is introduced to analyze the uncertainty in the statistics such as the mean. Each bootstrap pair is formed by randomly generating a sample from the distribution P representing error free data and a sample from the distribution $N(0, \sigma_e)$ representing measurement error. The bootstrap pairs form a bootstrap sample from which the bootstrap replicates of the mean or other statistics may be estimated. The sampling distributions of these statistics are the basis for estimating confidence intervals for these statistics.

In order to quantify uncertainty in the CDF of the error-free distribution for variability, a two-dimensional simulation is done to quantify uncertainty based upon the distribution P for the error free data. The results to this point only characterize the uncertainty arising from the random sampling errors. Therefore, it is necessary to incorporate the uncertainty from measurement error into the uncertainty analysis.

That the uncertainty associated with measurement error is incorporated is based upon the assumption that the variance in the sampling distribution of the i^{th} percentile of variability is the sum of the variance from random sampling errors at the i^{th} percentile for the error free dataset and the variance from the mean measurement error. ⁽²⁹⁾ Therefore, B samples are generated from the distribution $N(0, \sigma_e^2/n)$ representing uncertainty of the mean measurement error. Then the B samples are added into each percentile of the probability band for the error free data set. Such combinations will make the probability band include the uncertainty arising from the measurement error.

The methods of constructing confidence intervals based on bootstrap replications include the percentile method, bootstrap-t method and BC_a method. ⁽²⁸⁾ In this study, the percentile method is used because of its simplicity and wide application in practice.

3.0 CASE STUDY

To demonstrate the use of the methods for quantifying variability and uncertainty with measurement error presented in Section 2 and to investigate the effect of the size of measurement errors on the variability and uncertainty estimates, a synthetic dataset with mean of 107.3, standard deviation of 60.9 and sample size of 25 was generated to represent an observed data set. Let **r** represent the ratio of the standard deviation of the measurement error of the standard deviation of the observed data:

$$r = \frac{\sigma_e}{\sigma_{\text{Total}}} \quad (13)$$

Where,

σ_e = Standard deviation of measurement error

σ_{Total} = Standard deviation of the observed error-contaminated data set

In the case study, the standard deviation of the measurement error is assumed to be 10.0, 20.0, 40.0 and 55.0, respectively; therefore, the corresponding values of **r** are 0.16, 0.33, 0.66 and 0.90. The first two cases represent a situation where measurement error is not a main source of uncertainty; the last two cases represent a situation where measurement error is a main source of uncertainty. The measurement errors are assumed to follow normal distribution models with a mean of 0.0 and the specified standard deviation.

Table 1 lists the observed data set and estimated error free data sets under the use of different measurement error models. These error free data sets are estimated based upon Equation (7). Figure 2 displays the lognormal distributions fitted to the error free data under the use of different measurement error model and the lognormal distribution fitted to the observed data set. Table 2 shows the parameter estimation and goodness-of-fit results for the fitted lognormal distributions. All of the fitted distributions passed the Kolmogorov-Smirnov (K-S) test of goodness-of-fit, and thereby they are good representatives of the corresponding data sets. When the ratio $r < 0.5$, there are no substantial observable differences between the fitted distributions to the error free data and the fitted distribution to the observed data. However, when the ratio $r > 0.5$, the shapes and ranges of the distributions representing the error free data are substantially different from the distribution representing the observed data. For example, the 95% probability range of the fitted distributions for the error free data is from 45 to 215 for $\sigma_e = 40.0$, from 65 to 162 for $\sigma_e = 55.0$, and from 34 to 254 for the observed data set. These examples illustrate that the distribution representing the observed data set substantially deviates from the distribution for the true variability of a quantity due to the effect of measurement error. It implies that variability will be overestimated if measurement error is large. Therefore, in such situations, the distribution for the observed dataset cannot be used to directly describe the variability of a quantity, especially when measurement error is a main source of uncertainty.

To investigate the effect of measurement error on the shape and range of probability bands from two-dimensional simulations, two-dimensional simulations were done for the observed data set and error free data sets with and without consideration of

measurement error. Figure 3 displays the probability band for the observed data set. In this case, it is assumed that no measurement error exists in the observed dataset. In Figure 3 and the following probability band graphs, the empirical distribution plotted with open circles represents the observed dataset, and the fitted distributions to the observed dataset or error free datasets is drawn with a dashed line.

Figure 4 shows the uncertainty due to random sampling error in the CDF of the fitted distribution for error free data in the case of an assumed error of $N(0,10)$. Figure 5 shows the uncertainty due to both random sampling error and measurement error for the same error free distribution. There is no perceptible difference in these two results, which indicates that the measurement error is very small in this case. Similarly, a comparison of Figures 6 and 7 reveals no perceptible difference in the width of the confidence intervals when only random sampling error is considered compared to when both random sampling and measurement error are considered, for the case of a measurement error of $N(0,20)$. The results in Figures 4 through 7 indicate that a measurement error as large as $N(0,20)$ does not contribute substantially to total uncertainty. Furthermore, Figures 4 and 6, which are based only upon random sampling error for the fitted distribution of the error free data set, are not substantially different from Figure 3, which is based upon the distribution of error-contaminated data. This comparison again suggests that a measurement error of $N(0,20)$ is sufficiently small that it does not substantially affect the estimate of variability, as reflected by similarities of the fitted distributions and confidence intervals based upon random sampling error for the observed data and the error-free data based upon measurement errors of $N(0,10)$ and $N(0,20)$.

For the case of a measurement error of $N(0,40)$, Figure 8 displays the distribution fitted to the error free data and the confidence intervals on the fitted distribution based only upon random sampling error. Figure 9 includes uncertainty associated with random sampling error and measurement error for the same case. The distribution of error free data has less variability than the observed data because the contribution of measurement error to the observed data has been removed. Because the error-free distribution has less variability than for the other cases discussed, there is less random sampling error and the confidence intervals on the fitted distribution, as shown in Figure 8, are narrower than for the other cases. However, when measurement error is included as a source of uncertainty, as shown in Figure 9, the confidence intervals on the fitted distribution become wider. Figure 10 shows the results of characterization of random sampling error only for distribution fitted to error free data based upon a measurement error of $N(0,55)$. Compared to Figure 8, the fitted distribution in Figure 10 has less variability and narrower confidence intervals. However, when measurement error is included in the estimation of confidence intervals, as shown in Figure 11, the width of the confidence interval increases substantially compared to Figure 10. For the case of a measurement error of $N(0,55)$, measurement error contributes more to the overall uncertainty than does random sampling error.

These results indicate that when measurement error is a major source of uncertainty, ignoring uncertainty from measurement error will lead to an underestimate of total uncertainty. Failing to separately characterize contributions from random sampling error and measurement error, as done in Figure 3, will bring bias in the variability and uncertainty estimates if the measurement error is significant. This implies that it is

necessary to separately take into account the contributions from random sampling error and measurement error to the total uncertainty when measurement error is a main source of uncertainty.

Table 3 summarizes 95 percent confidence intervals in the mean when the uncertainty from measurement error is characterized and not characterized for all cases analyzed. Both analytical and numerical solutions for confidence intervals in the mean are provided in Table 3. The percentile method was used to form numerical confidence intervals for the mean for all cases based upon the results from bootstrap simulation. Central limit theorem is used to construct analytical and asymptotic confidence intervals. The results from bootstrap simulation agree well with the analytical solutions. Thus the asymptotic performance of the bootstrap solution for the 95 percent confidence interval for the mean was verified to be correct.

For the observed data, assuming no measurement error, the 95 percent confidence interval of the mean is approximately 84 to 132. As the measurement error increases, the 95 percent confidence interval of the mean estimated based only upon random sampling error decreases in range. For example, for a measurement error of $N(0,55)$, the 95 percent confidence interval of the mean is approximately 97 to 118, which is substantially narrower than the confidence interval for the observed data. However, if both random sampling error and measurement error are considered simultaneously, there is no substantial change in the range of confidence intervals for the mean as shown in Figure 12. For example, for the case of $N(0,55)$, the 95 percent confidence interval for the mean is approximately 85 to 131, which is approximately the same as that for the observed data assuming no measurement error. An important result,

therefore, is that the estimate of uncertainty in the mean is the same for the observed data assuming no measurement error as it is for an error-free data set when both random sampling and measurement error are considered. There is also no substantial change in the shape of the sampling distribution for the mean as shown in Figure 12. Thus, if one is interested only in an estimate of uncertainty in the mean, it is possible to obtain it based upon the observed data without the need to separate random sampling and measurement error. However, in order to get an unbiased estimate of true variability, it is necessary to separate measurement error from the observed variability.

4.0 CONCLUSION

This paper demonstrates methods for improving the characterization of variability and uncertainty if there are known measurement errors in environmental data. A method for constructing an error free data set was developed based on the observed data set, known measurement error and measurement error models. Numerical methods based upon bootstrap pair techniques were successfully applied to characterize uncertainty in statistics for the measurement error problem.

The effect of measurement error on quantification of variability and uncertainty was evaluated and investigated with respect to the size of measurement error. The investigation results indicate that when measurement error is not a main source of uncertainty, there is no substantial difference between the distribution representing the observed data set and the distribution representing the error free data set. However, when measurement error is a main source of uncertainty, substantial difference will appear between the distribution representing variability of the observed dataset and the distribution representing variability of the error free data set. Variability will be

overestimated if it is based on the observed data without adjustments for measurement error. Therefore, in such situations, the distribution for the observed dataset cannot be used to directly describe the variability of a quantity.

The results from two-dimensional simulation indicate that uncertainty will be underestimated if uncertainty arising from measurement error is subtracted not characterized. When the contribution from measurement error to the total uncertainty is considered, there is no substantial difference among 95% confidence intervals and sampling distributions for the mean for the observed data set and the error free data sets. . Thus, if one is interested only in an estimate of uncertainty in the mean, it is possible to obtain it based upon the observed data without the need to separate random sampling and measurement error. However, when measurement error is a main source of uncertainty, there exist substantial bias in the estimates of true variability. Thus, in order to get an unbiased estimate of true variability, it is necessary to separate measurement error from the observed variability.

Future study include the further development of methods of constructing error free data sets, such as a multiplicative error model is used to describe known measurement errors. There is a need for methods for quantification of variability and uncertainty if there exists multiple and different measurement error models in an observed dataset.

ACKNOWLEDGEMENTS

The authors appreciate Dr. L.A. Stefanski from the Department of Statistics at North Carolina State University for his valuable hints and suggestions about the construction of error free data.

References

1. Cullen, A.C., and Frey, H.C.; *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press: New York, 1999.
2. Morgan, M.G., and M. Henrion; *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York, 1990.
3. Finkel, A.M., *Confronting Uncertainty in Risk Assessment: A Guide for Decision Makers*, Center for Risk Management, Resources for the Future, Washington DC, 1990.
4. Burmaster, D.E. and Harris, R.H., "The magnitude of compounding conservatisms in Superfund risk assessments", *Risk Analysis*, **1993**, 13(2):131-143.
5. Cullen, A.C., "Measures of conservatism in probabilistic risk assessment", *Risk Analysis*, **1994**, 14: 389-392.
6. Thompson, K.M. and Graham, "Going beyond the single number: using probabilistic risk assessment to improve risk management", *Human and Ecological Risk Assessment*, **1996**, 2(4):1008-1034.
7. Frey, H.C., J. Zheng, 2002, "Quantification of Variability and Uncertainty in Utility NO_x Emission Inventories," *J. of Air & Waste Manage. Assoc.*, accepted for publications.
8. Frey, H.C., "Variability and Uncertainty in Highway Vehicle Emission Factors," Emission Inventory: Planning for the Future (held October 28-30 in Research Triangle Park, NC), Air and Waste Management Association, Pittsburgh, Pennsylvania, pp. 208-219, October 1997.
9. NRC, *Science and Judgment in Risk Assessment*, National Research Council, National Academy Press, 1994, Washington, D.C.
10. Frey, H.C. and Rhodes, D.S.; "Characterizing, simulating and analyzing variability and uncertainty: An illustration of methods using an air toxics emissions example", *Human and Ecological Risk Assessment*. **1996**, 2(4): 762-79
11. Frey, H.C., and Rhodes, D.S., "Characterization and simulation of uncertain frequency distributions: Effects of distribution choice, variability,

uncertainty, and parameter dependence,” *Human and Ecological Risk Assessment*, **1998**, 4(2): 423-468

12. Frey, H.C., R. Bharvirkar, J. Zheng; “Quantitative Analysis of Variability and Uncertainty in Emissions Estimation”; Final Report, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, pp 1-10,1999
13. Frey, H.C., and S. Bammi, "Quantification of Variability and Uncertainty in Lawn and Garden Equipment NO_x and Total Hydrocarbon Emission Factors," *Journal of the Air & Waste Management Association*, accepted January 2002 for publication
14. Li, S., and H.C. Frey, 2002, “Methods and Example for Development of a Probabilistic Per-Capita Emission Factor for VOC Emissions from Consumer/Commercial Product Use”, *Proceedings, Annual Meeting of the Air & Waste Management Association*, Pittsburgh, PA, June 2002 (in press).
15. Chesher, Andrew, “The Effect of Measurement Error”, *Biometrika*, Vol. 78 (3), 1991, pp. 451-562
16. Rabinovich, S., *Measurement Errors and Uncertainties: Theory and Practice*, Spinger-Verlag, New York, 1999
17. R.H. Dieck, *Measurement Uncertainty, Methods and Applications*, ISA, 1992
18. B.D. Ellis, *Basic Concepts of Measurement*, Cambridge University Press, Cambridge, 1966
19. Barford, N. C., 1985: *Experimental Measurements: Precision, Error, and Truth*. John Wiley and Sons, New York, 1985
20. Carroll, R.J., D. Ruppert, L.A. Stefanski, *Measurement Error in Nonlinear Models*, Chapman & Hall, New York, 1995
21. Fuller, Wayne A., *Measurement Error Models*, John Wiley & Sons, New York, 1987
22. Aigner, D.J., C.A.K. Lovell and P. Schmidt, “Formulation and estimation of stochastic frontier production functions.” *Journal of Econometrics*, **1977**, 6, 21-37.
23. Horowitz, J.L. and M. Markatou, “Semiparametric Estimation for Regression Models for Panel Data.” *Review of Economic Studies*, 1999 63, 145-168.

24. Stefanski, L and R.J. Carroll , “Deconvoluting Kernel Density Estimators.” *Statistics*, 1990, 21, 169-184.
25. Zheng, J., and H.C. Frey, 2001, "Quantitative Analysis of Variability and Uncertainty in Emission Estimation: An Illustration of Methods Using Mixture Distributions," *Proceedings, Annual Meeting of the Air & Waste Management Association*, Orlando, FL.
26. Efron, B., 1979, “Bootstrap Method: Another Look at the Jackknife,” *The Analysis of Statistics*, 7(1): 1-26.
27. Frey, H.C., J. Zheng; “Method and Example Case Study for Analysis of Variability and Uncertainty in Emission Estimation (AUVVE) ”, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, pp 1-10,1999.
28. Efron,B., Tibshirani,R.J.; *an Introduction to the Bootstrap*, Chapman & Hall, London, UK,1993.
29. Casella, G., R. L. Berger, *Statistical Inference*, Duxbury Press: Belmont, CA, 1990.
30. Frey, H.C., J. Zheng *et al.*, “Technical Documentation of the AuvTool Software Tool for Analysis of Variability and Uncertainty,” Prepared by North Carolina State University for Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, February, 2002.

Table 1. The Observed Data Set and Estimated Error Free Data Sets under Different Measurement Error Models

Index	Observed Data set	Estimated Error Free Data Sets *			
		$N(0,10)^a$	$N(0, 20)^a$	$N(0, 40)^a$	$N(0, 55)^a$
1	95.022	95.189	95.702	98.035	102.007
2	85.157	85.457	86.383	90.592	97.758
3	43.382	44.248	46.922	59.074	79.761
4	102.205	102.274	102.487	103.455	105.102
5	141.472	141.009	139.579	133.080	122.017
6	117.685	117.544	117.110	115.134	111.770
7	124.200	123.971	123.264	120.049	114.577
8	80.632	80.993	82.108	87.178	95.808
9	44.411	45.264	47.894	59.851	80.205
10	10.649	11.958	16.001	34.377	65.660
11	122.097	121.896	121.277	118.462	113.671
12	40.353	41.260	44.060	56.789	78.456
13	20.158	21.339	24.984	41.552	69.757
14	96.820	96.962	97.400	99.391	102.782
15	258.736	256.683	250.349	221.553	172.533
16	119.963	119.791	119.261	116.852	112.751
17	141.939	141.470	140.020	133.433	122.219
18	117.154	117.021	116.608	114.733	111.542
19	185.651	184.589	181.311	166.412	141.049
20	213.246	211.810	207.378	187.232	152.937
21	77.302	77.709	78.963	84.666	94.374
22	62.725	63.329	65.193	73.668	88.094
23	127.307	127.036	126.199	122.394	115.915
24	54.245	54.964	57.183	67.270	84.441
25	199.831	198.577	194.707	177.111	147.158
Mean	107.294	107.294	107.294	107.294	107.294
Standard Deviation	60.945	60.119	57.570	45.981	26.254
Sampling Variance	3714.292	3614.295	3314.289	2114.290	689.292

Note: *: Error free data sets are estimated based upon Equations (7).

^a: $N(0, \sigma_e)$, a normal distribution with mean of 0 and standard deviation of σ_e .

Table 2. The Fitted Lognormal Distributions under Different Measurement Error Models

Measurement Error model	First Parameter ^a	Second Parameter ^b	Kolmogorov-Smirnov Test Value
N(0.0, 0.0)*	4.541	0.519	0.146
N(0.0, 10.0)	4.544	0.513	0.145
N(0.0, 20.0)	4.553	0.494	0.141
N(0.0, 40.0)	4.594	0.403	0.123
N(0.0, 55.0)	4.678	0.236	0.092

Note: At the significant level of 0.05, the critical value of $n=25$ is 0.18⁽³⁰⁾

*: No measurement error is considered for this case.

^a: Mean of $\ln x$

^b: Standard deviation of $\ln x$

Table 3. Uncertainty in the Mean under Different Measurement Error Models

Measurement Error model	Analysis Method ^a	Confidence Intervals for Mean (Random Sampling Error only)			Confidence Intervals for Mean (Both Random Sampling and Measurement Error)		
		2.5%	Mean	97.5%	2.5%	Mean	97.5%
N(0.0, 0.0)*	Analytic	83.4	107.3	132.4	83.4	107.3	132.4
	Numerical	85.8	107.3	132.0	85.8	107.3	132.0
N(0.0, 10.0)	Analytic	83.7	107.3	132.1	83.4	107.3	132.4
	Numerical	86.4	107.2	132.3	85.8	107.2	132.4
N(0.0, 20.0)	Analytic	84.7	107.3	131.1	83.4	107.3	132.4
	Numerical	87.4	107.5	131.4	84.9	107.2	132.3
N(0.0, 40.0)	Analytic	89.3	107.3	126.3	83.4	107.3	132.4
	Numerical	90.8	107.3	126.3	84.1	107.3	132.2
N(0.0, 55.0)	Analytic	97.0	107.3	118.1	83.4	107.3	132.4
	Numerical	97.5	107.3	117.7	84.2	107.4	130.7

Note: The results listed here are the average values of 10 different simulations for each case.

*: No measurement error is considered for this case.

^a: Analytical solutions are based upon central limit theorem; numerical solutions are estimated from bootstrap simulation.

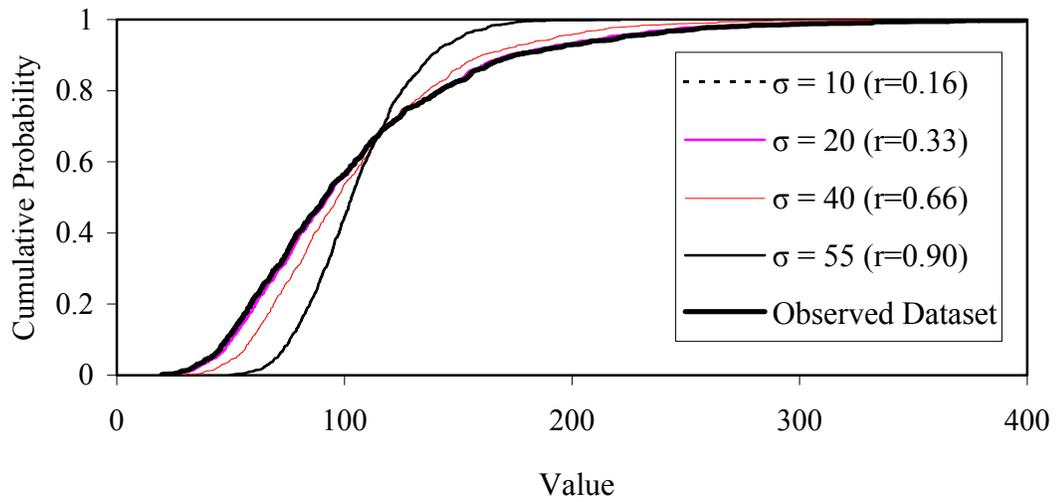


Figure 2. The Fitted Lognormal Distributions to the Error Free Data Set under Different Measurement Errors and the Observed Data Set

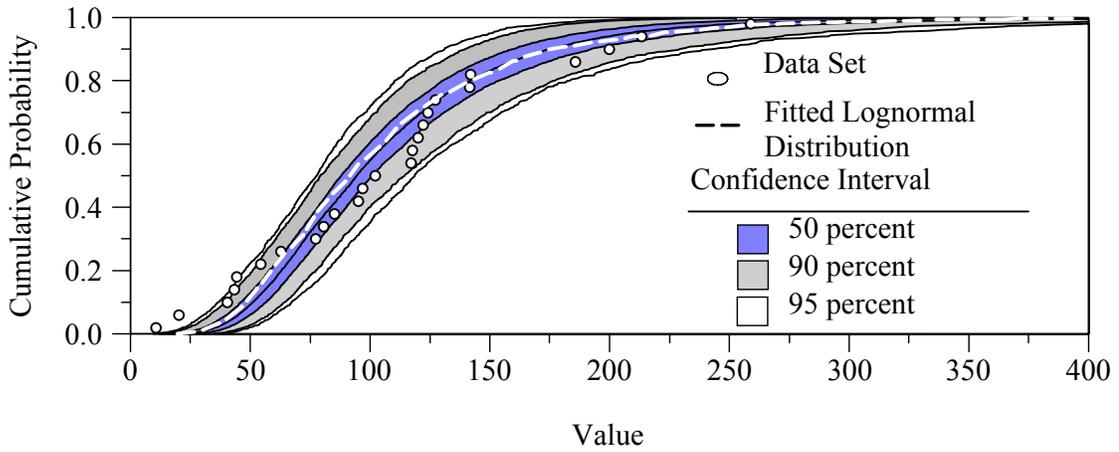


Figure 3. The Probability Band Based upon the Fitted Lognormal Distribution to the Observed Data Set (no measurement error is assumed)

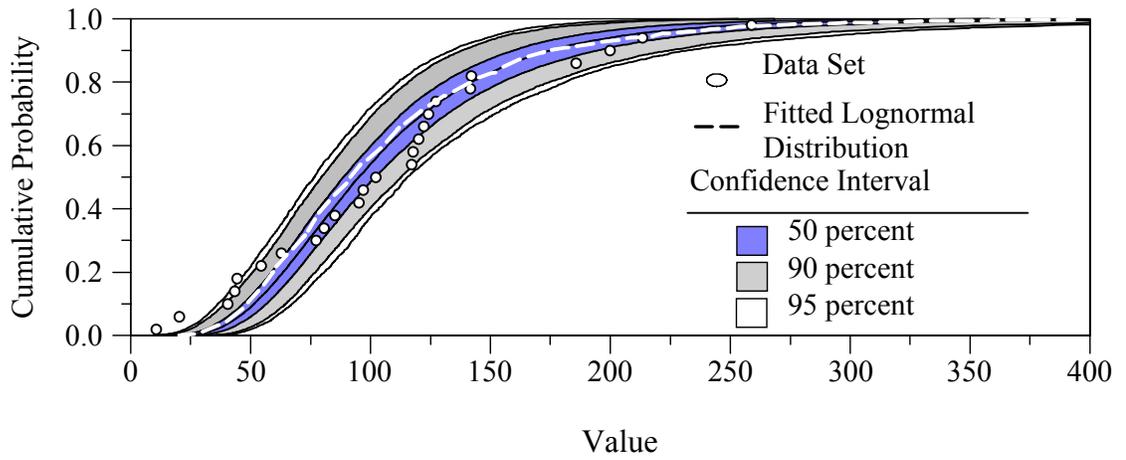


Figure 4. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=10.0$ without the Inclusion of Measurement Error

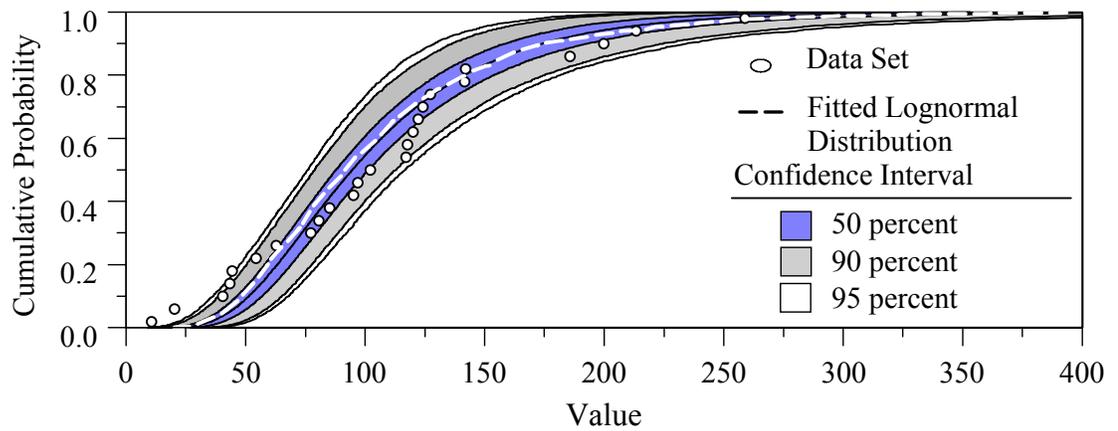


Figure 5. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=10.0$ with the Inclusion of Measurement Error

