In *Proceedings, Annual Meeting of the Air & Waste Management Association*, June 24-28, 2001, held in Orlando, Florida, and published byA&WMA, Pittsburgh, PA

Quantitative Analysis of Variability and Uncertainty in Emission Estimation: An Illustration of Methods Using Mixture Distributions

2001-Abstract #11

Junyu Zheng

Research Assistant, Department of Civil Engineering, North Carolina State University

H. Christopher Frey

Associate Professor, Department of Civil Engineering, North Carolina State University

ABSTRACT

Air pollutant emission inventories are a vital component of environmental decision-making. Errors in emission factor estimation can lead to errors in emission inventory estimation. Potential sources of error include unaccounted for variability and uncertainty. Variability refers to diversity over time or space. Uncertainty is a lack of knowledge about the true value of a quantity. Probability distribution models can be used to describe variability in a data set and as a starting point for characterizing uncertainty, such as for mean values. Mixture distributions have the potential to be useful in the quantification of variability and uncertainty because they can improve the goodness of fit to a dataset compared to the use of a single parametric distribution. In this paper, parameter estimation of mixture distributions is discussed. An approach for quantifying the variability and uncertainty based on mixture distributions by using Bootstrap simulation is developed. 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 method. Results from the use of single parametric distributions are compared with results from the use of a mixture distribution. The case study results indicate that a mixture lognormal distribution is a better fit to the selected case compared to single distributions. Furthermore, the estimate of the range of uncertainty in the mean is narrower with the mixture distribution than with the single component distribution, indicating that the mixture distribution has the potential to yield more "efficient" statistical estimates. This project is one component of a larger effort aimed at developing improved methods for characterizing uncertainty in emission inventories.

Keywords: Variability, uncertainty, mixture distribution, bootstrap simulation, parameter estimation

.

1.0 INTRODUCTION

Air pollutant emission inventories are a vital component of environmental decision-making. In practice, emission inventories are obtained by multiplying an emission factor (e.g., grams of pollutant per unit of product produced) with an activity factor (e.g., number of units of product produced in one year) to obtain total emissions for each emission source category, and then by summing emissions over all source categories. Emission factors are typically assumed to be

representative of the average emission rate from a population of pollutant sources in a specific category for a given time period [1], however, there may be uncertainty in the population average emissions because of random sampling error, measurement errors, or possibly because the sample from which the emission factor was developed was not a representative sample. These errors in the emission factor or activity factor can lead to errors in emission inventory estimation. Errors in emission inventories may have an adverse impact on decision making regarding air quality management. 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 to a source category where emissions are under-estimated, then air quality objectives cannot be achieved in an efficient and cost-effective manner.

1.1 Variability and Uncertainty

In this paper, we focus primarily on evaluation of uncertainty in emission factors, because these factors are a key input to emission inventories. Emission factors are often reported as single numbers. However, in reality, emission rates vary from one specific facility to another within a given source category, and even vary at a single facility from one time to another. The variability in emissions is observable based upon measurements made at multiple facilities, or based upon repeated measurements at a single facility. Variability refers to heterogeneity across different elements of a population over time or space. For example, emissions may vary from one power plant to another (inter-plant variability) because of differences in design and operating conditions. Moreover, emissions may vary for a single plant over time (intra-plant variability) because of temporal changes in feedstock composition or operating conditions (e.g., fuel to air ratio, furnace temperature, etc.). Variability can be represented by a frequency distribution showing the variation in a characteristic of interest over time, space.[2]

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; (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. For example, there may be uncertainty regarding the true average emission rate for a source category. Uncertainty can be quantified as a probability distribution representing the likelihood that the unknown quantity falls within a given range of values.[2]

1.2 Increasing Demand for Probabilistic 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 recently released a report on mobile source emissions estimation that calls for new efforts to quantify uncertainty in emissions.[3] 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, has developed a guideline document on Monte Carlo methods and has included guidance regarding probabilistic analysis in its most recent draft of Risk Assessment Guidance for Superfund.[4, 5, 6] In addition, there is a growing

track record of the demonstrated use of quantitative methods for characterizing variability and uncertainty applied to emission inventories. 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.[1, 3, 8] 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.[7, 8, 9, 10] 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.[11]

1.3 Importance of Input Assumptions in Probabilistic Analysis

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. Methods for developing input distributions for model inputs, propagating distributions through models, and analyzing the probabilistic results for model outputs, are well-described elsewhere.[12, 13]

In this paper, we focus on one typical situation and challenge encountered when trying to represent an emission inventory input with a probability distribution. Parametric distributions are a compact means for representing either variability or uncertainty in a quantity. 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.[12]

1.4 Benefits of Mixture Distributions

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 1998, 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.[4] In this work, the preference is for use of parametric distributions. 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 an emission factor or activity factor. However, in practice, 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. However, 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.