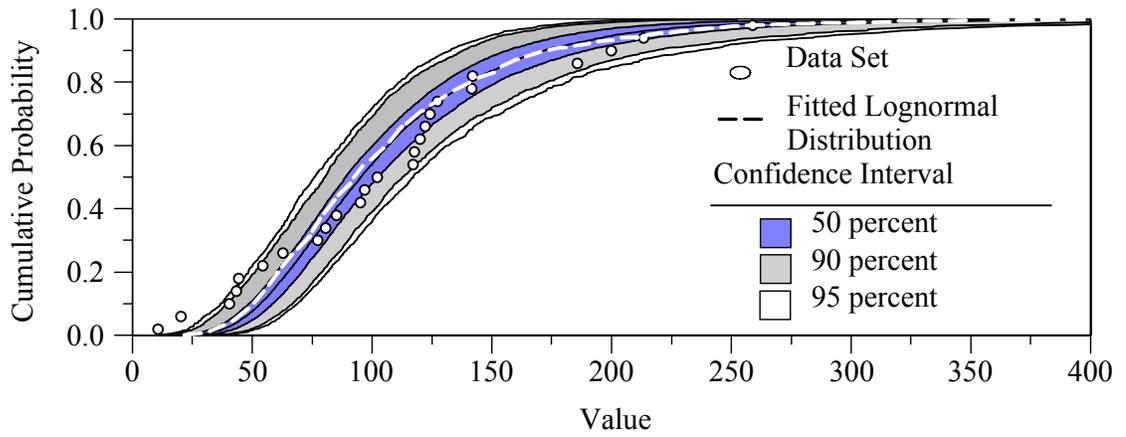


Figure 6. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=20.0$ without the Inclusion of Measurement Error

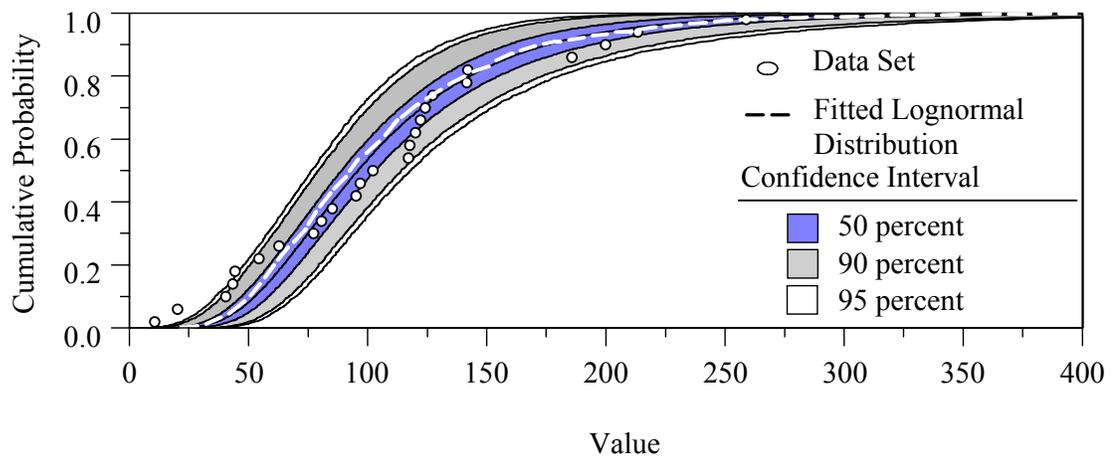


Figure 7. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=20.0$ with the Inclusion of Measurement Error

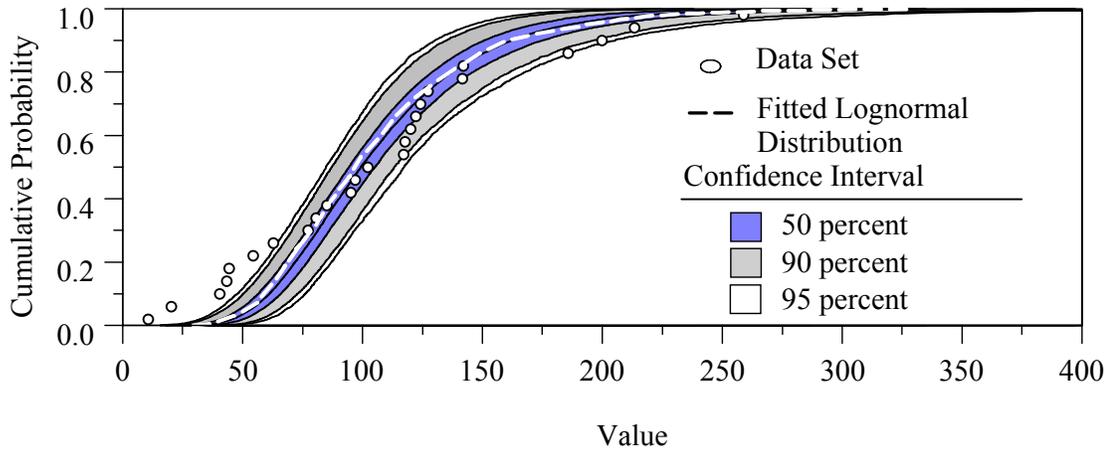


Figure 8. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=40.0$ without the Inclusion of Measurement Error

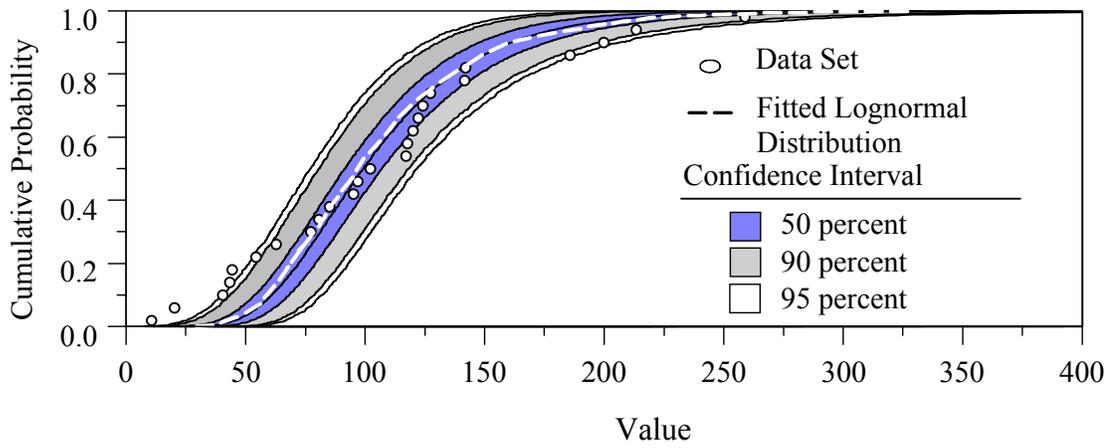


Figure 9. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=40.0$ with the Inclusion of Measurement Error

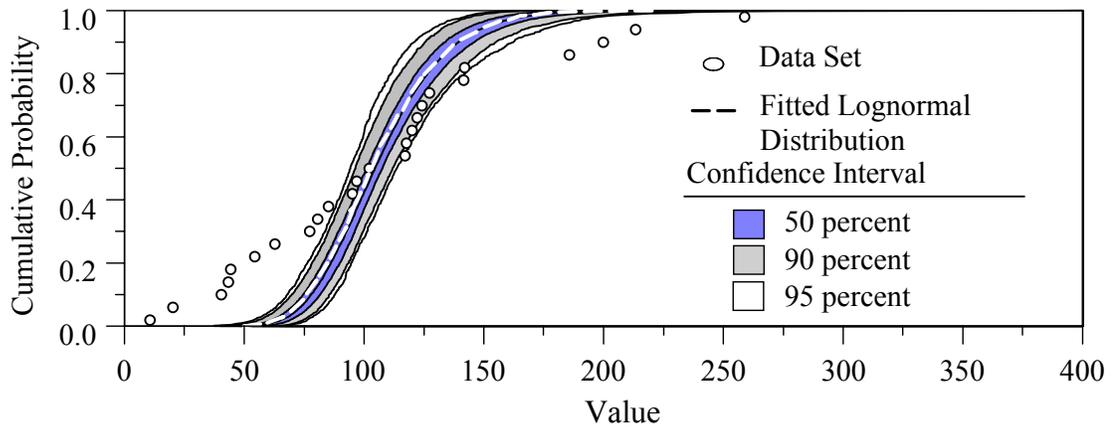


Figure 10. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=55.5$ without the Inclusion of Measurement Error

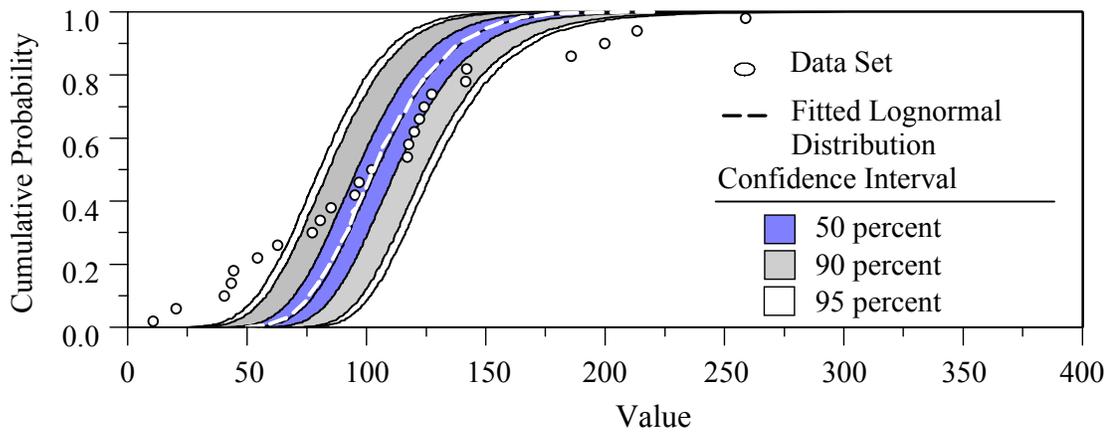


Figure 11. The Probability Band Based upon the Fitted Lognormal Distribution to the Error Free Data at $\sigma_\epsilon=55.5$ with the Inclusion of Measurement Error

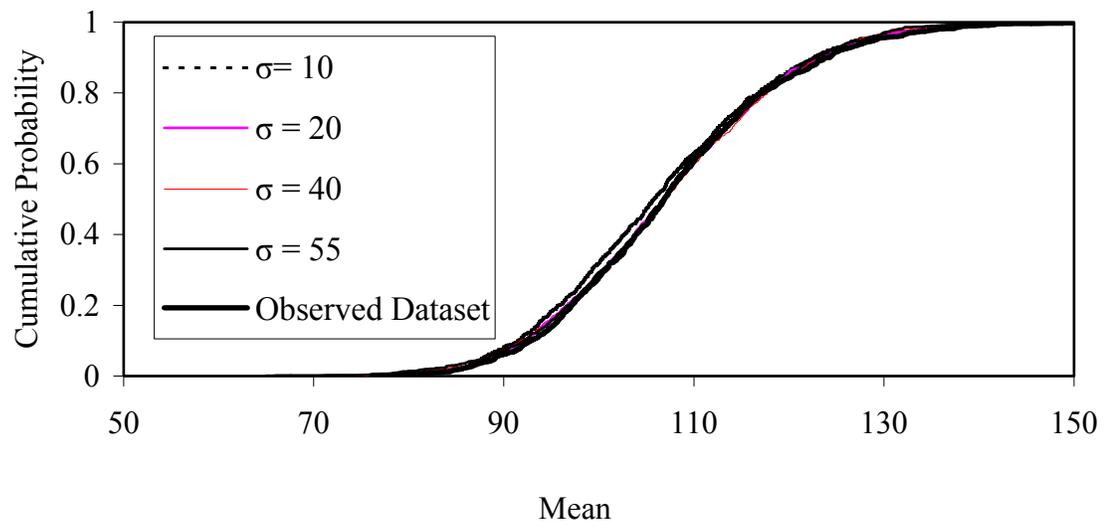


Figure 12. The Sampling Distributions for the Mean under Different Measurement Error Models.

PART VI

**QUANTIFICATION OF VARIABILITY AND
UNCERTAINTY IN AIR POLLUTANT EMISSION
INVENTORIES: METHOD AND CASE STUDY FOR
UTILITY NO_x EMISSIONS**

H. Christopher Frey and Junyu Zheng

Accepted by *Journal of Air and Waste Management Association* for Publication

Quantification of Variability and Uncertainty in Air Pollutant Emission Inventories: Method and Case Study for Utility NO_x Emissions

H. Christopher Frey

Junyu Zheng

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695-7908

ABSTRACT

The quality of stationary source emission factors is typically described using data quality ratings, which provide no quantification of the precision of the emission factor for an average source, nor of the variability from one source to another within a category. Variability refers to actual differences due to differences in feedstock composition, design, maintenance, and operation. Uncertainty refers to lack of knowledge regarding the true emissions. A general methodology for the quantification of variability and uncertainty in emission factors, activity factors, and emission inventories is described, featuring the use of bootstrap simulation and related techniques. The methodology is demonstrated via a case study for a selected example of NO_x emissions from coal-fired power plants. A prototype software tool was developed to implement the methodology. The range of inter-unit variability in selected activity and emission factors is shown to be as much as a factor of four, and the range of uncertainty in mean emissions is shown to depend on the inter-unit variability and sample size. The uncertainty in the total inventory of -16 % to +19 % was attributed primarily to one technology group, suggesting priorities for collecting data and improving the inventory. The implications for decision-making are discussed.

IMPLICATIONS

Emission factors, activity factors, and emission inventories are widely used for air quality management decisions. However, the uncertainty in these numbers is typically ignored or treated only qualitatively. This work represents an effort to develop and promote the use of quantitative techniques for characterizing both variability and uncertainty.

Knowledge of uncertainty allows analysts and decisions makers to evaluate the significance of time series trends, estimate the likelihood of meeting emissions budgets, estimate uncertainty in predicted air quality, and accurately estimate the cost savings of permit trading. Decision-making that takes into account uncertainty is likely to be more robust to uncertainty and, therefore, more effective.

Key Words: Variability, Uncertainty, Emissions, Emission Inventories, Emission Factors, Activity Factors, Monte Carlo simulation, Bootstrap Simulation, Nitrogen Oxides

1.0 INTRODUCTION

Emission inventories (EIs) are used at federal, state, and local governments and by private corporations for: (a) characterization of temporal emission trends; (b) emissions budgeting for regulatory and compliance purposes; and (c) prediction of ambient pollutant concentrations using air quality models. If random errors and biases in the EIs are not quantified, they can lead to erroneous conclusions regarding trends in emissions, source apportionment, compliance, and the relationship between emissions and ambient air quality.¹ Furthermore, an understanding of inter-unit variability in emissions can help identify cost saving opportunities where a unit with below average emissions can engage in an emission permit trade with units that are above average as part of possible future permit trading programs.

There is growing recognition of the importance of quantitative uncertainty analysis in environmental modeling and assessment. The National Research Council has recently recommended that quantifiable uncertainties be addressed in estimating mobile source emission factors, and in the past has addressed the need for understanding of uncertainties in emission inventories used in air quality modeling and in risk assessment.^{2,3} The U.S. Environmental Protection Agency (EPA) has developed guidelines for Monte Carlo analysis of uncertainty, and has also sponsored several workshops regarding probabilistic analysis.^{4,5,6} Probabilistic techniques have recently been applied to estimation of uncertainty in emission factors for mobile sources, major stationary sources and area sources.^{1,7-17}

Both variability and uncertainty should be taken into account in the process of developing a probabilistic emission inventory. Variability is the heterogeneity of values with respect to time, space, or a population. Variability in emissions arises from factors such as: (a) variation in feedstock (e.g., fuel) compositions; (b) inter-plant variability in design, operation, and maintenance; and (c) intra-plant variability in operation and maintenance. Uncertainty arises due to lack of knowledge regarding the true value of a quantity. It refers to statistical sampling error, measurement errors, and systematic errors. In most cases, emissions estimates are both variable and uncertain. Therefore, we employ a methodology for simultaneous characterization of both variability and uncertainty based upon previous work in emissions estimation, exposure assessment, and risk assessment.^{12,15-18} The method features the use of Monte Carlo and bootstrap simulation. However, this work expands on previous work by propagating uncertainties through an emission inventory, as opposed to focusing on emission factors. Thus, one of

the key contributions of this work is to synthesize specific techniques into one framework for estimating uncertainties in emission inventories and identifying key sources of uncertainty. The specific techniques include developing databases, fitting distributions to data, characterizing uncertainty in the fitted distribution, propagating both variability and uncertainty in inputs to predict uncertainty in the total emission inventory, and identifying the key sources of uncertainty. The synthesis of techniques was accomplished by developing a new prototype software tool that incorporates each of the techniques. The software tool is referred to as Analysis of Uncertainty and Variability in Emission Estimation (AUVVE). The probabilistic method for estimating emission inventories is illustrated using an case study of utility NO_x emissions.

The perspective of the case study is with respect to estimating emission in the near future. Clearly, with the prevalence of continuous emission monitoring (CEM) equipment for measuring hourly NO_x emissions from a large number of power plants in the U.S., it is possible in many cases to characterize recent emissions of these plants with a comparative high degree of accuracy.¹⁹ However, when making estimates of emissions any time into the future, it is more difficult to make a precise prediction. This is because there is underlying variability in the emissions of a single unit from one time period to another, even if the unit load is similar. Therefore, the purposes of this paper are: (1) to demonstrate a general probabilistic approach for quantification of variability and uncertainty in emission factors and emission inventories for utility electric power plant emission category; (2) to demonstrate the insights obtained from the general probabilistic approach regarding the ranges of variability and uncertainty in both emissions factors and emission inventories; (3) to demonstrate how probabilistic analysis can be used to

identify key sources of variability and uncertainty in an inventory for purposes of targeting additional work to improve the quality of the inventory; and (4) to illustrate the methodology developed in this paper by using a case study for electricity utility power plant NO_x emissions.

2.0 METHODOLOGY

The general approach employed to quantify variability and uncertainty in emission inventories and emission factors can be summarized as the following major steps:¹⁸

7. Compilation and evaluation of a database for emission and activity factors;
8. Visualization of data by developing empirical cumulative distribution functions for individual activity and emission factors;
9. Fitting, evaluation, and selection of alternative parametric probability distribution models for representing variability in activity data and emission factor data;
10. Characterization of uncertainty in the distributions for variability;
11. Propagation of uncertainty and variability in activity and emissions factors to estimate uncertainty in facility-specific emissions and/or total emissions from a population of emission sources; and
6. Calculation of importance of sources of uncertainty via sensitivity analysis.

2.1 Compilation and Evaluation of a Database

A starting point of probabilistic analysis is to collect and analyze data. This includes data combing, data screening and data evaluating. The data used for the case study are based upon Continuous Emission Monitoring (CEM) data for individual power plant units obtained through EPA. The data are from the quarterly "Preliminary Summary

Emissions Reports" of the Acid Rain Program of EPA.¹⁸ Two averaging times are considered: (1) 6-month; and (2) 12-month. The purpose of the 6-month averaging time is to characterize emissions that include the "ozone season." The purpose of the 12-month averaging time is to characterize annual emissions for emissions budgeting and other purposes. The 6-month averages are based upon combining data from the 2nd and 3rd quarters of the year, including the months from April through September. At the time that the data collection effort was made, quarterly data were available for the 1st quarter of 1997 through the 2nd quarter of 1999. Therefore, complete datasets of four quarters were available only for 1997 and 1998. Furthermore, data sets needed to characterize the 6-month period inclusive of the summer were available only for 1997 and 1998. Data from all ten available quarters were combined into a single database. In the database, each record represents a power plant unit or stack. Each record contains the following information:

- Unit/Stack Identification (Unit ID and ORISPL identifier)
- General Information (State, Region)
- Technology Group (Boiler Type, NO_x Control Technology)
- Operation Data (Capacity, Operating Time)
- NO_x Emission Data

For some units or stacks, there was not sufficient information regarding the maximum gross capacity, the control technology, or the emission rate. Without any of these pieces of information, it is not possible to completely characterize or categorize both the activity factors and the emission factor for that particular unit or stack. Activity factors include the heat rate and the capacity factor. Thus, records that are incomplete were screened out of the database to create a "clean" database comprised only of complete records.

Four quantities are calculated from the final database developed in this project: (1) unit/stack heat rate (BTU/kWh); (2) unit/stack capacity factor (actual kWh generated/maximum possible kWh); (3) NO_x emission rate on a fuel input basis (g/GJ); and NO_x emission rate on an energy output basis (g/GJ).

The emissions and activity data are calculated for selected technology groups. Four technology groups were selected based upon the most prevalent types of units in the database. These include: (1) dry bottom, wall-fired boilers with no NO_x control; (2) dry bottom, wall-fired boilers with low NO_x burners (LNB); (3) tangential-fired boilers no NO_x controls; and (4) tangential-fired boilers with low NO_x burners and overfire air option 1, referred to as LNC1. The number of data points for these four technology groups ranges from 36 to 136, depending upon the technology group and the averaging time used. In addition, the technology group of dry bottom, turbo-fired boilers with overfire air was selected because it has a small sample size (n=6). The reason for selecting this group was to demonstrate that the probabilistic method is able to deal with small data sets.

To simplify the database as much as possible, it is desirable to be able to select data for one representative year. The 12-month data for 1997 and 1998 were compared in Figure 1 to identify similarities and differences between them. The figure indicates that there is a strong correlation between NO_x emission rate data of 1997 and 1998 except that there are few units whose emission rates in 1998 is significantly higher than the ones in 1997. For example, there is a unit that had an emission rate of approximately 260g/GJ in 1997 but only 160 g/GJ in 1998. Units such as this have had a retrofit of emission control technology during either 1997 or 1998. Other data such as capacity factor and heat rate

were also similar for the two years, implying that data for either year could be used as the basis for probabilistic statistical analysis, as long as data are properly classified with respect to control technology. The 1998 data were selected for further analysis. Possible statistical dependencies between activity factors and emission factors were evaluated in detail.¹ No significant dependencies among emission factors and activity factors were identified. For example, Figure 2 illustrates that the 12-month average capacity factor for a technology group does not depend on the value of the heat rate. Therefore, it was not necessary to attempt to simulate statistical dependencies among emission factors and activity factors. However, it would be the case that a general increase or decrease in overall system load when comparing one six month or 12 month time period to another could lead to a systematic increase or decrease, respectively, in capacity factors among multiple units. Similarly, unit retirements or addition of new capacity could change the average capacity factor of the in-service units. These system expansion and dispatching considerations are not explicitly addressed here, since the main focus is on demonstrating a method for quantifying uncertainty in emission inventories. The implications of such considerations are discussed in the Conclusions.

2.2 Visualization of Data

Some of the key purposes of visualizing data sets include: (1) evaluation of the central tendency and dispersion of the data; (2) visual inspection of the shape of the empirical distribution of the data as a potential aid in selecting parametric probability distribution models to fit to the data; (3) identification of possible anomalies in the data set (e.g., outliers); and (4) identification of possible dependencies between variables. Specific techniques for evaluating and visualizing data include calculation of summary statistics,

development of empirical cumulative distribution functions, and generation of scatter plots for the evaluation of dependencies between pairs of activity and emission factors. An assumption is that all the quantities considered in this study are treated as continuous random variables.¹⁷

A Cumulative Distribution Function (CDF) is a relationship between “cumulative probability” and values of the random variable. Cumulative probability is the probability that the random variable has values less than or equal to a given numerical value.

Development of empirical CDFs of data are discussed elsewhere^{17,20,21}

2.3 Fitting, Evaluating, and Selecting Parametric Probabilistic Distribution Models

A probability distribution model is typically represented as a probability density function (PDF) or a CDF for a continuous random variable. The PDF for a continuous random variable indicates the relative likelihood of values. The CDF is obtained by integrating the PDF.¹⁷

Probability distribution models may be empirical, parametric, or combinations of both. A parametric probability distribution model is described by parameters. Data sets can be described in a compact manner based on a particular type of parametric distribution function and the values of its parameters. A potential advantage of parametric probability distributions compared to empirical distributions is that it is possible to make predictions in the tails of the distribution beyond the range of observed data and to interpolate within the range of the observed data. In contrast, using conventional empirical distributions, the minimum and maximum values of the distribution are limited to the minimum and maximum values, respectively, of the data set. These values typically change as more data are collected.

In choosing a distribution function to represent either variability or uncertainty, it is often useful to theorize about processes that generate both the data and particular types of distributions. *A priori* knowledge of the mechanisms that impact a quantity may lead to the selection of a distribution to represent that quantity. For example, an underlying mechanism based on the central limit theorem (CLT) may lead to the selection of the Normal or Lognormal distribution. Other factors to consider may be whether values must be non-negative, which rules out infinite two-tailed distributions such as the Normal, or whether or not the distribution is symmetric. Discussions of distribution selection criteria can be elsewhere.^{17,22,26-28}

Once a particular parametric distribution has been selected, a key step is to estimate the parameters of the distribution. The method of Maximum Likelihood Estimation (MLE) and the Method of Matching Moments (MoMM) are among the most typical techniques used for estimating the parameters.^{22,23} In this work, MLE is employed, and non-linear optimization was used to find the parameter values that maximize the likelihood function.

The fitted parametric distributions that are hypothesized to represent the population from which the available data were drawn may be evaluated for goodness-of-fit using probability plots and test statistics. Probability plots are a subjective method for determining whether or not data contradict an assumed model based upon visual inspection. Statistical tests can be used in conjunction with probability plots to provide a numerical indication of the goodness-of-fit.^{17, 22,24,25} However, each goodness-of-fit method focuses on only one measure and may not capture the preference of an analyst or decision maker. In this study, the empirical distribution of the actual data set is compared

visually with the cumulative probability functions of the fitted distributions to aid in selecting the probability distribution model which best describes the observed data. The bootstrap technique described in the next section is also used to check the adequacy of the fit.

2.4 Characterization of Uncertainty in the Distributions for Variability

Bootstrap simulation is used to quantify uncertainty in the inputs to the emission inventory. Bootstrap simulation was introduced by Efron as a means for calculating confidence intervals for statistics in a general manner for situations in which analytical solutions are not available.²⁹ A probabilistic framework for calculating uncertainty in emissions estimation using bootstrap simulation is described in detail elsewhere.^{1,12-18} In bootstrap simulation, a probability distribution is assumed to be a best estimate of the true but unknown population distribution for a quantity. Using random Monte Carlo sampling, synthetic data sets of the same sample size as the original observed data set, referred to as “bootstrap samples,” are simulated from the assumed population distribution. For each bootstrap sample, a value of the statistic of interest, such as the mean, standard deviation, distribution parameters, or fractiles of the distribution, is calculated. A statistic estimated from a bootstrap sample is referred to as a *bootstrap replicate* of the statistic. Typically, many bootstrap samples are simulated to yield hundreds or more bootstrap replications. The bootstrap replications describe a *sampling distribution* for the statistic. A sampling distribution is a probability distribution for a statistic. From the sampling distribution, probability ranges can be inferred. For example, the 95 percent probability range for the mean can be estimated from 2.5th and 97.5th percentiles of the bootstrap replicates.

2.5 Propagation of Uncertainty and Variability through a Model

For a power plant unit, activity data included the unit heat rate, unit capacity factor, and unit capacity. Based on the different types of NO_x control technology and boiler types, units in the inventory are classified into different technology groups. An annual emission inventory for all units in a technology group is given by:

$$TE_j = c \sum_{i=1}^N (EF_{i,j})(HR_{i,j})(f_{c,i,j})(C_{i,j}) \quad (1)$$

Where:

$EF_{i,j}$ = emission factor for unit i of technology group j (lb/10⁶ BTU)

$HR_{i,j}$ = heat rate for unit i of technology group j (BTU/kWh)

$f_{c,i,j}$ = Annual capacity factor for unit i of technology group j (actual kWh generated/maximum possible kWh)

$C_{i,j}$ = capacity for unit i of technology group j (MW)

N = number of units in technology group j

TE_j = Total emissions for all units in the j^{th} technology group (lb/year)

c = Conversion factor for dimensional consistency, 8.76 lb/year.

For a given technology group, N random samples are generated for heat rate, capacity factor, and NO_x emission factor from the corresponding parametric probability distributions for each of these three quantities. Each of the N random samples represents one unit in the emission inventory for the selected technology group. The calculation is repeated for each of the N units in the technology group to arrive at total emissions for the group.

The process of randomly simulating heat rate, capacity factor, and emission factor values for all of the N units is repeated to arrive at another estimate of total emissions for the technology group. The second estimate of total emissions will differ from the first because of random sampling fluctuations in the inputs. This process is repeated B times, to arrive at B estimates of the emission inventory of the technology group. The B estimates of total emissions for a technology group characterize a distribution for uncertainty in the total emissions. This process was conducted for each technology group. The emission inventory is calculated as the sum of emissions for all technology groups, and this process is repeated many times in the numerical simulation to characterize a distribution of the sum of emissions for all technology groups.

$$SE = \sum_{j=1}^G TE_j \quad (2)$$

where:

G: = Number of technology group

SE: = Sum of emissions from all technology groups (e.g., lb/year)

2.6 Identifying Key Sources of Uncertainty

The calculation of the importance of uncertainty from different model inputs is useful because it can indicate which model input makes the most contribution to uncertainty in a selected model output. Such information helps determine where to target additional research or data collection to reduce uncertainty in a model input, thereby leading to a reduction in uncertainty in the model output.

There are a variety of measures for evaluating the relative importance of uncertainties in model inputs.^{17, 26} The approach employed here is to calculate the sample correlation coefficient between the distribution of uncertainty in a technology group

emission inventory and the total emission inventory. The sample correlation coefficient is a measure of the linear dependence of the model output with respect to the selected model input.

3.0 INTRODUCTION TO AUVÉE

The AUVÉE prototype software provides features for fitting distributions to data, characterizing variability and uncertainty in emission and activity factors, calculating probabilistic emission inventories, and identifying key sources of uncertainty.^{18,30} The functional design of AUVÉE, the composition of the main modules and the relationships among them are briefly described.

Figure 3 shows the conceptual design of AUVÉE. AUVÉE is composed of 3 databases, which include an internal database, a user input database and an interim database. In addition, AUVÉE includes four main modules: (1) fitting distributions; (2) characterizing uncertainty; (3) calculating emission inventories; and (4) user data input. AUVÉE features an interactive Graphical User Interface (GUI). The program itself is written in FORTRAN, and the GUI was developed using C++ and support applications.

The internal database for AUVÉE includes emission and activity factors obtained from CEMS data, as described in the previous section. The user may select either a 6-month average or a 12-month average database as the basis for developing either a 6-month or 12-month emission inventory, respectively. The user input database stores data that the user provides regarding the number of power plant units, the boiler and emission control technology for each unit, and the capacity of each unit. This database can be edited by the user via the user data input module.

The interim database in AUVÉE is used to store the results from the fitting distribution module and to store project information. The interim database provides fitted

distribution information needed by the uncertainty analysis and emission inventory modules. A default interim database is provided so that the user can proceed to calculate emission inventory results even without making a new selection of parametric distributions to represent each input to the emission inventory.

The fitting distribution module implements all calculations for fitting parametric distributions to emission factor and activity factor data. This module provides graphs comparing the CDF of the fitted distributions to an empirical distribution of the data, allowing the user to visually evaluate the goodness of fit of parametric distributions fitted to datasets from the internal database. The user has the option, via a pull-down menu, to select alternative parametric distributions for fit to the data. The user may choose from normal, lognormal, gamma, Weibull, and beta distributions. When the user exits the fitting distribution model, the current set of fitted distributions are saved to the interim database for use by other modules in the program.

The characterizing uncertainty module uses data from the interim database to get distribution information including distribution type and the parameters of the fitted distributions for emission and activity factors. Uncertainty estimates of the mean emission and activity factors, of other statistics (i.e. standard deviation, and distribution parameters) and confidence intervals of the CDF of the fitted distribution, are calculated using the numerical method of bootstrap simulation. The results of the uncertainty analysis are displayed in the GUI. The user can graphically compare the empirical distribution of the original data with the confidence intervals of the fitted CDF as a technique for evaluating goodness-of-fit as described further in the discussion of the case study. Because this module uses data from the internal database, which may contain a

relatively large number of power plant units compared to an individual state emission inventory, the estimates of uncertainty in the mean and in other statistics are typically a lower bound on the range of uncertainty in the same statistic applicable to an emission inventory that includes a smaller number of power plant units.

The emission inventory module: (1) allows the user to visit the user database and append, modify or delete user input data; (2) characterizes the uncertainty in emission factors and activity factors based on user project data; (3) calculates uncertainty in the emission inventory; and (4) calculates the key sources of uncertainty from among the different technology groups. The estimates of uncertainty in the emission inventory module are based upon the number of power plant units of each technology group specified by the user. For example, although there may be 36 power plant units of a given type in the internal database, the user may have only 10 units of that type in the emission inventory of interest. The uncertainty in the emission and activity factors for that technology group will be estimated based upon a sample size of 10, not 36.

A user's manual is available for AUVVEE that describes in more detail the key databases and modules.³⁰

4.0 CASE STUDY: A PROBABILISTIC EMISSION INVENTORY FOR UTILITIES IN A SINGLE STATE

The case study is based on the number of units of each technology group in a single state. The specific case study was selected because the number of units representing each of four power plant technologies is dissimilar. Specifically, the following numbers of units are included in the case study:

- 19 tangential-fired boilers with no NO_x controls (T/U)

- 11 tangential-fired boilers using Low NO_x Burners and overfire air option 1(T/LNC1)
- 12 dry bottom wall-fired boilers with no NO_x controls (DB/U)
- 3 dry bottom wall-fired boilers using low NO_x burners (DB/LNB)

No units of the technology group with dry bottom turbo-fired boilers and overfire air are present in the example state. Therefore, data for this technology group were not used in the case study. The case study is based upon a 12-month period. Parametric probability distributions were fitted to each activity and emission factor required for the inventory. The results are summarized in Table 1, estimated by AUVEE using MLE. Examples of the fitted distributions for the example of one technology group are shown in Figures 4, 5, and 6 for an emission factor, a capacity factor, and a heat rate, respectively. The goodness-of-fit was evaluated by comparing confidence intervals of the fitted distribution, obtained from bootstrap simulations, with the data. For example, the lognormal distribution fitted to the heat rate data agrees well with the tails of the distribution of the data. There are some deviations of the fitted distribution from the data in the regions of the 40th to 60th percentiles. However, more than half of the data are enclosed by the 50 percent confidence interval and almost 90 percent of the data are enclosed by the 95 percent confidence interval. Although on average it is expected that 95 percent of the data should be enclosed by the 95 percent confidence interval if the data are consistent with the assumed probability distribution model, some random variation of this percentage is also expected. Therefore, in this case the fitted distribution is deemed to be an adequate, although not perfect, match with the data.

The 50 percent confidence interval for the Beta distribution fitted to the capacity factor data encloses 57 percent of the data, and 98 percent of the data are enclosed by the

95 percent confidence interval. Similarly, for the Gamma distribution fit to the emission factor data, 67 percent of the data are enclosed by the 50 percent confidence interval and 90 percent of the data are enclosed by the 95 percent confidence interval. Both of these comparisons indicate a good fit, although the fit is not perfect in either case.

Figures 4 through 6 reveal substantial inter-unit variability in activity factors and the emission factor for the example technology group. The range of heat rate variability is from 9,100 BTU/kWh to 13,200 BTU/kWh. The capacity factor varies from 0.18 to slightly over 0.90. The emission factor varies from 70 g/GJ to 290 g/GJ. Thus, in some cases, the range of variability is more than a factor of four from the low to high end of the range. The range of the confidence intervals is influenced both by the range of variability of the data and by the sample size.

In the example inventory, there are only 3 units of the specific technology group represented in Figures 4, 5, and 6. Thus, although there are a total of 98 such units represented in the database, the uncertainty estimate specific to the example inventory must account for the fact that there are only 3 units in the inventory. An assumption is that the 3 units are a random sample of the population of all units of the same technology group. Therefore, the uncertainty in the mean emission rate among the 3 units should be based upon a sample size of 3 and not a sample size of 98. Bootstrap simulation with bootstrap samples of 3 synthetic data points was used to quantify uncertainty.

An example of results for uncertainty based upon the number of units actually in the inventory is shown in Figure 7 for the case of the NO_x emission rate. In comparing Figure 7 with Figure 6, it is apparent that the confidence intervals are much wider in the

former, corresponding to the smaller sample size. With a smaller number of units, the range of uncertainty is larger.

Figure 8 shows the sampling distribution of the mean emission inventory for the selected technology group. In this case, the emissions are from 3 units. The mean value of the inventory is 16,200 tons of NO_x emitted over a twelve-month period. The 95 percent probability range for this distribution is from 11,500 tons to 22,100 tons, or almost a factor of two range of uncertainty. Expressed on a relative basis, the 95 percent probability range for uncertainty is minus 29 percent to plus 37 percent with respect to the mean value. The range of uncertainty is slightly asymmetric, reflecting the fact that many of the inputs have skewed distributions. The range of uncertainty reflects the large amount of inter-unit variability in the inputs to the inventory and the small sample size (n=3).

The overall uncertainty in the total emission inventory, inclusive of all four technology groups considered, is shown in Figure 9. The estimated mean emission rate is 157,400 tons of NO_x emitted in a twelve-month period. The 95 percent probability range is enclosed by emissions of 132,200 tons and 186,600 tons. This is a range of -25,200 tons to +29,200 tons, or -16 percent to +19 percent, with respect to the mean. The asymmetry of the 95 percent probability range is a result of skewness in many of the input assumptions among the four technology groups.

A summary of the uncertainty results for the entire emission inventory is given in Table 2. Although the absolute range of uncertainty for the total inventory is greater than the absolute range of uncertainty for the selected technology group, the relative range of uncertainty is smaller. While this result may seem counter-intuitive, it occurs because the

uncertainty in emissions for each technology group is assumed to be statistically independent of the other technology groups. It is clear from previous analysis of the data that there is not a significant statistical dependence (e.g., correlation) between the distributions of inter-unit variability for any pairwise combination of emission factor, heat rate, and capacity factor within a technology group for averaging times of three months or greater.^{1.18} This is not to say that there could not be common causes that might systematically increase or decrease the average values of any of these factors within or between technology groups when comparing results for a well-defined geographic area from one time period to another, such as between one year and another year. However, such causes were judged not to be significant for this particular case study, since only one such time period was analyzed.

Examples of possible common causes that could affect averages include the influence of: (a) ambient conditions (e.g., temperature, humidity) on the average NO_x emission factor or the heat rate; (b) design features that might be common to multiple units that affect average emission factors and average heat rate; (c) operational and maintenance practices that might be common to multiple units at a single plant or within a particular utility that affect average emission factors and average heat rate; and (d) dispatching and system expansion considerations with respect to the role of all units in responding to statewide or system-wide changes in electricity demand, which affects average capacity factors. However, insufficient data were available to assess these possible common causes, and in many cases these causes would result in a relatively small impact (e.g., influence of annual ambient conditions on annual average heat rate). Common causes, to the extent that they are significant, would tend to systematically shift

the mean value of the emission factor, heat rate, and/or capacity factor from one time period to another, but would have less influence regarding the relative range of variation with respect to the mean or with regard to the relative range of uncertainty in the mean.

While it is possible that there could be some common causes that affect multiple units within some subgroup (e.g., in a particular utility), there was little evidence of this type of multi-modality in the data, with the possible exception of the uncontrolled dry-bottom furnace data base. Even in this latter case, there was some evidence of multi-modality only for the emission factor and not for the heat rate or capacity factor.

Therefore, there was little evidence of subgroups among all of the technology groups to suggest that differences in ambient conditions, operation, or maintenance could explain some of the inter-unit variability in the emission factors or the activity factors.

Furthermore, the case study is based upon data for a particular year and therefore does not take into account inter-annual variability in system dispatching. The analysis was done for a relatively long averaging time (e.g., six-months, and 12-months) such that the effects of short term variability in ambient conditions and in operating practices would be dampened compared to, say, an analysis based upon one hour averages. The analysis was done for a large enough geographic area that it would not be expected to have the same ambient conditions statewide at any given time. For these reasons, the role of common causes for this particular case study was judged not to be important, although it could be important in other case studies

A property of probabilistic simulations is that, in general, it is not possible to sum the values of selected percentiles of each model input to obtain an estimate of the same percentile of the model output. For example, the 2.5th percentile of the total emission

inventory, which is 132,200 tons, does not correspond to a sum of the 2.5th percentile of each of the four technology groups. However, for linear models, the sum of the means is usually the same as the mean of the sum, unless there is a correlation among the model inputs.¹⁷

Figure 10 shows the relative importance of uncertainty in emissions from individual technology groups with respect to overall uncertainty in the total emission inventory. Of the four technology groups, the dry-bottom, uncontrolled (DB/U) group has the strongest correlation with uncertainty in the total emission inventory, with a correlation coefficient of approximately 0.7. In contrast, the controlled dry-bottom boiler group (DB/LNB) has a correlation of approximately 0.2. Thus, any imperfections in the fitted distributions for this technology group are not likely to contribute significantly to errors in the estimated overall uncertainty. The sensitivity analysis results imply that the most effective way to reduce uncertainty in the overall emission inventory is to begin by reducing uncertainty in the estimated emissions from DB/U technology group.

5.0 CONCLUSIONS

This paper has demonstrated a general methodology to quantify variability and uncertainty in emission factors, activity factors, and emission inventories, with application to the example of utility power plant emissions. A prototype software environment for calculation of probabilistic emission inventories was developed to illustrate the methodology and to demonstrate the case study. The prototype enables a user to visualize, in the form of empirical probability distributions, the data used to develop the inventory. This is sharp contrast from typical emission inventory work, in which point estimate values of emission factors are used to calculate a single estimate of the inventory. The range of variability in the example datasets was shown to be large.

Although it is not possible to quantify all sources of uncertainty, it is important to quantify as many sources of uncertainty as is practical. The case study demonstrates that the range of uncertainty attributable to random sampling error is substantial. For individual technology groups, the range of uncertainty is as large as approximately plus or minus 30 percent, and for the total inventory the range of uncertainty is approximately plus or minus 15 percent. These ranges of uncertainty are likely to be substantially larger than measurement errors in the data for this particular source category. The case study is based upon a relatively large sample of continuous emission monitoring data. Therefore, it is likely that the data used in the case study are reasonably representative of actual emissions among the population of units for the technology groups studied. In this case, it is likely that random sampling error is the most important contributor to overall uncertainty. The specific results will differ for other emission source categories.

Inspection of the distributions of inter-unit variability reveals that emission factors can vary by a factor of four or more within a technology group. This wide range of variability suggests the possibility of cost-saving opportunities for low emitting units to sell emissions permits to high emitting units under possible future permit trading programs. Furthermore, decision-makers can use data such as this to estimate the probability that they would exceed allowable emissions under a permit trading framework and to make judgments regarding whether to buy or sell permits.

It is possible to have a high degree of certainty regarding recent actual emissions at power plants equipped with CEM equipment. However, it is not possible to have certainty regarding what the emissions will be at a future time, whether in the near or distant future. In estimating distant future emissions, an additional refinement that may

be needed in the case study would be to consider changes in capacity factor and the effects of capacity expansion, as well as the possible impacts of other common causes that might influence average emission factors and average heat rates. For relatively short term future estimates (e.g., a year or two into the future), the methodology employed may provide a reasonable estimate of absolute emissions. However, the relative range of uncertainty estimated using the methods presented here are likely to be indicative of the relative range of uncertainty in a future emission inventory, unless there is a large shift in the relative contributions of different technology groups to the total inventory.

In addition to quantifying the substantial range of uncertainty in the inventory, the case study demonstrates the capability to identify key sources of uncertainty in the inventory. The largest contribution to uncertainty comes from one technology group. Therefore, resources could be focused on collecting more or better data for the most sensitive technology group. Knowledge of key sources of uncertainty can also aid in identifying where it is not necessary to target additional data collection. For example, even though there were some discrepancies in the fit of parametric distributions to some of the data as shown in Figure 3, that particular technology group does not contribute substantially to uncertainty in the overall inventory. Therefore, there would not be a large benefit associated with improving the characterization of uncertainty for that particular input.

The quantification of uncertainty has many important implications for decisions. For example, it enables analysts and decision makers to evaluate whether time series trends are statistically significant or not. It enables decision makers to determine the likelihood that an emissions budget will be met. Inventory uncertainties can be used as

input to air quality models to estimate uncertainty in predicted ambient concentrations, which in turn can be compared to ambient air quality standards to determine the likelihood that a particular control strategy will be effective in meeting the standards. In addition, using probabilistic methods, it is possible to compare the uncertainty reduction benefits of alternative emission inventory development methods, such as those based upon generic versus more site-specific data. Thus, the methods presented here allow decision makers to assess the quality of their decisions and to decide on whether and how to reduce the uncertainties that most significantly affect those decisions.

ACKNOWLEDGMENTS

The authors acknowledge the support of the Office of Air Quality Planning and Standards (OAQPS) of the U.S. Environmental Protection Agency, which funded most of this work. Some support for the methodological components of this work was also provided via U.S. EPA STAR Grants Nos. R826766 and R826790. The authors appreciate the guidance and encouragement of Mr. Steve Rhomberg, formerly with U.S. EPA, and Ms. Rhonda Thompson of U.S. EPA. The authors also thank Mr. Zhen Xie for his contributions to the development of the internal database used in the AUVÉE prototype software. This paper has not been subject to any EPA review. Therefore, it does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

REFERENCES

1. Frey, H.C., R. Bharvirkar, J. Zheng; "Quantitative Analysis of Variability and Uncertainty in Emissions Estimation"; Final Report, Prepared by North Carolina State University for Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1999.
2. NRC, *Modeling Mobile Source Emissions*, National Academy Press, Washington, D.C., 2000.
3. NRC, *Science and Judgment in Risk Assessment*, National Academy Press: Washington, D.C., 1994.
4. U.S. EPA, Guiding Principles for Monte Carlo Analysis, EPA/630/R-97/001, U.S. Environmental Protection Agency, Washington, DC, May, 1997.
5. U.S. EPA, Report of the Workshop on Selecting Input Distributions for Probabilistic Assessment, EPA/630/R-98/004, U.S. Environmental Protection Agency, Washington, DC, January 1999.
6. U.S. EPA, *Summary Report for the Workshop on Monte Carlo Analysis*, EPA/630/R-96/010, Risk Assessment Forum, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. September, 1996.
7. Kini, M.D., and H.C. Frey, Probabilistic Evaluation of Mobile Source Air Pollution: Volume 1, Probabilistic Modeling of Exhaust Emissions from Light Duty Gasoline Vehicles, Prepared by North Carolina State University for Center for Transportation and the Environment: Raleigh, December 1997.
8. Pollack, A.K., P. Bhave, J. Heiken, K. Lee, S. Shepard, C. Tran, G. Yarwood, R.F. Sawyer, and B.A. Joy, "Investigation of Emission Factors in the California EMFAC7G Model. PB99-149718INZ, Prepared by ENVIRON International Corp, Novato, CA, for Coordinating Research Council, Atlanta, GA. 1999.
9. Bammi, S.; Frey, H.C., "Quantification Of Variability and Uncertainty in Lawn And Garden Equipment NO_x and Total Hydrocarbon Emission Factors,"

Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, June 2001.

10. Frey, H.C., and S. Li; "Quantification of Variability and Uncertainty in Natural Gas-fueled Internal Combustion Engine NO_x and Total Organic Compounds Emission Factors," Proceedings of the Annual Meeting of the Air & Waste Management Association, Orlando, FL, June 2001.
11. Rubin, E.S., M. Berkenpas, H.C. Frey, and B. Toole-O'Neil , "Modeling the Uncertainty in Hazardous Air Pollutant Emissions," *Proceedings, Second International Conference on Managing Hazardous Air Pollutants*, Electric Power Research Institute, Palo Alto, CA, 1993.
12. Frey, H.C. and Rhodes, D.S.; "Characterizing, Simulating and Analyzing Variability and Uncertainty: An Illustration of Methods Using an Air Toxics Emissions Example", *Human and Ecological Risk Assessment*. 2(4): 762-797 (1996).
13. Frey, H.C., and R. Bharvirkar, "Quantification of Variability and Uncertainty: A Case Study of Power Plant Hazardous Air Pollutant Emissions," in *The Risk Assessment of Environmental and Human Health Hazards: A Textbook of Case Studies*, D. Paustenbach, Ed., John Wiley and Sons: New York. In press, 2001
14. Rhodes, D.S., and H.C. Frey, "Quantification of Variability and Uncertainty in AP-42 Emission Factors Using Bootstrap Simulation," Emission Inventory: Planning for the Future (held October 28-30 in Research Triangle Park, NC), Air and Waste Management Association: Pittsburgh, Pennsylvania, October 1997, pp. 147-161.
15. Frey, H.C., and D.S. Rhodes, "Characterization and Simulation of Uncertain Frequency Distributions: Effects of Distribution Choice, Variability, Uncertainty, and Parameter Dependence," *Human and Ecological Risk Assessment*, 4(2):423-468 (April 1998).

16. Frey, H.C., and D.E. Burmaster, "Methods for Characterizing Variability and Uncertainty: Comparison of Bootstrap Simulation and Likelihood-Based Approaches," *Risk Analysis*, 19(1): 109-130, February, 1999.
17. Cullen, A.C., and Frey, H.C.; *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press: New York, 1999.
18. Frey, H.C, J. Zheng, "Quantitative Analysis of Variability and Uncertainty in Emissions Estimation: Development of a Database, Prototype Software and Example Case Study," Draft, Prepared by North Carolina State University for U.S. Environmental Protection Agency, Research Triangle Park, NC, January 2001.
19. Frey, H.C., and L.K. Tran, *Quantitative Analysis of Variability and Uncertainty in Environmental Data and Models: Volume 2. Performance, Emissions, and Cost of Combustion-Based NO_x Controls for Wall and Tangential Furnace Coal-Fired Power Plants*, Report No. DOE/ER/30250--Vol. 2, Prepared by North Carolina State University for the U.S. Department of Energy, Germantown, MD, April 1999.
20. Harter, L.H. "Another Look at Plotting Positions," *Communications in Statistical-Theoretical Methods*, 13(13):1613-1633, 1984
21. Hazen, A., "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply," *Transaction of the Americal Society of Civil Engineers*, 77:1539-1640, 1914.
22. Hahn, G.J., and S.S. Shapiro, *Statistical Models in Engineering*, John Wiley and Sons, New York, 1967.
23. Cohen, A.C., and B. Whitten, *Parameter Estimation in Reliability and Life Span Models*, M. Dekker: New York, 1988.
24. Ang A. H.-S., and W. H. Tang, *Probability Concepts in Engineering Planning and Design, Volume 2*, John Wiley and Sons, New York.

25. D'Agostino, R.B., and M.A. Stephens, eds. , *Goodness-of-Fit Techniques*, M. Dekker: New York, 1986.
26. Morgan, M.G., and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York, 1990.
27. Hattis, D., and D.E. Burmaster, "Assessment of Variability and Uncertainty Distributions for Practical Risk Analyses," *Risk Analysis*, 14(5):713:729, 1994.
28. Seiler, F.A., and J.L. Alvarez, "On the Selection of Distributions for Stochastic Variables," *Risk Analysis*, 16(1):5-18, 1996.
29. Efron, B.; Tibshirani, R. J., *An Introduction to Bootstrap*; Chapman and Hall: New York 1993.
30. Frey, H.C., J. Zheng, *User's Guide for the Prototype Software for Analysis of Variability and Uncertainty in Emissions Estimation (AUVÉE)*, Prepared by North Carolina State University for the U.S. Environmental Protection Agency, Research Triangle Park, NC, 2000.

Table 1. Summary of 12-Month NO_x Emission and Activity Factors and of Fitted Distributions for Five Power Plant Technology Groups

Technology Group	Input Variables	Summary of Data			Fitted Distributions		
		Number of Data Points	Mean	Standard Deviation	Distributions	1 st parameter*	2 nd parameter*
DB/U	Heat Rate	84	11,150	1,450	Lognormal	9.31	0.124
DB/U	Capacity Factor	84	0.53	0.19	Beta	3.30	2.89
DB/U	NO _x Emission Factor	84	293	83	Weibull	323	4.22
DB/LNB	Heat Rate	98	10,610	890	Lognormal	9.27	0.0792
DB/LNB	Capacity Factor	98	0.67	0.14	Beta	6.94	3.36
DB/LNB	NO _x Emission Factor	98	177	41	Gamma	18.7	9.48
T/U	Heat Rate	134	10,780	1,290	Lognormal	9.28	0.113
T/U	Capacity Factor	134	0.56	0.18	Beta	3.62	2.84
T/U	NO _x Emission Factor	134	198	54	Gamma	13.5	14.8
T/LNC1	Heat Rate	36	10,730	790	Lognormal	9.28	0.0746
T/LNC1	Capacity Factor	36	0.65	0.20	Beta	3.11	1.70
T/LNC1	NO _x Emission Factor	36	161	37	Gamma	18.5	8.70
DTF/OFA	Heat Rate	6	10,360	900	Lognormal	9.24	0.058
DTF/OFA	Capacity Factor	6	0.66	0.07	Normal	0.664	0.0660
DTF/OFA	NO _x Emission Factor	6	191	17	Gamma	127	1.50

* 1st parameter is the mean for Normal distribution, the geometric mean for the Lognormal distribution, scale parameter for the Gamma and the Beta distribution, and the shape parameter for the Weibull distribution.

* 2nd parameter is the standard deviation for the Normal distribution, the geometric standard deviation for the Lognormal distribution, and the shape parameter for Weibull, Gamma and Beta distributions.

Table 2. Summary of Uncertainty Results for the Emission Inventory Case Study

Technology Group	2.5 th Percent	Mean	97.5 th Percentile	Random Error (%) ^a	
				Negative	Positive
DB/U	40,200	56,900	75,100	-29	+32
DB/LNB	11,500	16,200	22,100	-29	+37
T/U	27,600	37,100	50,800	-26	+37
T/LNC1	33,400	47,200	62,300	-29	+32
Total	132,200	157,400	186,600	-16	+19

^a Results shown are the relative uncertainty ranges for a 95 percent probability range, given with respect to the mean value.

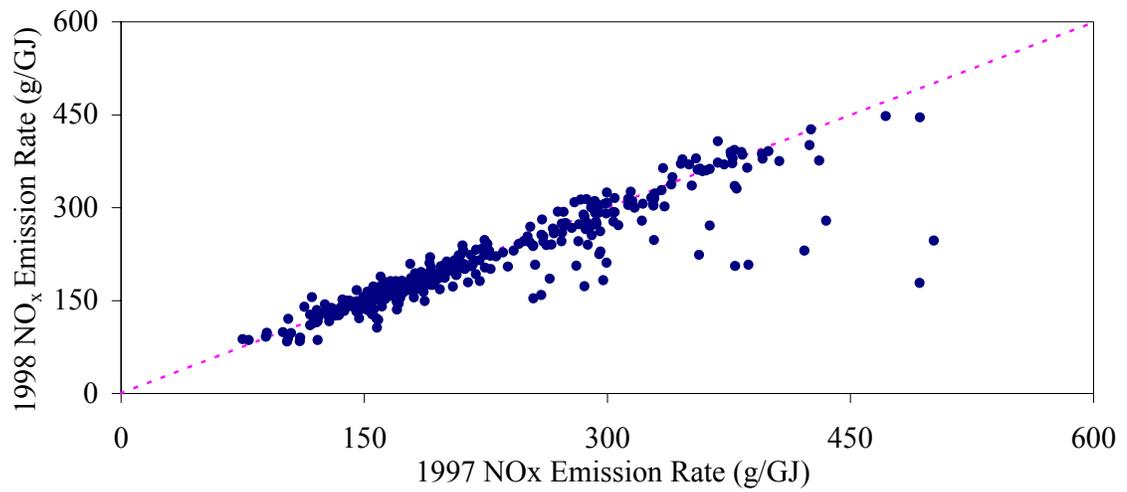


Figure 1. Scatter plot of 12-month NO_x Emission Rate of 1997 and 1998 (No. of Data=390)

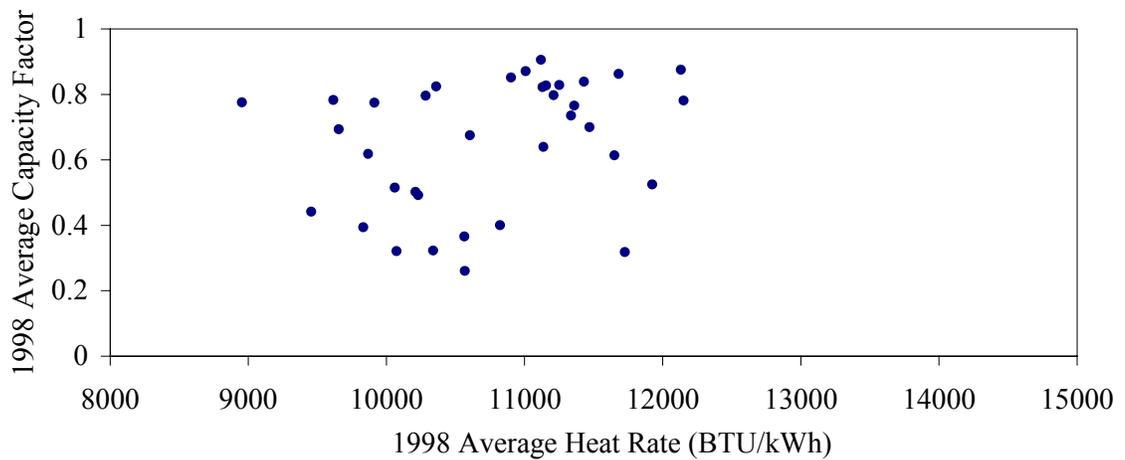


Figure 2. Scatter Plot for 12-month Average Heat Rate versus 12-month Average Capacity Factor for Tangential-Fired Boilers Using Low NO_x Burners and Overfire Air Option 1. (n=36)

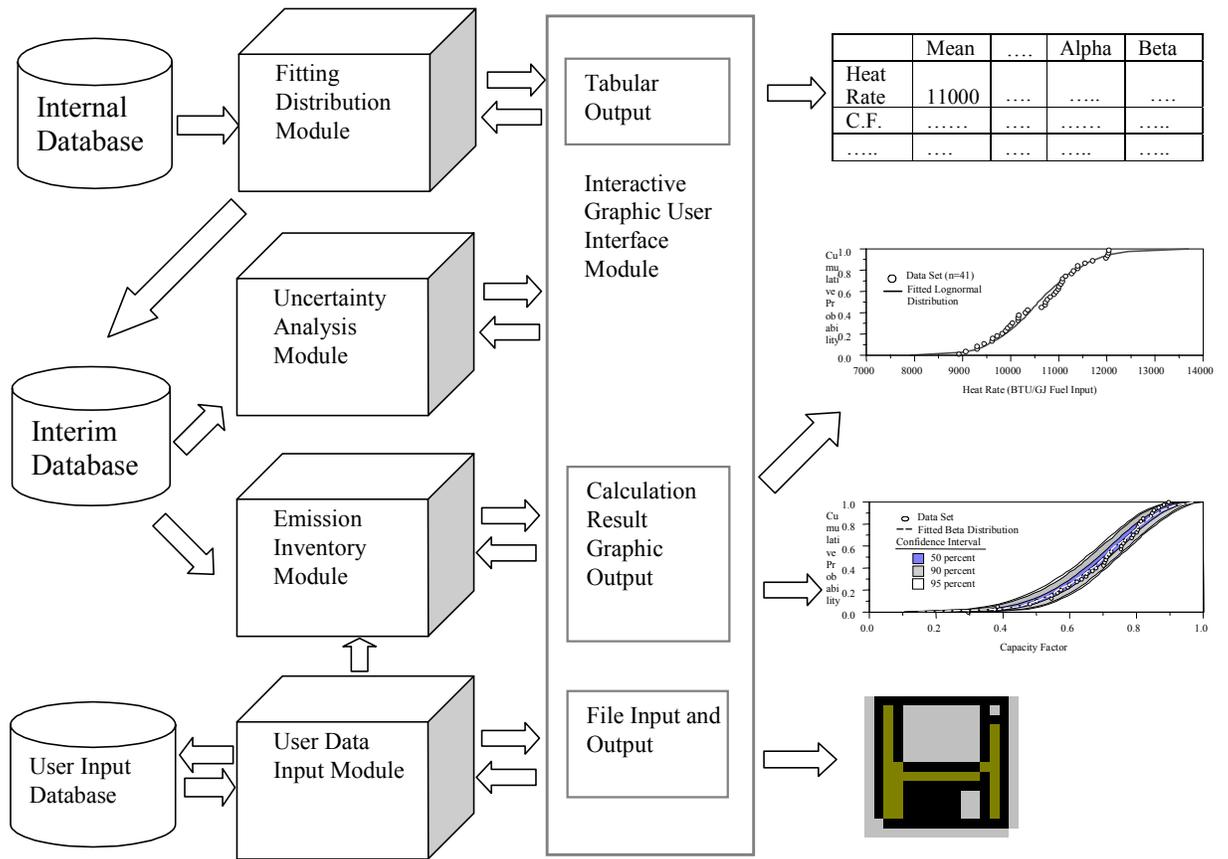


Figure 3. Conceptual Design of the Analysis of Uncertainty and Variability in Emissions Estimation (AUVEE) Prototype Software.

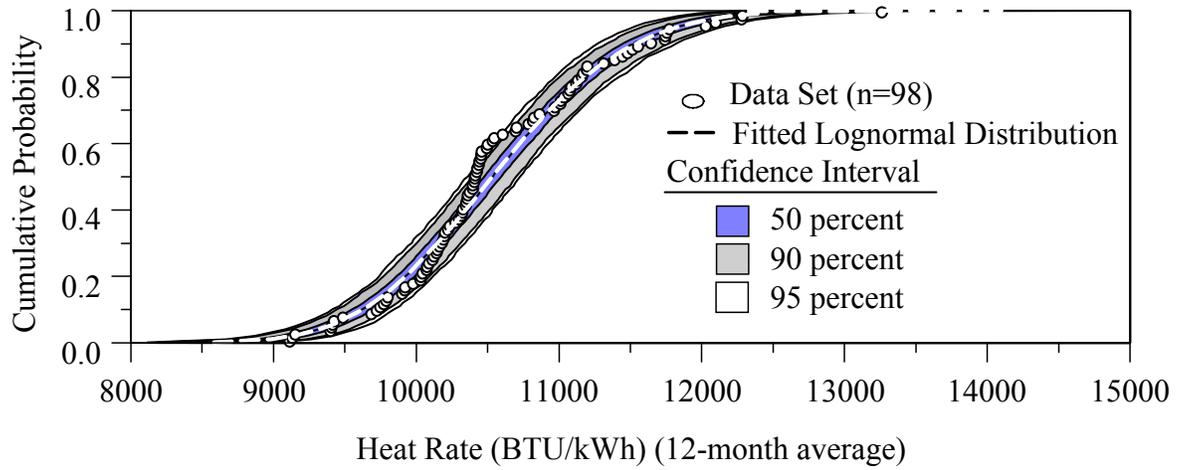


Figure 4. Probability Band for Distribution Fitted to Example Heat Rate Data for Dry Bottom Wall-fired Boilers Using Low NO_x Burners (n=98)

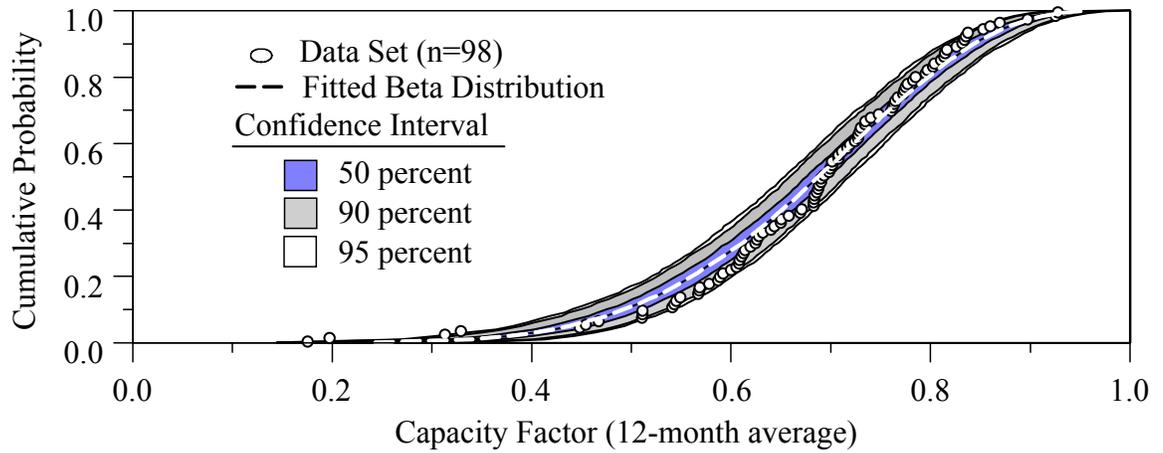


Figure 5. Probability Band for Distribution Fitted to Example Capacity Factor Data for Dry Bottom Wall-fired Boilers Using Low NO_x Burners (n=98)

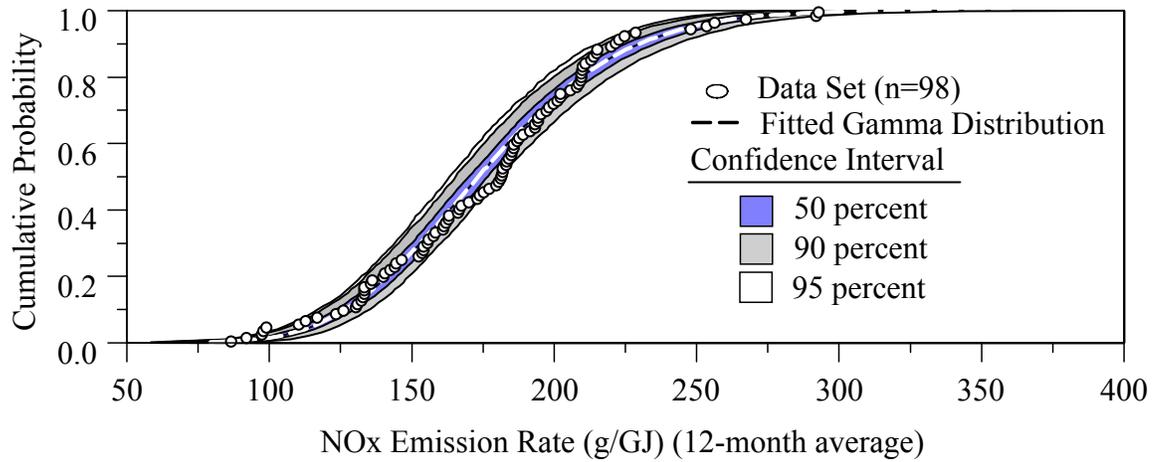


Figure 6. Probability Band for Distribution Fitted to Example NO_x Emission Rate Data for Dry Bottom Wall-fired Boilers Using Low NO_x Burners (n=98)

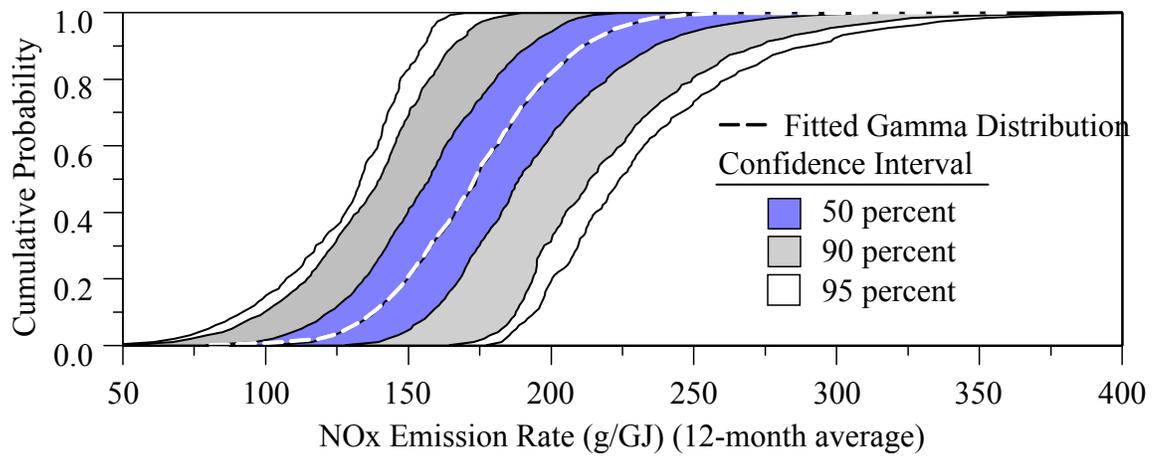


Figure 7. Probability Bands Based Upon Number of Units in the Emission Inventory (n=3) for the Example of NO_x Emission Rate.

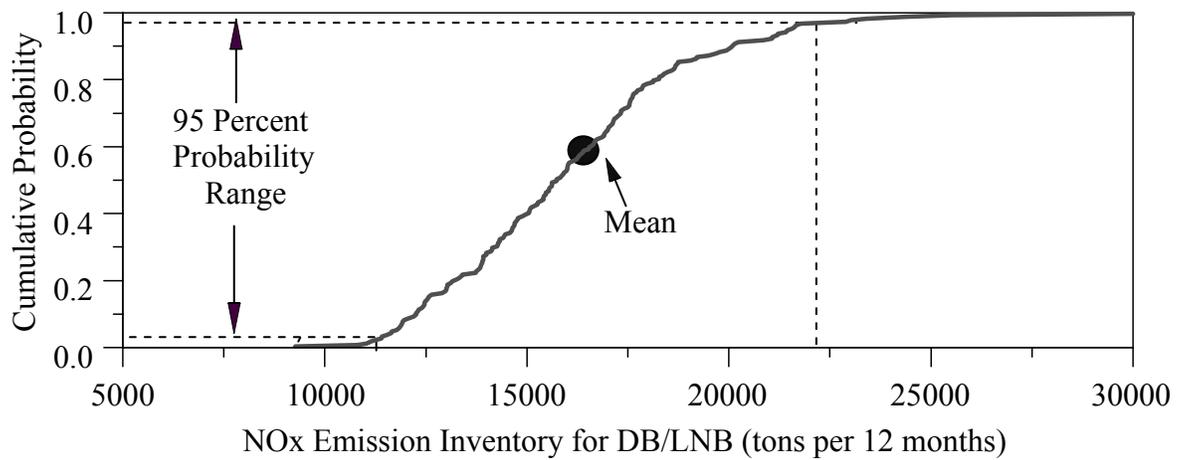


Figure 8. Uncertainty in a 12-Month NO_x Emission Inventory for an Individual Technology Group Comprised of 3 Units.

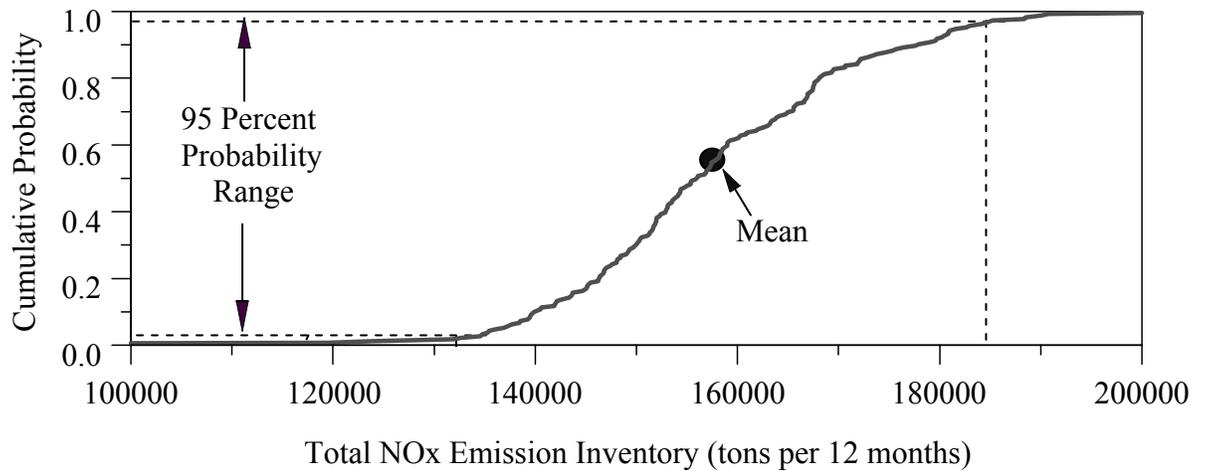


Figure 9. Uncertainty in a 12-Month NO_x Emission Inventory Inclusive of Four Technology Groups.

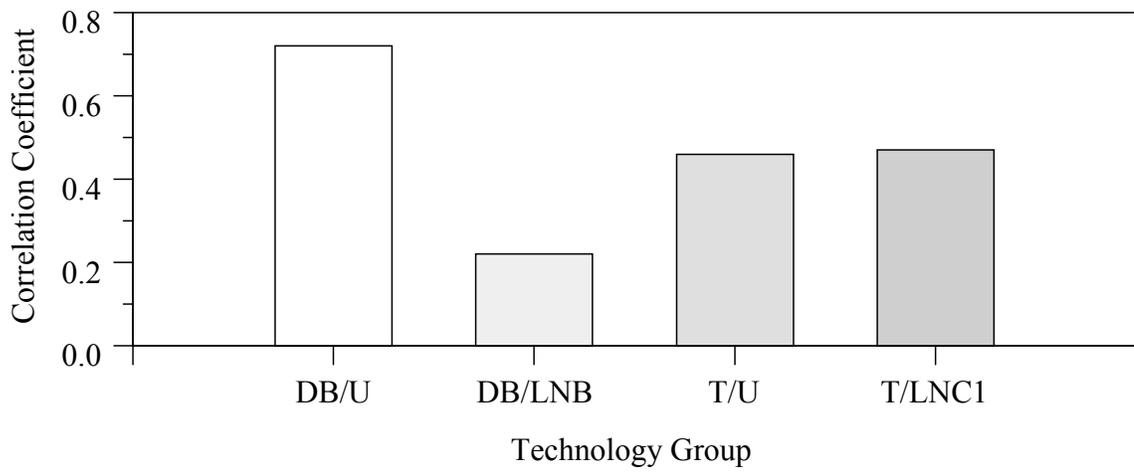


Figure 10. Relative Importance of Uncertainty in Emissions from Individual Technology Groups with Respect to Overall Uncertainty in the Total Emission Inventory

About the Authors

H. Christopher Frey is an Associate Professor and Mr. Junyu Zheng is a Graduate Research Assistant, both of the Department of Civil Engineering, Campus Box 7908, North Carolina State University, Raleigh, NC 27695-7908.

PART VII

**PROBABILISTIC ANALYSIS OF DRIVING CYCLE-
BASED HIGHWAY VEHICLE EMISSION FACTORS**

H. Christopher Frey and Junyu Zheng

Reviewed and Resubmitted to
Environmental Science and Technology

Probabilistic Analysis of Driving Cycle-Based Highway Vehicle Emission Factors

H. Christopher Frey* and Junyu Zheng
Department of Civil Engineering
North Carolina State University
Campus Box 7908
Raleigh, NC 27695-7908

Email: frey@eos.ncsu.edu
Tel: (919) 515-1155
Fax: (919) 515-7908

ABSTRACT

A probabilistic methodology for quantifying inter-vehicle variability and fleet average uncertainty in highway vehicle emission factors is developed. The methodology features the use of empirical distributions of emissions measurement data to characterize variability and the use of bootstrap simulation to characterize uncertainty. For the base emission rate as a function of mileage accumulation under standard conditions, a regression-based approach was employed in which the residual error terms were included in the probabilistic analysis. Probabilistic correction factors for different driving cycles, ambient temperature, and fuel Reid Vapor Pressure (RVP) were developed without interpolation or extrapolation of available data. The method was demonstrated for tailpipe carbon monoxide, hydrocarbon, and nitrogen oxides emissions for a selected light duty gasoline vehicle technology. Inter-vehicle variability in emissions was found to span typically two or three orders-of-magnitude. The uncertainty in the fleet average emission factor was as low as plus or minus 10 percent for a 95 percent probability range, in the case of standard conditions, to as much as minus 90 percent to plus 280 percent

when correction factors for alternative driving cycles, temperature, and RVP are applied. The implications of the results for methods selection and for decision making are addressed.

Key Words: Variability, Uncertainty, Mobile5b, Emission Factors, Highway Vehicle

1.0 INTRODUCTION

The National Research Council (NRC) recommends that efforts be conducted to quantify uncertainties in highway emission estimates (1). Such estimates are widely used at the state and federal level for regulatory, planning, and other decision making purposes involving substantial resources (2). Thus, there is incentive to understand the range of uncertainty in the estimates and to make air quality management decisions that are robust to uncertainty.

Kini and Frey (3) and Pollack et al. (4) have reported results for probabilistic analysis pertaining to aspects of the U.S. Environmental Protection Agency's (EPA's) Mobile5b and the California Air Resources Board EMFAC7F emission factor models, respectively. Both studies focused on a bottoms-up approach to assessing uncertainty in emission factors based upon statistical analysis of emission test data used to develop the model. Others (5) used a bootstrap approach to calculate confidence intervals for the speed correction factor in Mobile5a, but retained the functional form of a curve fit employed by US EPA in their analysis.

Compared to the Kini and Frey study, this paper introduces additional methodological tools required to deal with correction factors for which only relatively small datasets are available, with case study examples for temperature and Reid vapor pressure (RVP) corrections. To demonstrate the method and insights obtained from it,

detailed estimates of uncertainty are provided for a Light Duty Gasoline Vehicle (LDGV) technology group for carbon monoxide (CO), nitrogen oxides (NO_x), and hydrocarbon (HC) emission factors for each of 11 driving cycles at standard temperature and RVP, and for situations in which emissions are corrected for other values of temperature and RVP.

The case study for this paper is based upon the Mobile5b emission factor model (6-9). A new emission factor model, Mobile6 is expected to be publicly released soon (10). Both models share similar approaches regarding the use of multiplicative correction factors to adjust a base emission rate to non-standard conditions. Therefore, the methodological issues regarding uncertainty analysis are similar for both models. However, Mobile6 will typically provide different emission factor estimates than were obtained with Mobile5b (11). A key question is regarding whether such differences are significant. In order to answer this question, it is critically important to understand the range of uncertainty inherent in the Mobile5b emission factor estimates.

1.1 Sources of Variability and Uncertainty

Variability refers to the heterogeneity across different element of a population over time or space. Uncertainty is a lack of knowledge about the true value of a quantity. Uncertainty in emissions are typically attributable to (1) random measurement errors (lack of precision); (2) systematic errors (bias or lack of “accuracy”) such as would be caused by imprecise calibration or use of surrogate data (e.g., laboratory tests of vehicles rather than on-road measurements); (3) lack of empirical basis such as would occur when measurements have not been taken or when estimating emissions for a future source; and (4) human error, such as random mistakes in entering or processing data. Variability can be represented by a frequency distribution. Uncertainty can be quantified as a probability

distribution representing the likelihood that the unknown quantity falls within a given range of values (12-19).

1.2 Variability and Uncertainty in Highway Vehicle Emission Factors

Emissions vary from one vehicle to another because of differences in design, operation, maintenance, and fuel composition. Emissions measurements using specific driving cycles attempt to control for operation by imposing a specific speed versus time profile and for fuel composition. However, variations in vehicle operation within the allowable tolerances of the test procedure can lead to as much as an order-of-magnitude difference in emissions (20). Several researchers describe the inherent variability in emissions measurements obtained using a variety of testing methods (21-23). The main focus of the uncertainty analysis is on characterizing random and systematic errors associated with estimates of fleet average emissions. Random errors are characterized based upon statistical analysis of random sampling error. Systematic errors are characterized based upon deviations of the point estimate predictions of the model when compared to mean values inferred directly from available data.

1.3 Modeling Assumptions and Input Data

The methodological approach includes the following elements: (1) development of a simplified empirical emission factor model, similar to that of Mobile5b; (2) collection of emission test data for an example case study; and (3) probabilistic analysis and modeling techniques. The first two are described here, and the third is presented with the case study results.

1.4 Brief Review of the Mobile5b Model.

Mobile5b estimates emission factors for CO, HC, and NO_x by calculating a base emission rate (BER) for a standard driving cycle and standard conditions (e.g., ambient

temperature, RVP) associated with a given mileage accumulation (odometer reading). The BER is adjusted to other conditions, such as different driving cycles, ambient temperature, and RVP, using correction factors (7-10). The BER is developed separately for different vehicle types (e.g., light duty gasoline vehicles) based upon an assumed mix of technology groups, the latter of which are typically characterized based upon fuel delivery and emission control systems (e.g., throttle-body injected engines, three way catalysts). Emission control systems are assumed to undergo "deterioration" as a function of mileage accumulation. Curve fits are used in Mobile5b for the BER and for each correction factor. The curve fits are typically based upon regression analysis of driving cycle data. Mobile5b determines point estimates for each step in the calculation process.

1.5 Simplified Probabilistic Emission Factor Model

The BER in Mobile5b is intended to represent emissions for Bag 2 of the FTP driving cycle, which is taken as the reference point to which correction factors are applied. However, in order to obtain a large data set representative of the on-road vehicle fleet, EPA used inspection and maintenance program data obtained using the IM240 test procedure. The IM240 test is based upon a portion of the speed profile used in the FTP. EPA developed a regression equation for each of CO, HC, and NO_x emissions to convert the IM240 measurements to an equivalent FTP emission estimate. EPA used logarithmic transformations to develop the IM240 to FTP regression equations for CO and HC, and used a linear formulation in the case of NO_x. EPA did not account for the residual error term of the regression equations, which reflects the inter-vehicle variability in emissions that is not explained by the model. The residual error term is multiplicative for CO and

HC, and it is additive for NO_x, because of the formulations assumed. Here, the residual errors were characterized as empirical distributions based upon analysis of the data sets used by EPA to develop the IM240 to FTP regression models.

The estimated FTP emissions were used by EPA to develop a linear regression equation for emissions versus mileage accumulation for each of the three pollutants, thereby introducing a second residual error term. However, the residual error term was not normally distributed, which violates the assumption of least squares regression. The residuals for the logarithm of emissions were more nearly normally distributed. Therefore, a log-linear regression is used here instead, which differs from the approach used by EPA. For CO and HC, the BER equation is:

$$BER = \exp\{ZML + DR \times MA + \varepsilon_1\} \varepsilon_2 \quad (1)$$

For NO_x emissions, the BER equation is:

$$BER = \exp\{ZML + DR \times MA + \varepsilon_1\} + \varepsilon_2 \quad (2)$$

Where :

- BER* = Base Emission Rate (grams/mile)
- ZML* = Zero Mile Level emission constant (logarithm of g/mi)
- DR* = Deterioration Rate constant for mileage less than or equal to 50,000 miles (logarithm of g/mi²)
- MA* = Mileage Accumulation less than or equal to 50,000 miles (miles)
- ε_1 = Residual error distribution for the BER regression equation (logarithm of g/mi)
- ε_2 = Residual error distribution for the IM240 to FTP regression equation dimensionless ratio for CO and HC, g/mi for NO)

The BER represents emissions for the FTP driving cycle under standard conditions, including ambient temperature of 75 °F and fuel RVP of 9.

The emission factor for non-standard conditions is estimated using multiplicative correction factors, each of which is a dimensionless ratio. Three correction factors are evaluated empirically based upon data analysis: (1) speed correction factor (SCF); (2) temperature correction factor (TCF); and (3) RVP correction factor (RVPCF):

$$EF = BER \times SCF \times TCF \times RVPCF \quad (3)$$

The SCF is the ratio of emissions on a non-FTP driving cycle to the emissions on the FTP driving cycle. The TCF is the ratio of emissions at ambient temperature T on an FTP test to the emissions on the standard FTP test, which has a temperature of 75 °F. The RVPCF is the ratio of emissions for a non-standard RVP to that of the standard RVP of 9.0, with both evaluated using the FTP test.

The functional form of equation (3) is similar to but not the same as that in Mobile5b. The Mobile5b model employs curve fits for correction factors, and the RVPCF curve fit includes temperature as an explanatory variable. In the approach used here, curve fits are not employed so as to avoid introduction of systematic errors associated with any particular model formulation.

1.6 Collection of Emission Test Data

Emissions test data used by EPA to develop the BER and the correction factors were obtained from EPA. Data for the case study are based upon LDGV Technology Group 8, which have a throttle body fuel injection system and a three-way catalyst and are typical of many on-road vehicles. These data were analyzed to determine empirical correction factors specific to each driving cycle, temperature, and RVP represented in the database. The data used for the SCF analysis involved measurements of multiple vehicles on multiple driving cycle tests. The tests include Bag 2 of the FTP, as well as the LSP1, LSP2, LSP3, NYCC, SCC12, SCCC36, HFET, HSP1, HSP2, and HSP3 test

procedures. Each procedure is characterized by a different speed trace. The average speeds vary from 2.5 mph for LSP1 to 64 mph for HSP3. For Technology Group 8, a set of 35 vehicles were tested on each of the NYCC, SCC12, FTP Bag 2, SCC36, and HFET procedures. Fourteen of the vehicles were also tested on the LSP1, LSP2, and LSP3 cycles. Eight of the vehicles were tested also on the HSP2 and HSP3 cycles, while four were tested on the HSP1 cycle. The ratio of each vehicle's emissions on a non-standard cycle to its emissions on the FTP Bag 2 cycle were calculated, and an empirical distribution of the inter-vehicle variability in the ratio was developed. The uncertainty in the average SCF was characterized based upon the sampling distribution of the mean, which is influenced both by the sample size and by variability.

The available data sets for estimating variability and uncertainty in the TCF and RVPCF are much smaller than for the SCF. For a selected set of typically only three or four vehicles, several repeated FTP tests were run at the standard temperature of 75 °F, and then several repeated FTP tests were run at a different temperature, such as 50 °F.

In developing empirical correction factors, extrapolation of the actual test data is avoided by considering conditions only for which test data are available. These include temperatures of both 75 °F and 50 °F at RVP = 9 psi, and a temperature of 50 °F at RVP = 13 psi. Thus, a temperature correction factor is first applied to represent the conditions of lower temperature, and then an RVP correction factor is applied to represent conditions of high RVP at the lower temperature.

2.0 QUANTIFICATION OF INTER-VEHICLE VARIABILITY IN CORRECTION FACTORS

Inter-vehicle variability in correction factors was estimated by sampling from the available emissions measurement data to construct cumulative probability distribution

functions. In the case of the speed correction factors, there typically was sufficient data to construct empirical distributions of inter-vehicle variability. In contrast, for the TCF and RVPCF, data were available only for three or four vehicles that were tested under both standard and non-standard conditions. For each vehicle, typically three or four replicate measurements were made for each condition. Thus, there is intra-vehicle variability reflected by differences in the replicate measurements for a given vehicle, and there is inter-vehicle variability reflected by differences in average measurements when comparing vehicles. A distribution for variability for a single vehicle was developed by analyzing the replicate measurements for that vehicle. A combined distribution of variability among the three or four vehicles was developed by combining their individual distributions into a single mixture distribution. The observed intra-vehicle variability could also be interpreted as variability in emissions associated with differences in operation, which is a factor that contributes to inter-vehicle variability. Even though vehicles are tested on with respect to a standard speed profile, the test driver is allowed to deviate from the speed trace within a tolerance, and such deviations can lead to variability in emissions (20). Thus, the combined mixture distribution is interpreted as an indication of overall inter-vehicle variability.

The development of the mixture distribution includes the following tasks: (1) simulation of a distribution of variability for an individual vehicle; (2) development of weighting factors for each vehicle; and (3) simulation of a mixture distribution including all available vehicles.

For the first task, for a given vehicle, one of the replicate measurements at non-standard conditions is sampled, with replacement, as is one of the measurements at

standard conditions. The ratio of the two describes one possible correction factor for the given vehicle. If there are m data points obtained at standard conditions, the probability of sampling one of the measurements is $1/m$. Similarly, if there are n data points obtained at non-standard conditions, the probability of sampling one of the measurements is $1/n$. The process of randomly selecting one random non-standard measurement and one random standard measurement, and calculating one possible correction factor value, was repeated 1,000 times. From these 1,000 estimates of correction factor values, an empirical distribution of the correct factor is characterized. This process is repeated separately for each vehicle in the database.

For the second task, there is not a unique basis for assigning weights to each vehicle. As a default, it was assumed that each vehicle is equally representative of the on-road fleet. Therefore, if the number of vehicles is v , the weight assigned to each vehicle is $1/v$.

The third task involves combining the individual distributions of variability for each vehicle into a single continuous distribution. The weight of $1/v$ assigned to each vehicle refers to the proportion of total samples that are drawn from that individual vehicle's distribution when constructing the combined mixture distribution. For example, if there are three vehicles, one-third of the 1,000 simulated correction factors in the combined distribution will be obtained, at random, from the distribution for the first vehicle. Similarly, one third of the simulated correction factors will be obtained from each of the other two vehicles.

As an example, the individual and mixture CO TCF distributions for three vehicles are shown in Figure 1. The distributions are shown in the form of empirical

cumulative distribution functions. The step-wise nature of the CDFs reflects the finite number of possible combinations of correction factors if m and n are approximately 3 or 4 each. The lower tail of the composite distribution asymptotically approaches the lower tail of the distribution that has the lowest values of TCF, and the upper tail of the composite asymptotically approaches the upper tail of the distribution which has the highest values. In this case, vehicles 608 and 609 have similar variability, whereas vehicle 304 has a wider range of variability than the other two. Like each of the three individual vehicle distributions, the combined distribution has approximately a 60 percent probability of values less than 1.0, and, conversely, approximately a 40 percent probability of values greater than one. The correction factor varies from approximately 0.13 to 7.3 over a 95 percent probability range, which is a span of more than one order-of-magnitude. The mean value is 1.43. This implies that, on average, emissions of CO are expected to increase by 43 percent if temperature decreases from 75 °F to 50 °F. However, it is possible that emissions may decrease as much as 90 percent or they may increase by a factor of 5 or more, depending on the specific vehicle and operating conditions. The TCF for HC varies from 0.18 to 4.1, and for NO_x varies from 0.30 to 2.7. The mean TCF for HC is 1.22 and for NO_x is 1.06.

For the CO RVPCF, the average value varies from approximately 1.49 to 4.85 among the three vehicles, with a weighted average of 2.73. The range of inter-vehicle variability is from approximately 0.23 to a value of 9.8 over a 95 percent probability range, which is a range of more than an order-of-magnitude. For HC, the RVP correction factor varies from 0.13 to 11.4, and for NO_x it varies from 0.35 to 3.5. Thus, in all three

cases there is a possibility that emissions may be lower, but on average it is expected that emissions will be higher if a higher RVP fuel is used.

3.0 QUANTIFICATION OF UNCERTAINTY IN MEAN CORRECTION FACTORS

When developing emission inventories for motor vehicles, average emission factors for the on-road fleet are more useful than emission rates for individual vehicles. The mean is a statistic calculated from a random sample of data; therefore, it is a random variable. A probability distribution for a statistic is referred to as a sampling distribution. Under idealized conditions, the sampling distribution for the mean can be approximated with a normal distribution if the sample size is sufficiently large and/or if there is a sufficiently small range of variability in the data. However, the data sets used by EPA for developing correction factors are small, and there is a high degree of variability and positive skewness in the data sets. Therefore, it is not reasonable to assume normality for the sampling distribution of the mean. Instead, the numerical technique of bootstrap simulation is employed to estimate the sampling distribution of the mean. Bootstrap simulation was introduced by Efron in 1979 for the purpose of estimating confidence intervals for statistics (13). Bootstrap simulation does not require any assumptions regarding the shape of the sampling distribution.

The version of bootstrap simulation employed here involves randomly simulating, with replacement, a dataset of the same sample size as the original data set to create a bootstrap sample, which is a randomized version of the original data set. For each bootstrap sample, a replicate of the statistic of interest (e.g., mean) is calculated. The process is repeated many times to obtain multiple randomized estimates of the statistic. Typically 200 bootstrap replications is sufficient to estimate confidence intervals (13).

However, since the intent here is to estimate the sampling distribution, 1,000 replications are used.

For the TCF and RVPCF cases, the number of measurements under non-standard conditions, m , may not be the same as the number of measurements under standard conditions, n . So as not to underestimate the uncertainty associated with random sampling error, we let $k = \min(m,n)$ where k is the effective sample size. Thus, each bootstrap sample is comprised of k alternative estimates of the correction factor, which are obtained by randomly sampling k pairs of the measurements at the standard and nonstandard conditions and calculating one possible value of the correction factor for each random pair of measurements. The bootstrap replication of the mean is obtained by calculating the mean of the k alternative correction factor estimates. This process is repeated 1,000 times to yield an empirical CDF of the replicated mean values for a given vehicle. The bootstrap simulation is repeated for each vehicle to create sampling distributions of the mean for individual vehicles. A combined mixture distribution of mean values is developed in a manner similar to that for variability.

Figure 2 illustrates the results of analysis of the mean values of the CO TCF, showing the individual distributions for uncertainty in the mean of each of three vehicles, and the equally weighted mixture of all three. The range of uncertainty in the average correction factor is from approximately 0.38 to 3.5 over a 95 percent probability range. Thus, the range of uncertainty in the mean is less than the range of inter-vehicle variability. However, because only three vehicles were used in the testing, and because only a small number of tests were conducted at each temperature (typically only 3 or 4), the range of uncertainty in the mean is substantial and spans a factor of approximately 10.

The 95 percent probability range of uncertainty in the mean TCF for NO_x is from 0.59 to 1.6 and for HC it is from 0.44 to 2.6. Thus, there is greater relative range of uncertainty in the mean correction factors for CO and HC than there is for NO_x. Furthermore, in all three cases there is a chance that the correction factor could be less than 1, indicating a possibility that average emissions may decrease.

For RVPCF for CO, the mixture of the three individual vehicle distributions has a range of uncertainty from approximately 0.58 to 7.2. The mean value of the mixture distribution is approximately 2.7. The overall uncertainty in the mean correction factor for NO_x varies from approximately 0.7 to 2.8, and for HC varies from approximately 0.56 to 6.2. Thus, there appears to be relatively more uncertainty for CO and HC than for NO_x.

4.0 QUANTIFICATION OF VARIABILITY AND UNCERTAINTY IN THE EMISSION FACTORS

The emission factors for CO and HC are calculated using Equations (1) and (3), and for NO_x they are calculated using Equations (2) and (3). To estimate the inter-vehicle variability in emission factors, distributions for inter-vehicle variability are used for the residual errors and correction factors as summarized in Tables S-1, S-3, and S-5 in the Supplemental Information for CO, NO_x, and HC, respectively. A software package, Analytica™, was used to simulate variability in the tailpipe emission factors using Monte Carlo simulation.

Fleet average uncertainty in LDGV tailpipe emission factors is calculated in a manner similar to that for inter-vehicle variability, except that sampling distributions for mean values are used instead of frequency distributions for inter-vehicle variability. The distributions for fleet average uncertainty used for the inputs to the emission factor model

are summarized in Tables S-2, S-4, and S-6 in the Supplemental Information for CO, NO_x, and HC, respectively.

For each of the three pollutants, three sets of results were developed, representing different combinations of probabilistic assumptions. Each of the latter differ regarding the combination of correction factors that were treated probabilistically. All three include probabilistic assumptions for the base emission rate. One set is based upon the use of only the speed correction factor. The second set is based upon adjustment of emissions for an ambient temperature of 50° F, using the probabilistic temperature correction factor. The third set is based upon additional adjustment of the emission factors for a fuel RVP of 13, using the probabilistic RVP correction factor. These three sets of results are presented for both inter-vehicle variability in emissions and for fleet average uncertainty in highway vehicle CO emissions in Tables 1 and 2, respectively. Results for inter-vehicle variability and fleet average uncertainty are given for HC in Tables S-7 and S-8, respectively, and for NO_x in Tables S-9 and S-10, respectively, in the Supplemental Information.

4.1 Inter-Vehicle Variability in Emission Factors.

Table 1 contains CO emission factor estimates for 11 driving cycles. For each set of results, a deterministic point estimate is reported. The point estimate is obtained based upon the methods and assumptions of Kini and Frey (3). A 95 percent probability range, bounded by the 2.5th percentile and 97.5th percentile values, is reported based upon quantitative analysis of inter-vehicle variability. In addition, the mean value of each probabilistic simulation is reported.

Selected results are discussed here to illustrate the types of findings obtained from the probabilistic analysis. For example, for the low average speed LSP1 cycle, the 95

percent probability range of inter-vehicle variability is from 1.27 g/mi to 241 g/mi when considering variability only in the BER and SCF. This range is more than two orders-of-magnitude. The mean value of 44.1 grams per mile is approximately a factor of 40 greater than the 2.5th percentile value. The point estimate is based upon an analysis in which the skewness of the residual error term of the IM240-to-FTP regression model is not considered, which is similar to the approach used in the Mobile5b emission factor model. Furthermore, the point estimate is based upon a curve fit used by EPA for the speed correction factor. There are biases in the point estimate based upon the use of a speed correction factor curve fit and failure to properly account for residual errors in the Mobile5b emission factor model.

When the temperature correction factor is applied, the estimated range of variability increases. For example, for the case of CO emissions on the LSP1 cycle, the predicted 95 percent probability range for variability is from 0.63 g/mi to 377 g/mi, which is substantially wider than the range of variability at the standard ambient temperature. When the emissions estimate is adjusted for a fuel RVP of 13 psi instead of 9 psi, the 95 percent range of variability increases to an interval from 0.55 g/mi to 1040 g/mi.

Similar trends regarding the increase in estimated inter-vehicle variability resulting from the application of additional correction factors are observed for the other driving cycle emission estimates for CO tailpipe emissions. These trends are also observed for HC and NO_x emission factor estimates. The effects of variability in TCF and RVPCF on variability in emission factors are substantial.

4.2 Uncertainty in Mean Emission Factors

Probabilistic estimates of fleet average uncertainty in CO emission factors are summarized in Table 2. A 95 percent probability range is reported for each emission factor estimated, as given by the 2.5th and 97.5th percentile values for the mean. An average estimate of the mean is also given. In addition, systematic and random errors are reported. The systematic error is the point estimate minus the mean. In general, the mean values tend to be higher than the point estimates obtained using the same deterministic modeling methodology as employed by Mobile5b. For CO and HC, the residual error distribution, ε_2 , has a mean of greater than one, implying that the Mobile5b model is systematically underestimating the average emission rate because the residual error was not properly accounted for. Furthermore, although many of the input distributions for the uncertainty analysis are symmetric, a multiplicative model will typically yield positively skewed distributions for the product.

The random error is described in terms of how the 95 percent confidence interval compares to the mean on a relative basis. When uncertainty is relatively small, the random error is approximately symmetric and can be described as a simply “plus or minus” range. For example, the 95 percent probability range of uncertainty in the mean CO emission factor for LSP1 without any additional corrections is plus 62 percent or minus 59 percent, or approximately plus or minus 60 percent. However, when uncertainty is relatively large, the random error becomes asymmetric. For example, for the LSP1 CO emission factor estimate with both TCF and RVPCF, the range of uncertainty is minus 90 percent to plus 282 percent. The asymmetric uncertainty range results from the fact that emission factors cannot be negative. Therefore, the sampling

distribution of the mean is positively skewed when the range of uncertainty is large relative to the mean value.

For all driving cycles, the range of uncertainty in the mean CO emission factors with all correction factors applied is approximately minus 90 percent to plus 240 percent in many cases. For HC, the range of uncertainty in the mean emission factors with all correction factors applied is approximately minus 90 percent to plus 220 percent in most driving cycles. The range of uncertainty in the NO_x emission factors is not quite as large, ranging from approximately minus 60 percent to plus 120 percent. In all cases, the range of uncertainty in the emission factor with both TCF and RVPCF is substantially larger than for the base case conditions of ambient temperature and fuel RVP.

4.3 Identifying Key Sources of Uncertainty

Tables S-11, S-12, and S-13 in the Supplemental Information show the sample correlations for the the uncertain emission factors for CO, HC, and NO, respectively, calculated for each driving cycle when the SCF, TCF and RVPCF are all applied with respect to uncertainty in each individual model input. For all 11 driving cycles and for all three pollutants, the largest sample correlation coefficients are associated with the input uncertainty assumptions for TCF and RVPCF, with values of approximately 0.5 to 0.75 in most cases. In contrast, the uncertainty in the residual error terms, ε_1 and ε_2 , typically have sample correlations of less than 0.1 in magnitude for CO and HC, and less than 0.3 in magnitude for NO_x. The uncertainty in the SCF also contributes only modestly to overall uncertainty in the emission factors as reflected by a sample correlation of approximately 0.1 to 0.3 in most cases. There are some exceptions to these general trends. For example, for NO_x emissions, the uncertainty in the SCF contributes more to the range of uncertainty in the case of the HSP3 driving cycle than does any other input.

5.0 RESULTS AND DISCUSSION

In this work, all input variability distributions to the emission factor models were characterized based upon empirical distributions, rather than based upon assumed parametric probability distributions (e.g., normal, lognormal).. Thus, the analysis was based directly upon available data, without additional assumptions regarding the shape of the probability distributions and without introducing biases associated with curve fits. For the uncertainty analysis, normal distributions were used only when justified. For the TCF and RVPCF, empirical distributions based upon bootstrap simulation were used to characterize uncertainty in mean values.

The analysis results indicate that the range of both variability and uncertainty in the TCF and RVPCF was generally large (e.g., more than a factor of 10 in most cases), which contributes to large ranges of variability and uncertainty in emission factors. The range and shape of the distributions for uncertainty was heavily influenced by small sample sizes, large variation in the data, and positive skewness of the data.

The uncertainty in fleet average emission factors is typically of more interest in terms of policy-relevant analysis. This is because analysts are typically interested in predicting average emissions for fleets of vehicles than in knowing emission rates for individual vehicles. In many cases, the range of uncertainty is so large that traditional simplifying assumptions based upon normality and symmetry cannot be employed. For example, some emission factors were found to have uncertainty ranges of minus 80 percent to plus 220 percent of the mean value. The asymmetry reflects the fact that the emission factors are non-negative quantities and is influenced by both the large inter-vehicle variability in emissions and the relatively small sample sizes of data sets from which the emission factors were developed. The wide range of uncertainty implies that

there is a need to selectively test more vehicles in future model development to help reduce uncertainty in emission factors. Uncertainty in the TCF and RVPCF were typically the dominant sources of uncertainty in the estimated emission factors.

The uncertainty analysis reported here can be used as a basis for developing probabilistic emission inventories, which in turn can be used to determine the likelihood that an emission budget will be met and as input to air quality models to determine the likelihood that air quality management goals will be achieved. By understanding the levels of uncertainty in the Mobile5b emission factor estimates, it will be possible to interpret whether differences in emissions estimated with Mobile6 are statistically significant. Furthermore, the levels of uncertainty in the emission factors can be evaluated in terms of data quality objectives, as recommended by the National Research Council. Is the range of uncertainty acceptable? If not, what can be done to reduce uncertainty? One approach for reducing uncertainty is to collect more data for those portions of the emission factor model that contribute most to uncertainty in the emission factor. Thus, the methods presented here allow decision-makers to assess the quality of their decisions and to decide on whether and how to reduce the uncertainty that most significantly affects the vehicle emissions.

One key challenge in this work was the difficulty of obtaining data and information regarding the inputs and structure of Mobile5b. The effort required to do the uncertainty analysis once the data were available and the model was specified was a relatively small portion of this work. A key lesson, therefore, is that uncertainty analysis is much more efficient when done as an integrated part of model development, rather

than in a post-hoc manner. Thus, we strongly support the NRC recommendation that uncertainty analysis should be an integral part of future emission factor models.

6.0 ACKNOWLEDGMENTS

This work was supported in parts by the Center for Transportation and the Environment (CTE) at North Carolina State University and by the Office of Air Quality Planning and Standards of the U.S. Environmental Protection Agency (EPA). The opinions, findings, and conclusions expressed represent those of the authors and not necessarily those of CTE or the EPA

7.0 REFERENCES

1. National Research Council. *Modeling Mobile-Source Emissions*; National Academy Press: Washington, DC, 2000.
2. TRB. *Expanding Metropolitan Highways: Implications for Air Quality and Energy Use*; Special Report 245; Transportation Research Board: Washington, DC, 1995.
3. Kini, M.D.; Frey, H.C. *Probabilistic Evaluation of Mobile Source Air Pollution, Volume 1: Probabilistic Modeling of Exhaust Emissions from Light Duty Gasoline Vehicles*; Center for Transportation and the Environment, North Carolina State University: Raleigh, NC, 1997.
4. Pollack, A.K.; Bhave P.; Heiken J.; Lee K.; Shepard S.; Tran C.; Yarwood G.; Sawyer R.F.; Joy B.A. *Investigation of Emission Factors in the California EMFAC7G Model*; PB99-149718INZ; Coordinating Research Council: Atlanta, GA, 1999.
5. Chatterjee, A.; Wholley, Jr., T.F.; Guensler, R.; et al. *Improving Transportation Data for Mobile Source Emissions Estimates*; Project 25-7; National Cooperative Highway Research Program: Washington, DC, 1996.
6. *User's Guide to Mobile5*; EPA-AA-TEB-94-01; U.S. Environmental Protection Agency: Ann Arbor, MI. 1994 (Chapter 2 updated 1996).
7. Heirigs P.L.; Dulla, R.G. *Investigation of Mobile5a Emission Factors: Evaluation of IM240-to-FTP Correlation and Base Emission Rate Equations*; API Publication No. 4605; American Petroleum Institute: Washington, DC, 1994.
8. Sierra Research, Inc. *Evaluation of Mobile Vehicle Emission Model*; Prime Contract # DTRS-57-8-D-00089; US Department of Transportation: Washington, DC, 1994.
9. Systems Applications International, *Investigation of Mobile5a Emission Factors*; Final Report SYSAPP94-93/21 Irl ; American Petroleum Institute: Washington, DC, 1994.
10. *Draft User's Guide to Mobile6: Mobile Source Emission Factor Model*; EPA420-D-01-003; U.S. Environmental Protection Agency: Ann Arbor, MI, 2001.

11. Beardsley, M. *Mobile6: EPA's Highway Vehicle Emissions Model*; Presented at North American Vehicle Emission Control Conference: Atlanta, GA. 2001.
12. Bogen, K.T.; Spear, R.C., *Risk Analysis*, **1987**, 7, 427-436.
13. Efron, B.; Tibshirani, R.J., *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.
14. Hoffman, F.O.; Hammonds, J.S., *Risk Analysis*, **1994**, 7, 707 - 712
15. Frey, H.C.; Rhodes, D.S., *Human Health and Ecological Risk Assessment*, **1996**, 2:762-797.
16. Burmaster, D.E.; Wilson, A.M., *Human and Ecological Risk Assessment*, **1996**, 2:892 - 919.
17. Cullen, A.; Frey, H.C., *Probabilistic Techniques in Exposure Assessment*, Plenum: New York, 1999.
18. Helton, J.C.; Bean, J.E.; Economy, K.; et al., *Reliability Engineering and System Safety*, **2000**, 69, 263-304.
19. Frey, H.C.; Bharvirkar, R.; Zheng, J., *Quantitative Analysis of Variability and Uncertainty in Emissions Estimation*, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1999.
20. Webster, W.J.; Shih, C., "A Statically-Derived Metric to Monitor Time-Speed Variability in Practical Emissions Testing," *Proceedings of the Sixth CRC On-Road Vehicle Emissions Workshop*, Coordinating Research Council: Atlanta, GA, 1996.
21. Zhang, Y.; Bishop, G.A.; Stedman, D.H., *ES&T*, **1994**, 28, 1370-1374.
22. Bishop, G.A.; Stedman, D.H.; Ashbaugh, L., *J. Air Waste Manage. Assoc.*, **1996**, 46,667-675.
23. Frey, H.C., and Eichenberger, D.A., *Remote Sensing of Mobile Source Air Pollutant Emissions: Variability and Uncertainty in On-Road Emissions Estimates of Carbon Monoxide and Hydrocarbons for School and Transit Buses*, FHwy/NC/97-005, North Carolina Department of Transportation: Raleigh, NC, 1997.

Table 1. Characterization of Inter-Vehicle Variability in Estimated Tailpipe CO Emission Factors for Technology Group 8

Driving Cycle	Speed (mph)	T=75F, RVP=9, (only SCF)				T=50F, RVP=9, (SCF & TCF)				T=50F, RVP=13, (SCF, TCF and RVPCF)			
		Point Estimate	2.5th Percentile	Mean	97.5th Percentile	Point Estimate	2.5th Percentile	Mean	97.5th Percentile	Point Estimate	2.5th Percentile	Mean	97.5th Percentile
LSP1	2.45	57.2	1.27	44.0	241	81.7	0.627	58.2	377	223	0.551	150	1040
LSP2	3.64	38.1	1.77	42.5	247	54.4	0.753	66.3	419	149	0.703	164	1290
LSP3	4.02	34.3	1.24	54.2	301	49.0	0.644	77.7	592	134	0.552	218	1470
NYCC	7.1	18.8	2.27	37.9	193	26.9	0.974	65.2	398	73.3	0.968	172	1060
SCC12	12.1	10.4	0.75	16.0	82.1	14.9	0.435	24.2	130	40.6	0.353	69.0	442
FTP BAG2	16.1	7.46	0.80	10.5	46.4	10.7	0.343	16.1	96.7	29.1	0.287	43.4	305
SCC36	35.9	5.87	0.40	8.16	44.3	8.39	0.218	11.5	81.9	22.9	0.173	30.8	212
HFET	48.4	4.06	0.31	4.74	25.6	5.80	0.143	7.48	51.4	15.8	0.106	19.8	137
HSP1	50.9		0.36	5.06	22.4		0.143	7.91	55.2		0.124	20.9	148
HSP2	57.6		0.02	0.29	1.38		0.00806	0.436	3.17		0.00642	1.15	8.53
HSP3	64.3		0.01	0.29	1.34		0.00729	0.479	3.10		0.00638	1.16	7.77

Emission factors are in grams per mile

Point estimate is a deterministic estimate of the emission factor obtained as described by Kini and Frey (1997)

The 2.5th and 97.5th percentiles describe a 95 percent probability range for the emission factor.

The mean emission factors were obtained from probabilistic simulation.

Table 2. Characterization of Fleet Average Uncertainty in Estimated Tailpipe CO Emission Factors for Technology Group 8.

Driving Cycle	Speed (mph)	T=75F, RVP=9 (only SCF)						T=50F, RVP=9 (SCF & TCF)						T=50F, RVP=13 (SCF, TCF and RVPCF)					
		2.5th %ile	Mean	97.5th %ile	Systematic Error	Random (-) %	Error (+) %	2.5th %ile	Mean	97.5th %ile	Systematic Error	Random (-) %	Error (+) %	2.5th P %ile	Mean	97.5th %ile	Systematic Error	Random (-) %	Error (+) %
LSP1	2.45	15.3	36.9	59.8	23.6	-59	62	12.5	53.4	156	28.3	-77	192	15.1	148	567	75	-90	282
LSP2	3.64	11.7	38.9	64.5	0.92	-70	66	9.09	55.6	159	-1.2	-84	187	14.2	154	583	-5	-91	279
LSP3	4.02	19.4	50.9	84.1	-16.4	-62	65	14.5	73.4	227	-24.4	-80	209	20.0	222	727	-88	-91	227
NYCC	7.1	24.9	33.1	40.8	-11.8	-25	23	13.5	47.2	124	-20.3	-71	163	16.1	130	464	-57	-88	256
SCC12	12.1	10.3	14.3	18.2	-2.89	-28	27	5.81	20.4	55.4	-5.5	-72	171	6.49	56.8	187	-16	-89	229
FTP BAG2	16.1	7.91	8.78	9.66	-0.71	-10	10	3.57	12.6	33.7	-1.9	-72	168	4.40	34.7	112	-6	-87	224
SCC36	35.9	5.30	6.65	8.15	-0.33	-20	23	2.71	9.52	25.3	-1.13	-72	165	3.10	26.3	88.4	-3	-88	236
HFET	48.4	3.27	4.16	5.08	0.21	-21	22	1.72	5.95	15.7	-0.15	-71	165	1.98	16.5	57.1	-1	-88	246
HSP1	50.9	2.80	4.50	6.44		-38	43	1.72	6.40	16.7		-73	161	2.12	18.2	65.7		-88	262
HSP2	57.6	0.15	0.26	0.37		-42	43	0.10	0.37	0.99		-72	170	0.12	1.03	3.46		-88	236
HSP3	64.3	0.12	0.27	0.42		-57	57	0.08	0.38	1.09		-79	183	0.11	1.04	3.85		-90	272

Emission factors are in grams per mile

Point estimate is a deterministic estimate of the emission factor obtained as described by Kini and Frey (1997)

Systematic Error = Point Estimate – Mean

The 2.5th and 97.5th percentiles describe a 95 percent probability range for the emission factor.

The mean emission factors were obtained from probabilistic simulation.

Random Error (-) = (2.5th Percentile-Mean)/Mean*100

(+) = (97.5th Percentile-Mean)/Mean

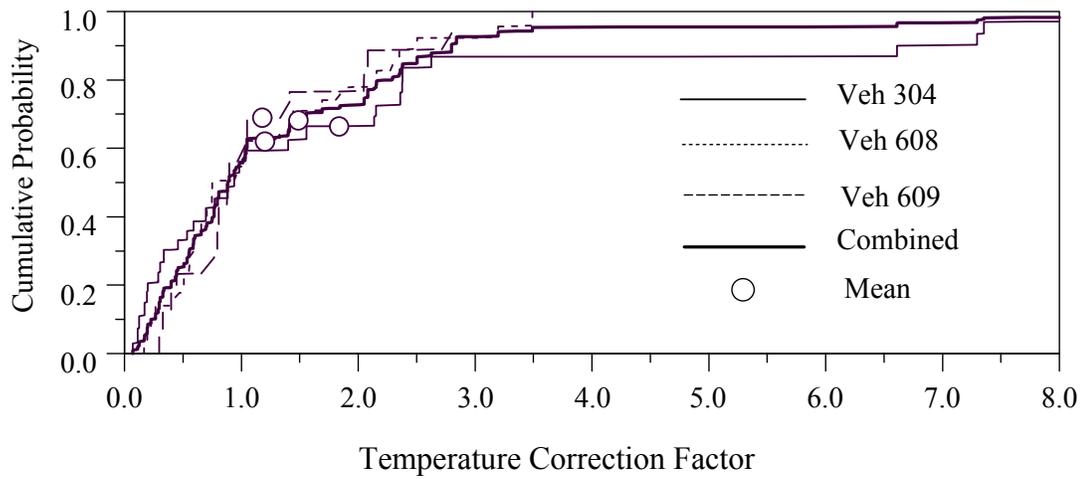


Figure 1. Estimated Inter-Vehicle Variability in Temperature Correction Factor For CO Technology Group 8

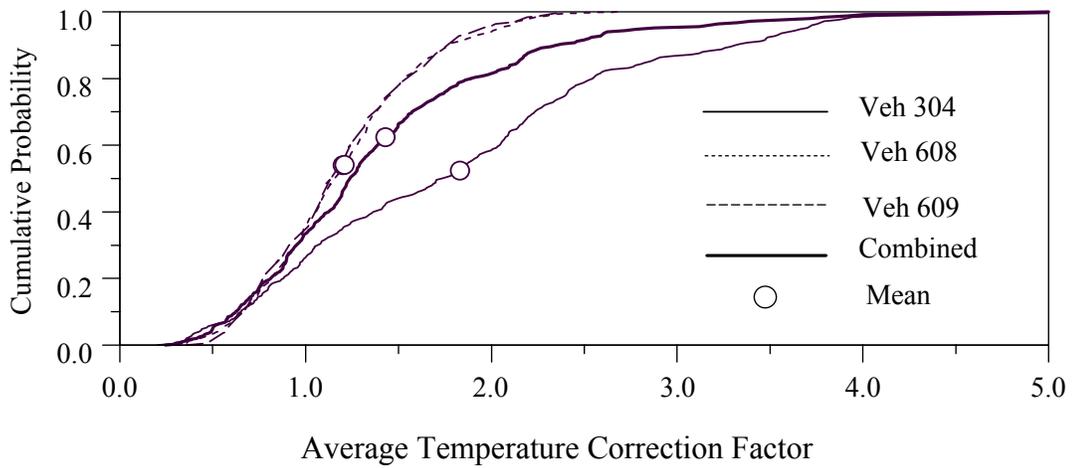


Figure 2. Estimated Fleet Average in Temperature Correction Factor For CO Technology Group 8

Support Information

(Table S1-13)

(Figure S1-S10)

Table S-1. Input Variability Assumptions for CO Emissions

Input ^a	Distribution ^c	Mean	2.5% Percentile	97.5% Percentile
BER ^b				
ϵ_1	Empirical	0	-1.32	1.52
ϵ_2	Empirical	1.38	0.18	4.92
SCF				
LSP1	Empirical	3.72	0.96	14.9
LSP2	Empirical	4.06	0.78	17.6
LSP3	Empirical	5.56	0.63	23.0
NYCC	Empirical	3.61	1.53	8.21
SCC12	Empirical	1.55	0.72	3.68
FTP Bag 2	n/a	1	1	1
SCC36	Empirical	0.75	0.23	1.62
HFET	Empirical	0.46	0.17	0.95
HSP1	Empirical	0.49	0.36	0.8
HSP2	Empirical	0.028	0.014	0.063
HSP3	Empirical	0.027	0.015	0.08
TCF	Empirical	1.43	0.13	7.30
RVPCF	Empirical	2.73	0.23	9.77

^a **BER**= Base Emission Rate, see equation (1) (gram/mile)

ϵ_1 = Residual error distribution for BER regression equation

ϵ_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycles

^b For BER, ZML=1.0372, DR=1.46e-5, MA=50,000 miles,

^c For normal distributions, the parameters μ (mean) and σ (standard deviation) are shown in parenthesis.

Table S-2. Input Uncertainty Assumptions for CO Emissions

Input ^a	Distribution ^c	Mean	2.5% Percentile	97.5% Percentile
BER ^b				
ε_1	Normal (0, 0.0263)			
ε_2	Normal(1.48, 0.0636)			
SCF				
LSP1	Normal(4.24, 1.230)			
LSP2	Normal(4.43, 1.490)			
LSP3	Normal(6.17, 1.93)			
NYCC	Normal(3.76, 0.4310)			
SCC12	Normal(1.63, 0.216)			
FTP Bag 2	n/a	1	1	1
SCC36	Normal(0.76, 0.073)			
HFET	Normal(0.474, 0.046)			
HSP1	Normal(0.514, 0.107)			
HSP2	Normal(0.0295, 0.0062)			
HSP3	Normal(0.0305, 0.0089)			
TCF	Empirical	1.43	0.38	3.5
RVPCF	Empirical	2.73	0.58	7.21

^a **BER**= Base Emission Rate, see equation (1) (gram/mile)

ε_1 = Residual error distribution for BER regression equation

ε_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycles

^b For BER, ZML=1.0372, DR=1.46e-5, MA=50,000 miles,

^c For normal distributions, the parameters μ (mean) and σ (standard deviation) are shown in parenthesis.

Table S-3. Input Variability Assumptions for HC Emissions

Input ^a	Distribution ^c	Mean	2.5% Percentile	97.5% Percentile
BER ^b				
ϵ_1	Empirical	0.0012	-0.630	1.41
ϵ_2	Empirical	1.32	0.230	3.820
SCF				
LSP1	Empirical	9.66	3.91	20.6
LSP2	Empirical	10.3	4.25	19.1
LSP3	Empirical	17.7	5.62	30.9
NYCC	Empirical	2.83	0.631	11.1
SCC12	Empirical	2.81	0.479	11.2
FTP Bag 2	n/a	1	1	1
SCC36	Empirical	0.555	0.176	1.22
HFET	Empirical	0.455	0.122	1.40
HSP1	Empirical	1.21	0.769	1.59
HSP2	Empirical	1.12	0.463	2.23
HSP3	Empirical	1.09	0.423	2.78
TCF	Empirical	1.22	0.179	4.08
RVPCF	Empirical	2.50	0.129	11.4

^a **BER**= Base Emission Rate, see equation (1) (gram/mile)

ϵ_1 = Residual error distribution for BER regression equation

ϵ_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycles

^b For BER, ZML=1.0372, DR=1.46e-5, MA=50,000 miles,

^c For normal distributions, the parameters μ (mean) and σ (standard deviation) are shown in parenthesis.

Table S-4. Input Uncertainty Assumptions for HC Emissions

Input ^a	Distribution ^c	Mean	2.5% Percentile	97.5% Percentile
BER ^b				
ϵ_1	Normal (0, 0.0185)			
ϵ_2	Normal(1.33,0.064)			
SCF				
LSP1	Normal(9,87, 1.540)			
LSP2	Normal(10.5, 1.38)			
LSP3	Normal(17.7, 2.57)			
NYCC	Normal(3.03, 0.603)			
SCC12	Normal(2.92, 0.526)			
FTP Bag 2	n/a	1	1	1
SCC36	Normal(0.570, 0.0594)			
HFET	Normal(0.463, 0.0496)			
HSP1	Normal(1.21, 0.199)			
HSP2	Normal(1.15, 0.218)			
HSP3	Normal(1.16, 0.294)			
TCF	Empirical	1.22	0.435	2.62
RVPCF	Empirical	2.53	0.563	6.20

^a **BER**= Base Emission Rate, see equation (1) (gram/mile)

ϵ_1 = Residual error distribution for BER regression equation

ϵ_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycles

^b For BER, ZML=1.0372, DR=1.46e-5, MA=50,000 miles,

^c For normal distributions, the parameters μ (mean) and σ (standard deviation) are shown in parenthesis.

Table S-5. Input Variability Assumptions for NO_x Emissions

Input ^a	Distribution ^c	Mean	2.5% Percentile	97.5% Percentile
BER ^b				
ε ₁	Empirical	0.00	-0.870	0.954
ε ₂	Empirical	0.002	-0.505	0.597
SCF				
LSP1	Empirical	3.11	1.98	6.49
LSP2	Empirical	3.10	1.40	5.32
LSP3	Empirical	3.66	0.923	6.61
NYCC	Empirical	2.27	1.02	4.35
SCC12	Empirical	2.33	1.11	8.89
FTP Bag 2	n/a	1	1	1
SCC36	Empirical	1.08	0.590	2.01
HFET	Empirical	0.881	0.493	1.63
HSP1	Empirical	1.74	0.841	2.89
HSP2	Empirical	0.221	0.0642	0.778
HSP3	Empirical	0.241	0.0776	1.08
TCF	Empirical	1.06	0.303	2.67
RVPCF	Empirical	1.43	0.350	3.53

^a **BER**= Base Emission Rate, see equation (2) (gram/mile)

ε₁= Residual error distribution for BER regression equation

ε₂= Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycles

^b For BER, ZML=1.0372, DR=1.46e-5, MA=50,000 miles,

^c For normal distributions, the parameters μ (mean) and σ (standard deviation) are shown in parenthesis.

Table S-6. Input Uncertainty Assumptions for NO_x Emissions

Input ^a	Distribution ^c	Mean	2.5% Percentile	97.5% Percentile
BER ^b				
ε ₁	Normal (0, 0.0161)			
ε ₂	Normal(0.0021, 0.109)			
SCF				
LSP1	Normal(3.22, 0.396)			
LSP2	Normal(3.11, 0.326)			
LSP3	Normal(3.67, 0.505)			
NYCC	Normal(2.30, 0.164)			
SCC12	Normal(2.56, 0.550)			
FTP Bag 2	n/a	1	1	1
SCC36	Normal(1.09, 0.0629)			
HFET	Normal(0.888, 0.0513)			
HSP1	Normal(1.78, 0.476)			
HSP2	Normal(0.250, 0.0955)			
HSP3	Normal(0.290, 0.141)			
TCF	Empirical	1.08	0.59	1.61
RVPCF	Empirical	1.44	0.69	2.76

^a **BER**= Base Emission Rate, see equation (2) (gram/mile)

ε₁= Residual error distribution for BER regression equation

ε₂= Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycles

^b For BER, ZML=1.0372, DR=1.46e-5, MA=50,000 miles,

^c For normal distributions, the parameters μ (mean) and σ (standard deviation) are shown in parenthesis.

Table S-7. Characterization of Inter-Vehicle Variability in Estimated Tailpipe HC Emission Factors for Technology Group 8.

Driving Cycle	Speed (mph)	T=75F, RVP=9, (only SCF)				T=50F, RVP=9, (SCF & TCF)				T=50F, RVP=13, (SCF, TCF and RVPCF)			
		Point Estimate	2.5th Percentile	Mean	97.5th Percentile	Point Estimate	2.5th Percentile	Mean	97.5th Percentile	Point Estimate	2.5th Percentile	Mean	97.5th Percentile
LSP1	2.45	2.40	0.61	6.31	26.2	2.93	0.225	8.53	49.0	7.42	0.133	20.2	127
LSP2	3.64	1.67	0.61	6.98	29.4	2.04	0.286	8.42	45.2	5.17	0.165	20.7	130
LSP3	4.02	1.52	0.94	12.3	46.7	1.85	0.463	12.80	70.3	4.69	0.253	32.1	199
NYCC	7.1	0.92	0.10	1.76	8.71	1.12	0.056	2.21	14.8	2.84	0.023	6.01	39.4
SCC12	12.1	0.60	0.10	1.97	10.4	0.73	0.043	2.11	16.5	1.85	0.0252	5.71	41.8
FTP BAG2	16.1	0.48	0.09	0.67	2.56	0.59	0.040	0.813	4.03	1.50	0.0175	1.98	12.0
SCC36	35.9	0.42	0.04	0.38	1.59	0.51	0.013	0.443	2.40	1.29	0.0083	1.05	7.94
HFET	48.4	0.35	0.03	0.32	1.45	0.43	0.012	0.380	2.07	1.09	0.0059	0.883	5.63
HSP1	50.9		0.10	0.81	3.08		0.047	0.977	4.68		0.0199	2.43	15.4
HSP2	57.6		0.08	0.75	3.04		0.034	0.816	4.60		0.0168	2.06	14.1
HSP3	64.3		0.07	0.73	3.33		0.035	0.886	4.77		0.0172	2.08	15.2

Emission factors are in grams per mile

Point estimate is a deterministic estimate of the emission factor obtained as described by Kini and Frey (1997)

The 2.5th and 97.5th percentiles describe a 95 percent probability range for the emission factor.

The mean emission factors were obtained from probabilistic simulation.

Table S-8. Characterization of Fleet Average Uncertainty in Estimated Tailpipe HC Emission Factors for Technology Group 8.

Driving Cycle	Speed (mph)	T=75F, RVP=9, (only SCF)						T=50F, RVP=9, (SCF & TCF)						T=50F, RVP=13, (SCF, TCF and RVPCF)					
		2.5th Percentile	Mean	97.5th %ile	Systematic Error	Random (-)%	Error (+)%	2.5th Percentile	Mean	97.5th %ile	Systematic Error	Random (-)%	Error (+)%	2.5th Percentile	Mean	97.5th %ile	Systematic Error	Random (-)%	Error (+)%
LSP1	2.45	3.81	5.55	7.36	-3.14	-31	33	2.21	6.81	15.6	-3.88	-68	129	2.37	17.2	55.2	-9.66	-86	221
LSP2	3.64	4.25	5.88	7.55	-4.20	-28	28	2.37	7.19	16.2	-5.15	-67	125	2.44	18.2	62.5	-13.0	-87	243
LSP3	4.02	7.03	9.97	12.9	-8.39	-30	30	4.11	12.2	27.8	-10.4	-66	127	4.03	30.9	99.4	-26.0	-87	222
NYCC	7.1	1.01	1.71	2.40	-0.78	-41	40	0.65	2.08	5.01	-0.96	-69	140	0.71	5.30	17.7	-2.41	-87	235
SCC12	12.1	1.06	1.64	2.22	-1.04	-35	36	0.61	2.01	4.58	-1.28	-69	129	0.68	5.07	16.7	-3.17	-87	230
FTP BAG2	16.1	0.51	0.56	0.62	-0.08	-10	10	0.24	0.69	1.46	-0.10	-65	113	0.25	1.74	5.72	-0.23	-86	229
SCC36	35.9	0.25	0.32	0.40	0.10	-22	24	0.13	0.39	0.86	0.12	-66	121	0.14	0.99	3.28	0.30	-86	231
HFET	48.4	0.21	0.26	0.32	0.09	-21	24	0.11	0.32	0.69	0.11	-66	118	0.11	0.80	2.53	0.29	-86	215
HSP1	50.9	0.45	0.68	0.91		-33	34	0.27	0.82	1.81		-68	119	0.28	2.08	6.62		-86	218
HSP2	57.6	0.40	0.65	0.88		-38	36	0.25	0.79	1.79		-68	126	0.28	2.00	6.62		-86	231
HSP3	64.3	0.33	0.65	0.98		-49	51	0.22	0.80	1.96		-72	145	0.24	2.00	6.72		-88	235

Emission factors are in grams per mile

Point estimate is a deterministic estimate of the emission factor obtained as described by Kini and Frey (1997)

Systematic Error = Point Estimate – Mean

The 2.5th and 97.5th percentiles describe a 95 percent probability range for the emission factor.

The mean emission factors were obtained from probabilistic simulation.

Random Error (-) = (2.5th Percentile-Mean)/Mean*100

(+) = (97.5th Percentile-Mean)/Mean*100

Table S-9. Characterization of Inter-Vehicle Variability in Estimated Tailpipe NO_x Emission Factors for Technology Group 8

Driving Cycle	Speed (mph)	T=75F, RVP=9, (only SCF)				T=50F, RVP=9, (SCF & TCF)				T=50F, RVP=13, (SCF, TCF and RVPCF)			
		Point Estimate	2.5th Percentile	Mean	97.5th Percentile	Point Estimate	2.5th Percentile	Mean	97.5th Percentile	Point Estimate	2.5th Percentile	Mean	97.5th Percentile
LSP1	2.45	1.61	0.29	2.93	8.31	1.71	0.177	3.00	11.4	2.44	0.180	4.20	17.6
LSP2	3.64	1.32	0.28	2.77	7.75	1.40	0.178	2.99	9.95	2.00	0.204	4.23	17.2
LSP3	4.02	1.26	0.19	3.36	8.84	1.34	0.227	3.52	13.3	1.91	0.176	4.91	20.3
NYCC	7.1	1.02	0.16	2.09	6.09	1.08	0.132	2.22	8.03	1.55	0.111	3.08	12.6
SCC12	12.1	0.89	0.13	2.27	10.2	0.943	0.128	2.30	11.1	1.35	0.108	3.16	14.6
FTP BAG2	16.1	0.84	0.11	0.91	2.27	0.890	0.089	0.974	3.05	1.27	0.0716	1.36	5.42
SCC36	35.9	0.82	0.07	1.00	2.60	0.869	0.0766	1.07	3.48	1.24	0.0548	1.49	6.11
HFET	48.4	0.79	0.08	0.78	2.06	0.837	0.0644	0.872	3.06	1.20	0.0624	1.24	5.48
HSP1	50.9		0.13	1.57	4.41		0.0964	1.66	5.68		0.105	2.34	10.50
HSP2	57.6		0.01	0.20	0.85		0.0093	0.2101	1.00		0.0081	0.295	1.42
HSP3	64.3		0.01	0.22	1.38		0.0094	0.244	1.46		0.0088	0.335	1.86

Emission factors are in grams per mile

Point estimate is a deterministic estimate of the emission factor obtained as described by Kini and Frey (1997)

The 2.5th and 97.5th percentiles describe a 95 percent probability range for the emission factor.

The mean emission factors were obtained from probabilistic simulation.

Table S-10. Characterization of Fleet Average Uncertainty in Estimated Tailpipe NO_x Emission Factors for Technology Group 8.

Driving Cycle	Speed (mph)	T=75F, RVP=9, (only SCF)						T=50F, RVP=9, (SCF & TCF)						T=50F, RVP=13, (SCF, TCF and RVPCF)					
		2.5th Percentile	Mean	97.5th %ile	System-atic Error	Random (-)%	Error (+)%	2.5th Percentile	Mean	97.5th %ile	System-atic Error	Random (-)%	Error (+)%	2.5th Percentile	Mean	97.5th %ile	System-atic Error	Random (-)%	Error (+)%
LSP1	2.45	1.80	2.67	3.63	-1.03	-33	36	1.39	2.91	4.90	-1.20	-52	68	1.55	4.13	9.04	-1.69	-62	119
LSP2	3.64	1.76	2.57	3.50	-1.24	-32	36	1.37	2.81	4.59	-1.41	-51	63	1.47	3.97	8.57	-1.97	-63	116
LSP3	4.02	1.99	3.05	4.24	-1.77	-35	39	1.57	3.31	5.72	-1.97	-53	73	1.68	4.70	10.5	-2.79	-64	123
NYCC	7.1	1.38	1.91	2.46	-0.89	-28	29	1.04	2.08	3.41	-1.00	-50	64	1.10	2.94	6.20	-1.39	-63	111
SCC12	12.1	1.13	2.13	3.37	-1.21	-47	58	0.92	2.31	4.02	-1.37	-60	74	1.03	3.24	7.58	-1.89	-68	134
FTP BAG2	16.1	0.61	0.83	1.05	0.01	-27	26	0.46	0.90	1.46	-0.010	-50	62	0.49	1.28	2.74	-0.010	-62	115
SCC36	35.9	0.64	0.90	1.14	-0.08	-29	26	0.49	0.98	1.62	-0.111	-51	65	0.53	1.39	3.01	-0.150	-62	116
HFET	48.4	0.53	0.74	0.95	0.06	-28	28	0.40	0.80	1.31	0.037	-50	63	0.42	1.13	2.44	0.070	-63	115
HSP1	50.9	0.64	1.47	2.42		-56	64	0.61	1.60	3.04		-62	90	0.65	2.25	5.27		-71	134
HSP2	57.6	0.07	0.21	0.37		-68	81	0.04	0.22	0.46		-83	104	0.05	0.32	0.80		-83	151
HSP3	64.3	0.00	0.24	0.48		-100	100	0.02	0.26	0.57		-94	117	0.02	0.38	1.09		-96	185

Emission factors are in grams per mile

Point estimate is a deterministic estimate of the emission factor obtained as described by Kini and Frey (1997)

Systematic Error = Point Estimate – Mean

The 2.5th and 97.5th percentiles describe a 95 percent probability range for the emission factor.

The mean emission factors were obtained from probabilistic simulation.

Random Error (-) = (2.5th Percentile-Mean)/Mean*100

(+) = (97.5th Percentile-Mean)/Mean*100

Table S-11. Correlation of Uncertain CO Emission Factor with Input Uncertainties

Driving Cycle	ϵ_1	ϵ_2	SCF	TCF	RVPCF
LSP1	0.0068	0.032	0.28	0.56	0.64
LSP2	-0.032	0.036	0.32	0.57	0.60
LSP3	-0.0053	0.052	0.30	0.56	0.62
NYCC	-0.010	0.050	0.084	0.60	0.68
SCC12	0.029	0.034	0.12	0.62	0.67
FTP Bag 2	-0.0023	0.052	NA	0.62	0.68
SCC36	-0.0038	0.047	0.067	0.60	0.69
HFET	-0.0031	0.047	0.075	0.59	0.68
HSP1	0.0014	0.051	0.20	0.58	0.65
HSP2	0.0087	0.035	0.21	0.58	0.65
HSP3	-0.017	0.056	0.31	0.56	0.64

ϵ_1 = Residual error distribution for BER regression equation

ϵ_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycle

Table S-12. Correlation of Uncertain HC Emission Factor with Input Uncertainties

Driving Cycle	ϵ_1	ϵ_2	SCF	TCF	RVPCF
LSP1	0.0098	0.073	0.20	0.61	0.66
LSP2	0.013	0.089	0.17	0.61	0.68
LSP3	0.0086	0.067	0.19	0.60	0.70
NYCC	0.013	0.061	0.24	0.59	0.67
SCC12	0.021	0.070	0.22	0.58	0.68
FTP Bag 2	0.011	0.070	NA	0.62	0.70
SCC36	0.0068	0.069	0.13	0.61	0.69
HFET	0.014	0.068	0.074	0.61	0.70
HSP1	0.0054	0.060	0.16	0.60	0.69
HSP2	0.0065	0.072	0.21	0.60	0.67
HSP3	-0.0042	0.079	0.25	0.55	0.68

ϵ_1 = Residual error distribution for BER regression equation

ϵ_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycle

Table S-13. Correlation of Uncertain NO_x Emission Factor with Input Uncertainties

Driving Cycle	ϵ_1	ϵ_2	SCF	TCF	RVPCF
LSP1	0.019	0.26	0.27	0.51	0.73
LSP2	0.018	0.27	0.24	0.52	0.74
LSP3	-0.0077	0.28	0.28	0.52	0.70
NYCC	0.0090	0.30	0.17	0.52	0.74
SCC12	0.0065	0.27	0.41	0.47	0.66
FTP Bag 2	0.012	0.030	NA	0.53	0.75
SCC36	0.016	0.30	0.010	0.53	0.74
HFET	0.11	0.29	0.12	0.52	0.75
HSP1	0.0088	0.22	0.49	0.42	0.65
HSP2	0.0076	0.21	0.61	0.37	0.56
HSP3	0.026	0.15	0.70	0.36	0.47

ϵ_1 = Residual error distribution for BER regression equation

ϵ_2 = Residual error distribution for IM240 to FTP regression equation

SCF= Speed Correction Factor (dimensionless)

TCF= Temperature Correction Factor (dimensionless)

RVPCF= Reid Vapor Pressure Correction Factor (dimensionless)

LSP1, LSP2, ..., refers to specific driving cycle

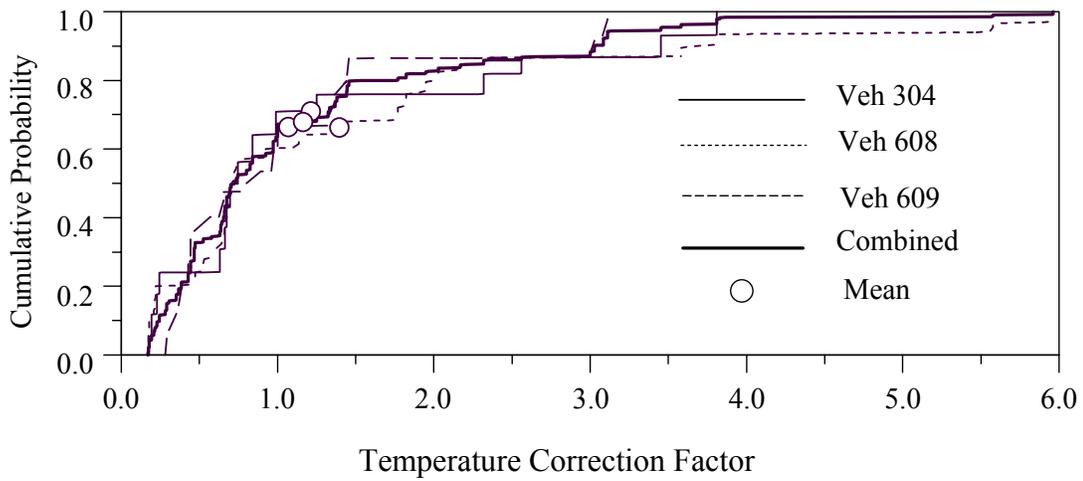


Figure S-1. Estimated Inter-Vehicle Variability in Temperature Correction Factor For HC Technology Group 8

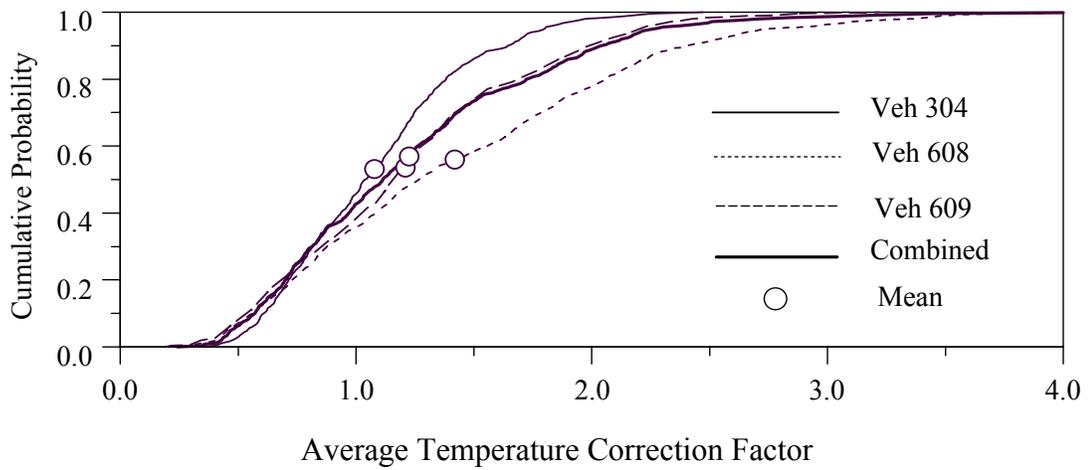


Figure S-2. Estimated Fleet Average in Temperature Correction Factor For HC Technology Group 8

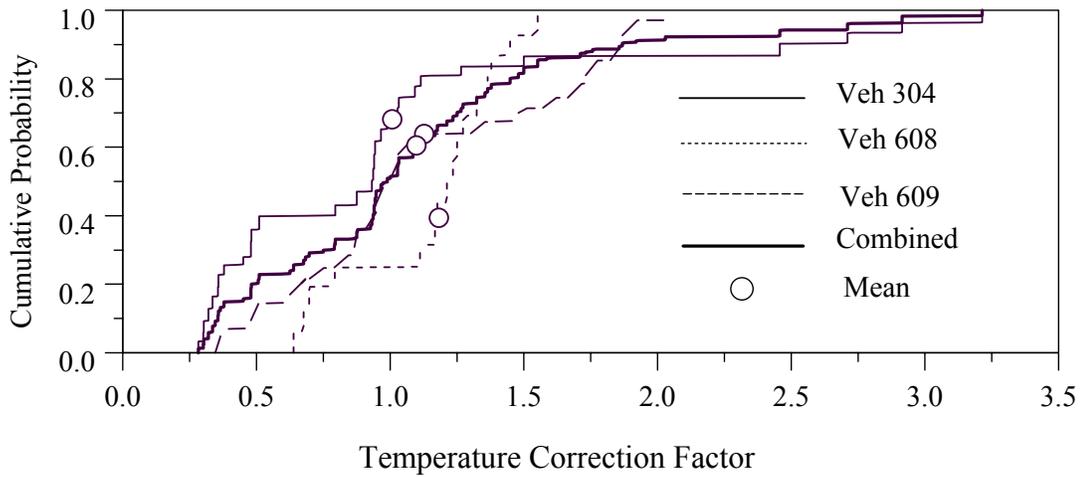


Figure S-3. Estimated Inter-Vehicle Variability in Temperature Correction Factor For NO_x Technology Group 8

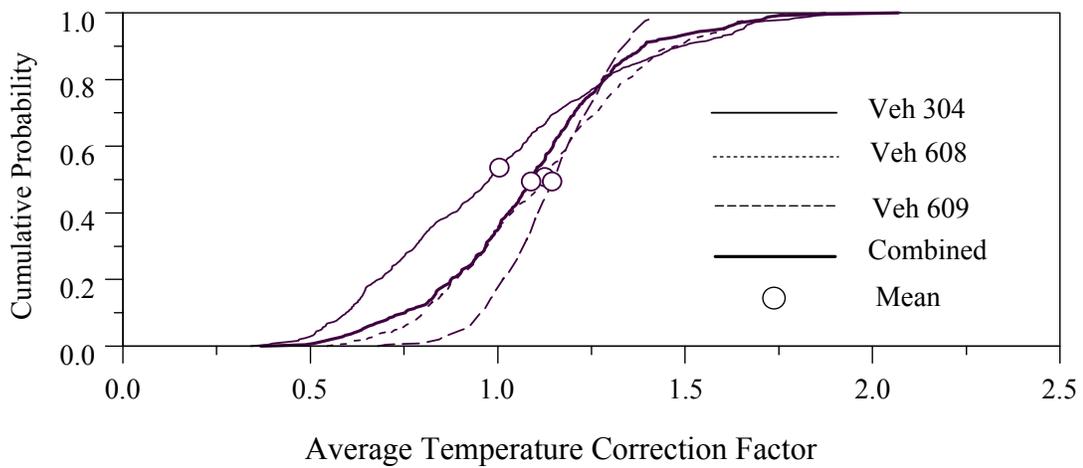


Figure S-4. Estimated Fleet Average in Temperature Correction Factor For NO_x Technology Group 8

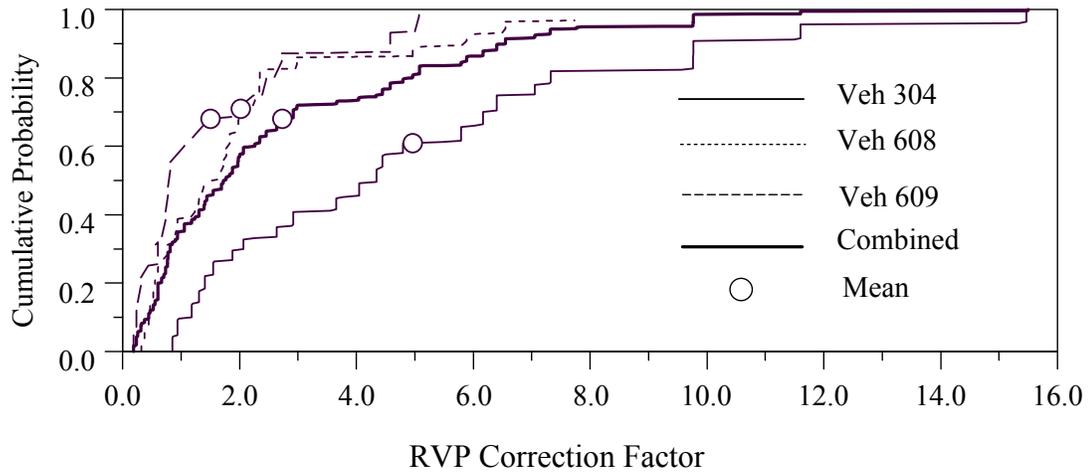


Figure S-5. Estimated Inter-Vehicle Variability in RVP Correction Factor For CO Technology Group 8

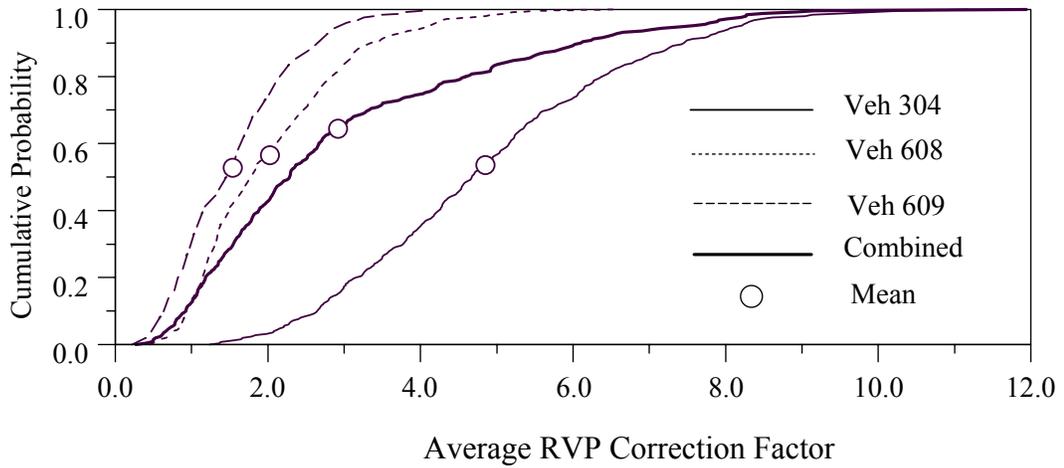


Figure S-6. Estimated Fleet Average in RVP Correction Factor For CO Technology Group 8

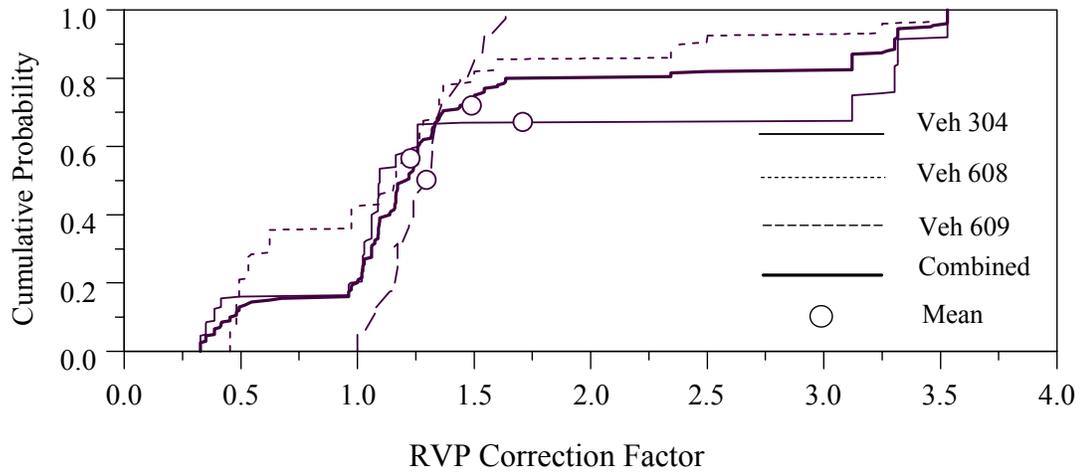


Figure S-7. Estimated Inter-Vehicle Variability in RVP Correction Factor For NO_x Technology Group 8

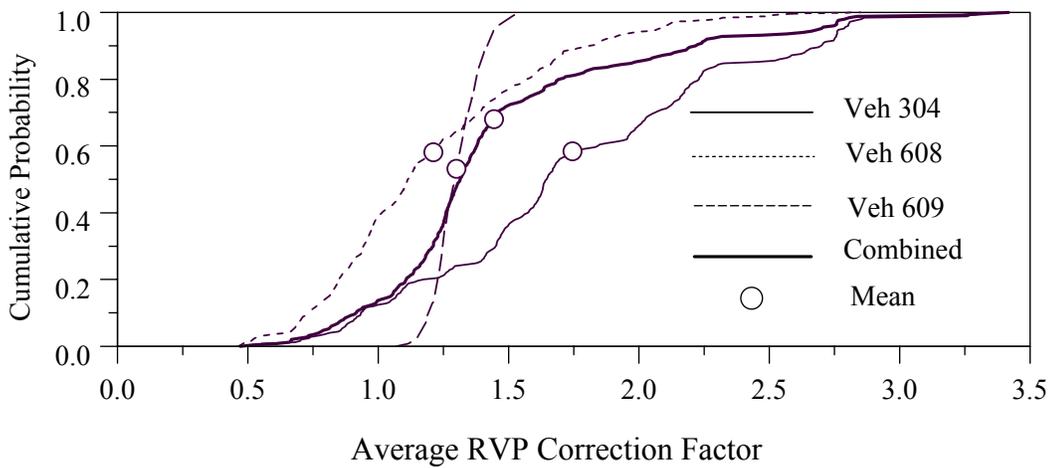


Figure S-8. Estimated Fleet Average in RVP Correction Factor For NO_x Technology Group 8

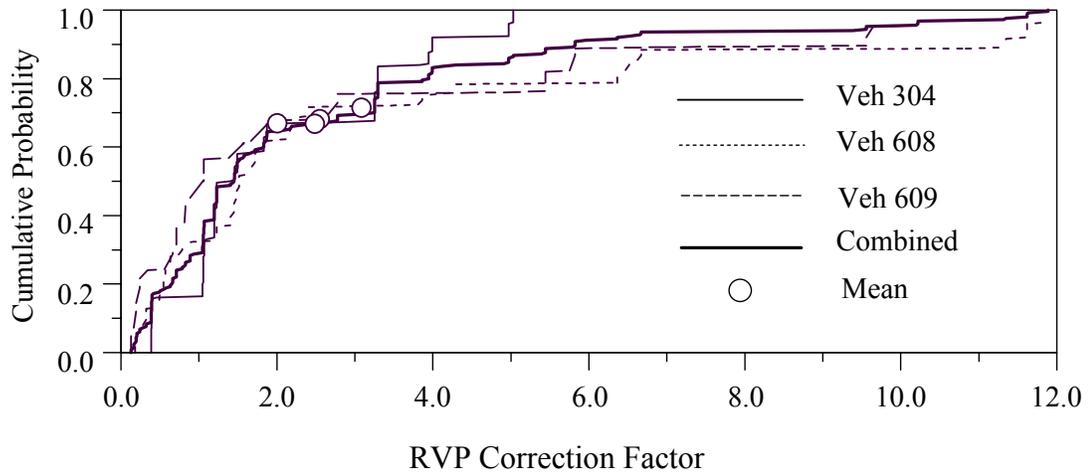


Figure S-9. Estimated Inter-Vehicle Variability in RVP Correction Factor For HC Technology Group 8

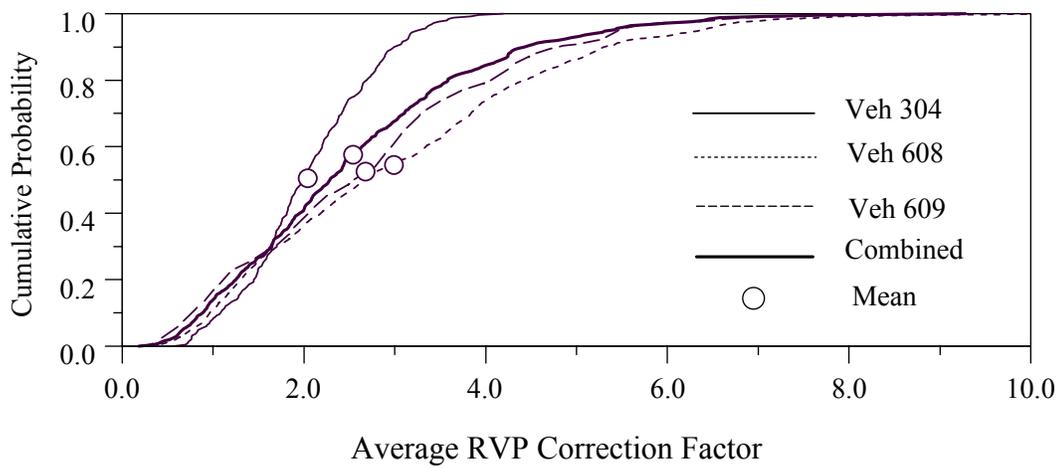


Figure S-10. Estimated Fleet Average in RVP Correction Factor For HC Technology Group 8

PART VIII

CONCLUSIONS AND RECOMMENDATIONS

Junyu Zheng

8.0 CONCLUSIONS AND RECOMMENDATIONS

This part presents a summary of work done in this dissertation, key conclusions from this dissertation and offers recommendations for future work.

8.1 Summary

The use of probabilistic quantitative methods for dealing with variability and uncertainty in the development of emission factors and emission inventories is being widely recognized as an improvement over qualitative techniques such as AP-42 data quality rating factors or semi-quantitative techniques such as DARS. Probabilistic analysis provides decision-makers with quantitative information about the confidence with which an emission factor may be used and helps decision-makers to determine the likelihood that an emissions budget will be met. Ignorance of the distinction between variability and uncertainty may lead to erroneous conclusions regarding emission factors and emission inventories. Variability refers to the heterogeneity of a quantity with respect to time, space, or different members of a population. Uncertainty refers to the lack of knowledge regarding the true value of an empirical quantity. This dissertation extensively and systematically discussed methodologies associated with quantification of variability and uncertainty in the development of emission factors and emission inventories, including methods based upon the use of mixture distribution and methods for accounting for the effect of measurement error on variability and uncertainty analysis. Some example case studies have been applied to demonstrate the methodologies. The main work done in this dissertation is summarized in the following subsections.

8.1.1 Methodologies

A general approach for developing a probabilistic emission inventory was presented. The general approach can be applied to the development of emission inventories for various air pollutant source categories. Associated methods for dealing with variability and uncertainty in the development of a probabilistic emission inventory were systematically introduced and summarized in the general approach framework.

These methods include:

- Development of database used to develop an emission inventory
- Selection or development of emission inventory models
- Numerical sampling techniques
- Plotting of data sets using the Hazen plotting position
- Visualization of the CDF of fitted distributions and graphical comparison of these with the data
- Estimation of parameters for parametric probability distributions using MoMM or MLE approaches
- Presentation of empirical step-wise CDFs.
- Generation of random numbers from empirical step-wise CDFs or from parametric probability distribution models.
- Calculation of test statistics as an aid in determined goodness-of-fit of a parametric probability distribution to a data set. The statistic tests include K-S and Anderson-Darling tests.

- Estimation of confidence intervals of the CDF of a parametric probability distribution fitted to a dataset and graphical comparison with the data as an aid in evaluating goodness-of-fit.
- Use of bootstrap simulation to characterize sampling distributions and confidence intervals for key statistics, such as the mean, standard deviation, and parameters of parametric probability distribution models. The methods of forming confidence interval include percentile and BC_a methods.
- Propagation of uncertainty and variability in model inputs through emission inventory models to estimate uncertainty in category-specific emissions and/or total emissions from a population of emission sources.
- Identification of key sources of variability and uncertainty and the associated methodological issues to address the variability and uncertainty

An approach for quantifying variability and uncertainty based on mixture distributions featuring the use of bootstrap simulation was developed in this dissertation. The approach includes a method for parameter estimation of mixture distributions and a method for generating bootstrap samples based upon mixture distributions and for forming confidence intervals based upon bootstrap samples.

A methodology for improving the characterization of variability and uncertainty with known measurement errors in environmental data was demonstrated in this study. The method for constructing an error free data set based on the observed data set, known measurement error and measurement error models was demonstrated. Numerical methods based upon bootstrap pair techniques were applied to characterize uncertainty in statistics for the measurement error problem.

8.1.2 Case Studies

An emission factor case study based upon NO_x emissions from coal-fired tangential boilers with low NO_x burners and overfire air was used to illustrate the use of the approach. To evaluate properties of quantification of variability and uncertainty based on mixture distributions with respect to variation in sample size, mixing weight and separation between components, 108 synthetic datasets generated from the selected population mixture lognormal distributions are investigated. A case study was used to demonstrate the use of methods for dealing with variability and uncertainty with known measurement errors. A few measurement error models with different sizes of measurement error were used to investigate and evaluate the effect of measurement error on variability and uncertainty estimates with respect to the size of measurement error.

An illustrative example case study from utility power plant emission source was used to demonstrate the development of a probabilistic emission inventory by following the general approach developed in this research.

A case study with the use of Mobile5b highway vehicle emission factor model was done to demonstrate an approach for conducting uncertainty analysis in highway vehicle emission factors with incorporation of variability and uncertainty from the temperature correction factor and Reid vapor pressure correction factor. Inter-vehicle variability and fleet average uncertainty was characterized for HC, CO, and NO_x emission factors for a selected light duty gasoline vehicle technology. Numerical methods based upon the use of Monte Carlo simulation, Latin Hypercube and bootstrap simulation were employed to develop emission factors for 11 driving cycles, and each

with a different average speed, and for adjustments to both temperature and fuel Reid vapor pressure compared to standard test conditions.

8.1.3 Software Development

A prototype software tool, AUVÉE, was developed to demonstrate the general approach for calculating a probabilistic emission inventory. The prototype software enables a user to visualize, in the form of empirical probability distributions, the data used to develop the inventory. Therefore, the user is able to observe the range of variability in the data and characterize the uncertainty in emission factors and activity factors. This is sharp contrast to typical emission inventory work, in which point estimate values of emission factors are used to calculate a single estimate of the inventory. Although AUVÉE is limited to the use of a particular power plant utility emission source category, the ideas and approaches demonstrated in AUVÉE provide a basic framework or prototype for the development of future integrated emission inventory system, in which various emission source categories can be covered.

A general software tool, AuvTool, was developed based upon the methodologies introduced and developed in this dissertation. AuvTool features the use of bootstrap simulation and two-dimensional Monte Carlo simulation for simultaneously quantifying variability and uncertainty. It provides a user-friendly environment for statistical analysis of variability and uncertainty in model inputs. It can be used in any applications in which there is a need to fit distributions to data and/or to characterize variability and uncertainty associated with the fitted distribution. AuvTool has the following main features: (1) an input data sheet similar to a spreadsheet for data input, which also can import or export Microsoft Excel and tab-limited text files; (2) parameter estimation for common single

component parametric distributions, including method of matching moment and maximum likelihood estimation method; (3) Kolmogorov-Smirnov and Anderson-Darling statistical test measures to help the user choose a good fit; (4) batch analysis to help the user to handle a large number of datasets efficiently; (5) uncertainty analysis for variables without original data; (6) further analysis of the sampling data of statistics of interests from bootstrap simulation; (7) variability and uncertainty analysis based upon the use of two component mixture distributions; (8) variability and uncertainty analysis for datasets with known measurement error; and (9) instant graphical presentation of simulation results and tabular summarization of variability and uncertainty analysis results. One application of AuvTool is that it has become a module of EPA/SHEDS models to provide variability and uncertainty analysis in SHEDS model inputs.

8.2 Conclusions

The key contributions and main conclusions from this dissertation are presented in the subsections.

8.2.1 Methodological Conclusions

The general approach presented in Part II of this dissertation for developing a probabilistic emission inventory provides a framework and guidance in dealing with variability and uncertainty in emission estimation. The use of the general approach will be able to provide probabilistic emission estimation in the development of an emission inventory, which has many important implications for decision analysis. For example, it enables analysts and decision makers to evaluate whether time series trends are statistically significant or not. It enables decision makers to determine the likelihood that an emissions budget will be met. Inventory uncertainties can be used as input to air quality models to estimate uncertainty in predicted ambient concentrations, which in turn

can be compared to ambient air quality standards to determine the likelihood that a particular control strategy will be effective in meeting the standards. In addition, using probabilistic methods, it is possible to compare the uncertainty reduction benefits of alternative emission inventory development methods, such as those based upon generic versus more site-specific data. Thus, the methods presented in this dissertation allow decision makers to assess the quality of their decisions and to decide on whether and how to reduce the uncertainties that most significantly affect those decisions.

Mixture distributions are potentially useful in the quantification of variability and uncertainty because they can improve the goodness of fit to datasets that cannot be adequately described by a single component parametric distribution. The method presented in this dissertation for dealing with the use of mixture distribution is a promising one for improving the fit of distributions to data and for obtaining improved estimates of variability and uncertainty in statistics estimated from the fitted distribution. The use of mixture distributions should be considered and evaluated in situations in which single component distributions are unable to provide acceptable fits to the data, or in situations in which it is known that the data arise from a mixture of distributions.

The appearance of measurement error potentially affects any statistical analysis because the distribution representing the observed data set possibly deviates from the distribution which would have generated an error free data set. The method developed in this dissertation for conducting variability and uncertainty analysis in a data set with known measurement error demonstrated an approach to improve estimates of variability and uncertainty of the true value of a quantity if measurement errors appear in a data set.

8.2.2 Conclusions Based Upon Case Studies

The case study presented Part IV of this dissertation investigated the properties of the use of mixture distributions in the quantification of variability and uncertainty. The findings from these investigations include: (1) the size of mixing weight will influence the stability and accuracy of variability and uncertainty estimates; (2) bootstrap simulation results tend to be more stable normally for larger sample size; (3) when two components are well separated, the stability and accuracy of quantification of variability and uncertainty are improved, however, a higher uncertainty arises regarding percentile of mixture distributions coinciding with the separated region; (4) when two components are not well separated, a single component distribution may often be a better choice because it has fewer parameters and better numerical stability; and (5) Dependencies may exist in sampling distributions of parameters of mixtures and are influenced by the amount of separation between the components.

Investigations of effect of measurement error size on variability and uncertainty, which was presented in the Part V of this dissertation, indicates that when measurement error is a main source of uncertainty, bias between the distribution representing variability of the observed dataset and the distribution representing variability of the error free data set will appear; and that the shape and range of the probability band based upon the observed data set is largely different from the ones for the error free data set. These results suggest that ignoring the effect of measurement error on the quantification of variability and uncertainty will bring bias in the estimates of both variability and uncertainty of the quantity. An approach as presented in this dissertation, in which the contribution from measurement error and random sampling error are separately

characterized, should be used in the quantification of variability and uncertainty of a quantity if there are known measurement errors.

The results from the example case study demonstrating the use of the general approach for developing a probabilistic emission inventory, which is given in the Part VI of this dissertation, indicate that the range of uncertainty attributable to random sampling error is substantial in emission estimation for utility power plant emission sources. For individual technology groups, the range of uncertainty is as large as approximately plus or minus 30 percent, and for the total inventory the range of uncertainty is approximately plus or minus 15 percent. Identification of key sources of uncertainty in the inventory indicates that the largest contribution to uncertainty comes from one technology group. The finding is very useful in that, for example, if it were an objective to reduce uncertainty in the overall inventory, resources could be focused on collecting more or better data for the most sensitive technology group.

The uncertainty analysis results with the use of MOBILE5b in highway vehicle emission factors, which are demonstrated in the Part VII of this dissertation, indicate that there is substantial inter-vehicle variability in emission factors, typically spanning an order-of-magnitude or more, and that the range of uncertainty in fleet average emission factors was also found to be substantial. In many cases, the range of uncertainty is so large that traditional simplifying assumptions based upon normality and symmetry cannot be employed. For example, some emission factors were found to have uncertainty ranges of minus 80 to plus 220 percent of the mean value. The asymmetry reflects the fact that the emission factors are non-negative quantities and is influenced by both the large inter-vehicle variability in emissions and the relatively small sample sizes of data sets from

which the emission factors were developed. It suggests that it is necessary to collect more vehicle emission data in developing highway vehicle emission factors in order to reduce the uncertainty in emission estimation. The methods presented in this study allow decision-makers to assess the quality of their decisions and to decide on whether and how to reduce the uncertainty that most significantly affects the vehicle emissions.

Uncertainty analysis is much more efficient when done as an integrated part of model development, rather than in a post-hoc manner

8.3 Recommendations for Future Work

The recommendations for future work regarding methodologies and development of probabilistic emission inventories and AuvTool are presented in the subsections.

8.3.1 Methodologies

The methodologies presented in this dissertation are focused on dealing with variability and uncertainty in model inputs. Further development of methods for quantification of variability and uncertainty should be extended to take account into the uncertainty arising from model structure itself.

This dissertation focused on the methodologies based upon two parameter distributions and mixture distributions with two components. However, in practice, three parameter distributions, for example, three parameter lognormal distribution, are also used in some situations to describe variation of a quantity; two component mixture distributions are sometimes not enough to be a good representative of a data set. Therefore, there is a need to develop methodologies of variability and uncertainty based upon three parameter distributions and mixture distributions with three components or more in future research.

Further research should be done for the methods for dealing with variability and uncertainty with measurement errors. These include development of methods of constructing an error and free data set, and the methods for quantification of variability and uncertainty if there exists multiple and different measurement error models such as multiplicative error model used in an observed dataset.

It is often the case that some censored datasets are encountered in the quantification of variability and uncertainty. Censored datasets refer to the datasets in which there are some measurements below detect limit (DL) reported as “less than detection limit” or non-detects (NDs) rather than as numerical values. Therefore, the methods for fitting parametric probability distributions to censored data sets and methods for making inferences regarding the mean and other statistics taking into account the non-detect data need to be developed and should be incorporated into the general approach.

Limits in data often preclude the use of common statistical techniques to produce probabilistic estimates in variability and uncertainty analysis. A way to deal with the situation is to ask experts for their best professional judgment. The methods for the elicitation of subjective probability distributions from experts and characterization of uncertainty of expert judgment should be developed and be incorporated into the general approach for calculating a probabilistic emission inventory in future study.

8.3.2 Probabilistic Emission Inventories

The general approach for developing a probabilistic emission inventory presented in this dissertation should be applied to additional case studies from other emission source categories. It will help improve the completeness of the general approach. The work needed to enhance probabilistic emission inventories includes the selection and development of emission inventory models for various emission sources categories and

construction of databases to support the development of probabilistic emission inventories for additional source categories.

The development of emission inventories has common characteristics. For example, an emission inventory is typically a product of an emission factor and an activity factor. Therefore, it is possible to expand the prototype software tool AUVEE to an integrated probabilistic emission estimation system in which various emission source categories such as non-road vehicle emission source category and highway emission source category can be covered. The establishment of such a system will be a very useful supplement for the existing AP-42 and DARS systems in which qualitative techniques or semi-quantitative techniques are used in the emission estimation, and thus will improve the emission estimation using emission inventory.

8.3.3 Development of AuvTool

The current AuvTool can provide variability and uncertainty analysis for common probabilistic distributions such as normal, lognormal, beta, gamma, Weibull, uniform and symmetric triangle distributions, and for mixture normal and lognormal distributions with two components and for measurement error problems with known measurement errors in which additive error model is used. Future recommended development include variability and uncertainty analysis capacities for other distributions such as exponential, Pareto distributions and three parameter distributions such as asymmetric triangle distribution, three parameter lognormal or gamma distribution and additional mixture distributions with two components or more.

Future version of AuvTool should have a component that can deal with variability and uncertainty for the censored data sets. The current component that can handle

variability and uncertainty analysis for measurement error problem need to be enhanced in future development such as the use of multiplicative error model.

The batch method presented in the AuvTool for automatically selecting a "best" fitting distribution is based upon only one criterion. Uncritical application of this criterion can lead to potentially incorrect results. There is a need to enhance the criteria for selecting a "best" fitting distribution and to give users more participation to utilize the batch analysis feature so that a selected "best" distribution is reasonable in terms of both statistical and physical insights.

Because AuvTool is modular and based upon an object-oriented programming approach, it is possible to extend AuvTool as a common tool for quantifying variability and uncertainty in model inputs, propagating variability and uncertainty to model outputs, and further analysis of sampling results. Incorporation of an ability to allow users to interface their own models into AuvTool will help it become a more general tool useful any quantitative analysis fields where characterization of variability and uncertainty are needed in both model inputs and outputs.

(APPENDIX)

APPENDIX A.

1. Normal Distribution

As defined in Table 2-1, the parameters for the normal distribution are the arithmetic mean, μ , and the arithmetic variance, σ^2 . The MoMM estimator of the mean is the sample mean, \bar{X} . The MoMM estimator of the variance is the unbiased sample variance, s^2 (Morgan and Henrion, 1990; Casella and Berger, 1990).

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (\text{A-1})$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (\text{A-2})$$

2 Lognormal Distribution

The parameters of the lognormal distribution can be defined as: (1) the geometric mean, μ_g , and geometric standard deviation, σ_g , estimated by $\hat{\mu}_g$ and $\hat{\sigma}_g$, respectively; or (2) the mean and standard deviation of the logarithm of X , $\mu_{\ln(x)}$, and $\sigma_{\ln(x)}$, estimated by $\hat{\mu}_{\ln(x)}$ and $\hat{\sigma}_{\ln(x)}$, respectively (Morgan and Henrion, 1990; Casella and Berger, 1990).

$$\hat{\mu}_{\ln x} = \ln(\bar{X}) - \frac{1}{2} \hat{\sigma}_{\ln x}^2 \quad (\text{A-3})$$

$$\hat{\sigma}_{\ln x} = \sqrt{\ln(\bar{X}^2 + s^2) - 2 \ln(\bar{X})} \quad (\text{A-4})$$

In AuvTool, the mean of $\ln x$, $\mu_{\ln x}$, and the standard deviation of $\ln x$, $\sigma_{\ln x}$, are used as the parameters to define the lognormal distribution.

3. Beta Distribution

The beta distribution has two shape parameters. The parameters can be estimated through relationships with the sample mean and the unbiased sample variance, \bar{X} and s^2 (Hahn and Shapiro, 1967; Morgan and Henrion, 1990):

$$\hat{\alpha} = \bar{X} \left[\bar{X} \frac{(1 - \bar{X})}{s^2} - 1 \right] \quad (\text{A-5})$$

$$\hat{\beta} = (\bar{X} - 1) \left[\bar{X} \frac{(1 - \bar{X})}{s^2} - 1 \right] \quad (\text{A-6})$$

4. Gamma Distribution

The parameters of the gamma distribution are the shape parameter α , and the scale parameter β , where $\hat{\alpha}$ is an estimate of α , and $\hat{\beta}$ is an estimate of β . These parameters are estimated through relationships with the sample mean and unbiased sample variance, \bar{X} and s^2 (Morgan and Henrion, 1990; Casella and Berger, 1990).

$$\hat{\alpha} = \frac{\bar{X}^2}{s^2} \quad (\text{A-7})$$

$$\hat{\beta} = \frac{s^2}{\bar{X}} \quad (\text{A-8})$$

5. Weibull Distribution

For the Weibull distribution, the relationship between the parameters and the central moments of the data are (Morgan and Henrion, 1990):

$$\bar{X} = \hat{\beta} \Gamma \left(1 + \frac{1}{\hat{\alpha}} \right) \quad (\text{A-9})$$

$$s^2 = \hat{\beta}^2 \left[\Gamma \left(1 + \frac{2}{\hat{\alpha}} \right) - \Gamma^2 \left(1 + \frac{1}{\hat{\alpha}} \right) \right] \quad (\text{A-10})$$

There is no closed form solution for the MoMM estimator of the parameters of the Weibull distribution. Therefore, as an alternative, a parameter estimation method based upon regression analysis of a probability plot is used.

In the probability plot method, if a data set is reasonably described by a Weibull distribution, then the following transformation may be used to plot the data (Cullen and Frey, 1999):

$$\ln \left\{ \ln \left[\frac{1}{\bar{F}(x_i)} \right] \right\} = c \ln(x_i) - c \ln(k) \quad (\text{A-11})$$

where,

c = shape parameter

k = scale parameter

$$\bar{F}(x_i) = 1 - F(x_i) \quad (\text{A-12})$$

$\bar{F}(x_i)$ is the complementary CDF of x. An empirical estimate of the CDF can be obtained using Equation (2-1), presented by Hazen (1914). Thus, it is possible to plot the data set and to calculate the scale and shape parameters from the intercept and slope of a best fit

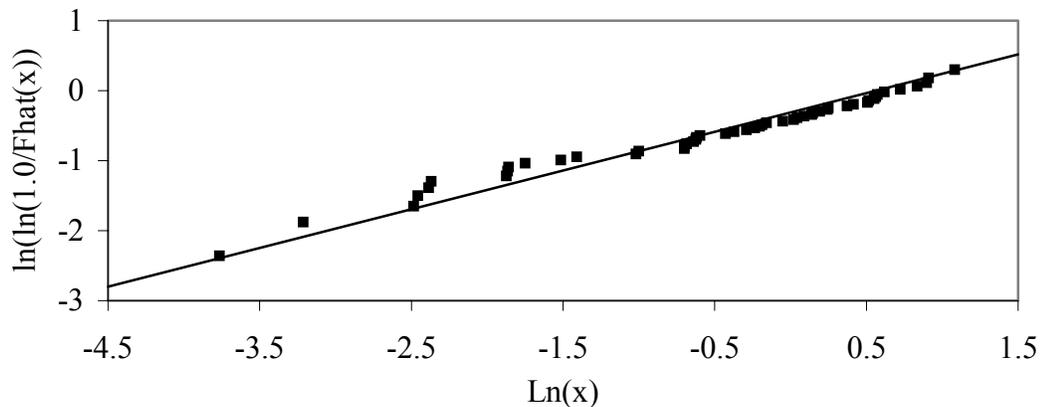


Figure 1. Example of a Probability Plot for a Weibull Distribution (n=50)

regression line obtained using conventional least-squares regression. An example is shown in Figure 2-4 for $n=50$. In this example, the best fit equation was:

$$\ln\left\{\ln\left[\frac{1}{\bar{F}(x_i)}\right]\right\} = 0.47313 \ln(x_i) - 0.3644 \quad (\text{A-13})$$

Therefore, the shape parameter is $c=0.47313$. The scale parameter can be found by solving the expression:

$$k = \exp\left(\frac{0.3644}{c}\right) \quad (\text{A-14})$$

From Equation (A-14), it can be inferred that k is equal to 2.17.

6. Uniform Distribution

The parameters of the uniform distribution are the endpoints, a and b , which are estimated by \hat{a} and \hat{b} . The parameter estimation formulae using MoMM are as follows (Morgan and Henrion, 1990):

$$\hat{a} = \bar{X} - \sqrt{3} s \quad (\text{A-15})$$

$$\hat{b} = \bar{X} + \sqrt{3} s \quad (\text{A-16})$$

7. Symmetric Triangle Distribution

The parameters of symmetric triangle distribution are a and b , which are estimated by \hat{a} and \hat{b} . MoMM parameter estimation formulas for these two parameters are (Morgan and Henrion, 1990):

$$\hat{a} = \bar{X} \quad (\text{A-17})$$

$$\hat{b} = \sqrt{6} s \quad (\text{A-18})$$

APPENDIX B.

1. Normal Distribution

Generation of random variables from a normal distribution is simplified by the fact that any normal distribution can be written in terms of the standard normal distribution, with a mean of zero and standard deviation of one. The symbol “ \sim ” denotes “is distributed as.” If $X \sim N(\mu, \sigma^2)$, and if $x' \sim N(0,1)$, which is the standard normal distribution, then

$$X = \mu + \sigma x' \quad (\text{B-1})$$

Therefore, it is only necessary to generate random numbers from the standard normal. Standard normal random samples can be generated using an acceptance-rejection method developed by Box and Muller (1958). In this method, two uniformly distributed $U(0,1)$ random variates, U_1 and U_2 , are used to generate two $N(0,1)$ random variates, X_1 and X_2 . The Box and Muller method is used to calculate X_1 and X_2 as follows:

$$\begin{aligned} X_1 &= \sqrt{-2 \ln U_1} \cos(2\pi U_2) \\ X_2 &= \sqrt{-2 \ln U_1} \sin(2\pi U_2) \end{aligned} \quad (\text{B-2})$$

However, a more efficient version of the Box-Muller method, called the polar method, was developed by Marsaglia and Bray (1964). The polar method is used in this study. The algorithm is presented in Law and Kelton (1991) as follows:

Step1: Generate U_1 and U_2 as independent and identically distributed (IID) uniform random samples on the interval $[0,1]$. Therefore, $U_1 \sim (0,1)$ and $U_2 \sim (0,1)$

Step2: Let $V_i = 2U_i - 1$ for $i = \{1, 2\}$, and let $W = V_1^2 + V_2^2$. If $W > 1$, go back to

Step 1. Otherwise, let $Y = \sqrt{(-2 \ln(W))/W}$, $X_1' = V_1 Y$, and $X_2' = V_2 Y$.

Step3: Then X'_1 and X'_2 are IID $N(0,1)$ random variates. $X_1 = \mu + \sigma X'_1$ and $X_2 = \mu + \sigma X'_2$ so that X_1 and X_2 are IID $N(\mu, \sigma^2)$.

Since two normal random samples are generated with each call of this subroutine, in principle the procedure only needs to be implemented once for every two normal distributions that are to be simulated. If U_1 and U_2 were truly IID random variables from a uniform distribution $U(0,1)$, then using X_1 followed by X_2 on subsequent calls to the subroutine would be valid. It has been shown, however, that if U_1 and U_2 are sequential pseudo random numbers (as is the case in this implementation) then X_1 and X_2 will fall on a spiral in (X_1, X_2) space, rather than being truly IID. In order to ensure that all normal random variates are truly IID in this implementation, only X_1 is used and X_2 is discarded. Another option would be to generate U_1 and U_2 from separate and independent pseudo-random number streams.

2. Lognormal Distribution

Lognormal random samples are generated by using a special property of the lognormal distribution. Namely, if $Y \sim N(\mu_{\ln x}, \sigma_{\ln x}^2)$, then $e^Y \sim LN(\mu_{\ln x}, \sigma_{\ln x}^2)$. Therefore, lognormal random samples are generated by the following algorithm:

Generate $Y \sim N(\mu_{\ln x}, \sigma_{\ln x}^2)$,

$X = e^Y$, so that $X \sim LN(\mu_{\ln x}, \sigma_{\ln x}^2)$,

Note that $\mu_{\ln x}$ and $\sigma_{\ln x}^2$ are the mean of $\ln x$ and standard deviation of $\ln x$.

3. Beta Distribution

The method used in this study for generating beta random samples relies upon a special property of the beta distribution. The beta distribution can be described as a ratio comprised of gamma distributions. If $Y_1 \sim G(\alpha,1)$ and $Y_2 \sim G(\beta,1)$ and Y_1 and Y_2 are

independent, then $X = Y_1/(Y_1+Y_2) \sim B(\alpha, \beta)$ (Law and Kelton, 1991). Thus, the methods described for generating random samples from a gamma distribution are used as a basis for generating random samples for the beta distribution

4. Gamma Distribution

Like the normal and lognormal distributions, the gamma distribution has no closed form solution for its CDF or inverse CDF. Therefore, the method of inversion is not feasible for generating random variables in this case. An acceptance-rejection method is used here to generate gamma random variables.

In generating $G(\alpha, \beta)$ random variables, it is noted that if $X' \sim G(\alpha, 1)$, then $X = \beta X' \sim G(\alpha, \beta)$. Therefore, only the $G(\alpha, 1)$ distribution needs to be simulated and the results can be easily transformed to that of any $G(\alpha, \beta)$ distribution. Furthermore, a gamma distribution with $\alpha = 1$, $G(1, \beta)$, is simply an exponential distribution with a mean of β . Exponential random variables can be easily generated by the method of inversion as shown below (Morgan and Henrion, 1990):

$$X = -\frac{1}{\beta} \ln(U) \quad (\text{B-3})$$

where U is a random sample from the $U(0,1)$ distribution and β is the parameter of the exponential distribution.

Gamma distributions for which $\alpha < 1$ are shaped significantly differently than gamma distributions for which $\alpha > 1$. Therefore, two distinct acceptance-rejection algorithms are necessary.

For $\alpha < 1$, an acceptance-rejection algorithm by Ahrens and Deiter is used in this study. A description of this method is provided in Law and Kelton (1991), where the following algorithm is also presented:

- Step 1. Let $b = (e + \alpha)/e$ (e is a constant, and $e = \exp(1.0) = 2.718282$)
- Step 2. Generate $U_1 \sim U(0,1)$, and let $P = bU_1$. If $P > 1$, go to step 4. Otherwise proceed to Step 3
- Step 3. Let $Y = P^{1/\alpha}$, and generate $U_2 \sim U(0,1)$. If $U_2 \leq e - Y$, return $X = Y$.
Otherwise, go back to Step 1.
- Step 4. Let $Y = -\ln[(b - P)/\alpha]$ and generate $U_2 \sim U(0,1)$. If $U_2 \leq Y^{\alpha-1}$, return $X = Y$. Otherwise, go back to Step 1.

For $\alpha > 1$, a modified acceptance-rejection algorithm by Cheng (1977) is used to sample random samples from a Gamma distribution. A description of the method is provided in Law and Kelton (1991). Only the algorithm is presented here:

- Step1. Let $a = 1/\sqrt{2\alpha-1}$, $b = \alpha - \ln 4$, $q = \alpha + 1/a$, $\theta = 4.5$, and $d = 1 + \ln \theta$.
- Step 2. Generate U_1 and U_2 as IID $U(0,1)$.
- Step 3. Let $V = a \ln[U_1/(1 - U_1)]$, $Y = \alpha e^V$, $Z = (U_1^2 U_2)$, and $W = b + qV - Y$.
- Step 4. If $W + d - \alpha Z \geq 0$, return $X = Y$. Otherwise, proceed to Step 5.
- Step 5. If $W \geq \ln Z$, return $X = Y$. Otherwise, go back to Step 1.

5. Weibull Distribution

The CDF for the Weibull distribution can be written as (Morgan and Henrion, 1990):

$$F(x) = 1 - \exp^{-(x/k)^c} \quad (\text{B-4})$$

A random sample, X , from a $W(k,c)$ can therefore be generated directly by the method of inversion using the inverse CDF:

$$X = F^{-1}(U) = k[-\ln(1-U)]^{1/c} \quad (\text{B-5})$$

where U is a random sample from the $U(0,1)$ distribution.

6. Uniform distribution

The method of inversion is used in this study for generating uniform distributions with any arbitrary endpoints. The method is as follows (Morgan and Henrion, 1990):

$$X = a + (b - a)U \quad (\text{B-6})$$

where U is a random sample from the $U(0,1)$ distribution.

7. Symmetric Triangle Distribution

The method of inversion is used in this study for generating symmetric triangle distribution, as follows as (Morgan and Henrion, 1990):

$$\begin{aligned} X &= (a - b) + b(2U)^{1/2} & 0 \leq U \leq 0.5 \\ X &= (a + b) - b(2.0 - 2U)^{1/2} & 0.5 < U \leq 1.0 \end{aligned} \quad (\text{B-7})$$

where U is a random sample from the $U(0,1)$ distribution.

REFERENCES FOR APPENDIX A & B

- Box, G.E.P., M.E. Muller, 1958, "A Note on the Generation of Random Normal Deviates," *Annals of Mathematical Statistics*, 29: 610-611.
- Casella, G., R. L. Berger, 1990, *Statistical Inference*, Duxbury Press: Belmont, CA
- Cheng, R. C. H., 1977, "The Generation of Gamma Variables with Non-Integral Shape Parameter," *Applied Statistics*, 26:71-75.
- Cullen, A.C., H.C. Frey, 1999, *Use of Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press: New York.
- Hahn, G.J., S.S. Shapiro, 1967, *Statistical Models in Engineering*, John Wiley and Sons: New York.
- Hazen, A., 1914, "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply," *Transactions of the American Society of Civil Engineers*, 77: 1539-1640.
- Law, A.M., W.D. Kelton, 1991, *Simulation Modeling and Analysis* 2d ed., McGraw-Hill: New York.
- Marsaglia, G., T.A. Bray, 1964, "A Convenient Method for Generating Normal Variables," *SIAM Review*, 6:260-264.
- Morgan, M.G., and M. Henrion, 1990, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York.