1.5 Experience with Mixture Distributions in Environmental Analysis

Mixture distributions have been extensively used in a wide variety of important practical situations because they provide a powerful way to extend common parametric families of distributions to fit datasets not adequately fit by single common parametric distributions. Mixture models have been used in the physical, chemical, social science, biological and other fields. As examples, Hariris [14] applied mixture distributions to modeling crime and justice data, Kanji [15] described wind shear data using mixture distributions. In human exposure and risk assessment, David E. Burmaster [16] used mixture lognormal models to re-analyze data set collected by the U.S. EPA for the concentration of $Radon^{222}$ 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.

1.6 Quantification of Uncertainty in Statistics Estimated Based Upon Mixture Distributions

Frey and Rhodes have presented a two-dimensional probabilistic modeling framework for simultaneously quantifying variability and uncertainty in the context of emissions estimation by using bootstrap simulation[7,17]. However, these studies are focused on the use of single component distributions to represent variability in model inputs. The general framework for quantification of variability and uncertainty as presented by Frey and Rhodes is adopted here. However, the approach is extended to include mixture distributions as one method for representing variability in a model input.

Because a mixture distribution has a more complicated mathematical form and more parameters then a single component distribution, the process of parameter estimation is typically more challenging. For example, there are often no analytic parameter estimators available for any mixture distributions. Therefore, 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.

1.7 Purpose of this Paper

The purpose of this paper is to: (1) discuss parameter estimation of mixture distribution models; (2) demonstrate the methodology for the quantification of variability using mixture distributions; (3) demonstrate the methodology for quantification of uncertainty in statistics inferred from the mixture distribution, such as the mean; (4) illustrate the methods using a case study; and (5)

comparatively analyze the results from single distribution-based estimates and the mixture distribution-based estimates.

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 Distributions

According to the definition from Titterington *et al.*[18], 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:

Equation 1.

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_k f_k(x) \qquad \text{for } x \text{ in the sample space}$$

With

$$w_j > 0 \qquad \text{for } j = 1, \ldots, k$$

And

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

Where

- $f(x)$: Probability density function for the mixture model
- $f_j(x)$: Probability density function (PDF) for the jth component of the mixture.
- w_j : The mixing weight for the jth component of the mixture

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 means 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:

Equation 2.

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

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.

The simplest case of a mixture distribution is one that has only two components. A mixture model with two components can be expressed in the following form:

Equation 3.

$$f(x) = wf_1(x) + (1-w)f_2(x)$$

with $0 < w < 1$.

A mixture distribution comprised of two components, each of which is a two-parameter distribution, will have a total of five parameters. For example, a mixture of two lognormal distributions will have a weight parameter, a logarithmic mean and standard deviation for the first component, and a logarithmic mean and standard deviation for the second component.

In this paper, we will focus on discussion of mixture distributions with two components as a means for quantifying variability in a dataset, and the application of bootstrap simulation as a means for quantifying uncertainty in statistics estimated based upon the mixture. Although the analysis here focused on two component mixtures, the method introduced here can be easily expanded to mixture models with more than two components.

2.2 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. 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.[19, 20] However, the assumption cannot be met in most cases.

With the advent of high-speed computers, there has been more interest in recent years in using likelihood estimation for the parameters in a mixture distribution. Therefore, MLE will be considered as the preferred method for estimating parameters in a mixture distribution in this paper. 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. 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:

Equation 4.

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

where

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

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 jth component in a mixture distribution

There are three approaches that can be used to find the maximum of Equation (4) 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.*.[21] 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,[19] and its results are strongly depend upon the initial guesses assumed for the parameters.[20] 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.[19, 20] 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:

Equation 5.

$$\text{Maximize} \qquad L = \sum_{i=1}^{n} \ln\left[w\, f_1(x_i|\mu_1,\sigma_1) + (1-w) f_2(x_i|\mu_2,\sigma_2)\right]$$

$$\text{Subject to} \qquad 0 \le w \le 1$$

$$\mu_1,\ \mu_2 > 0$$

$$\sigma_1,\ \sigma_2 > 0 \qquad \text{for mixture log normal}$$

Where n= the number of samples

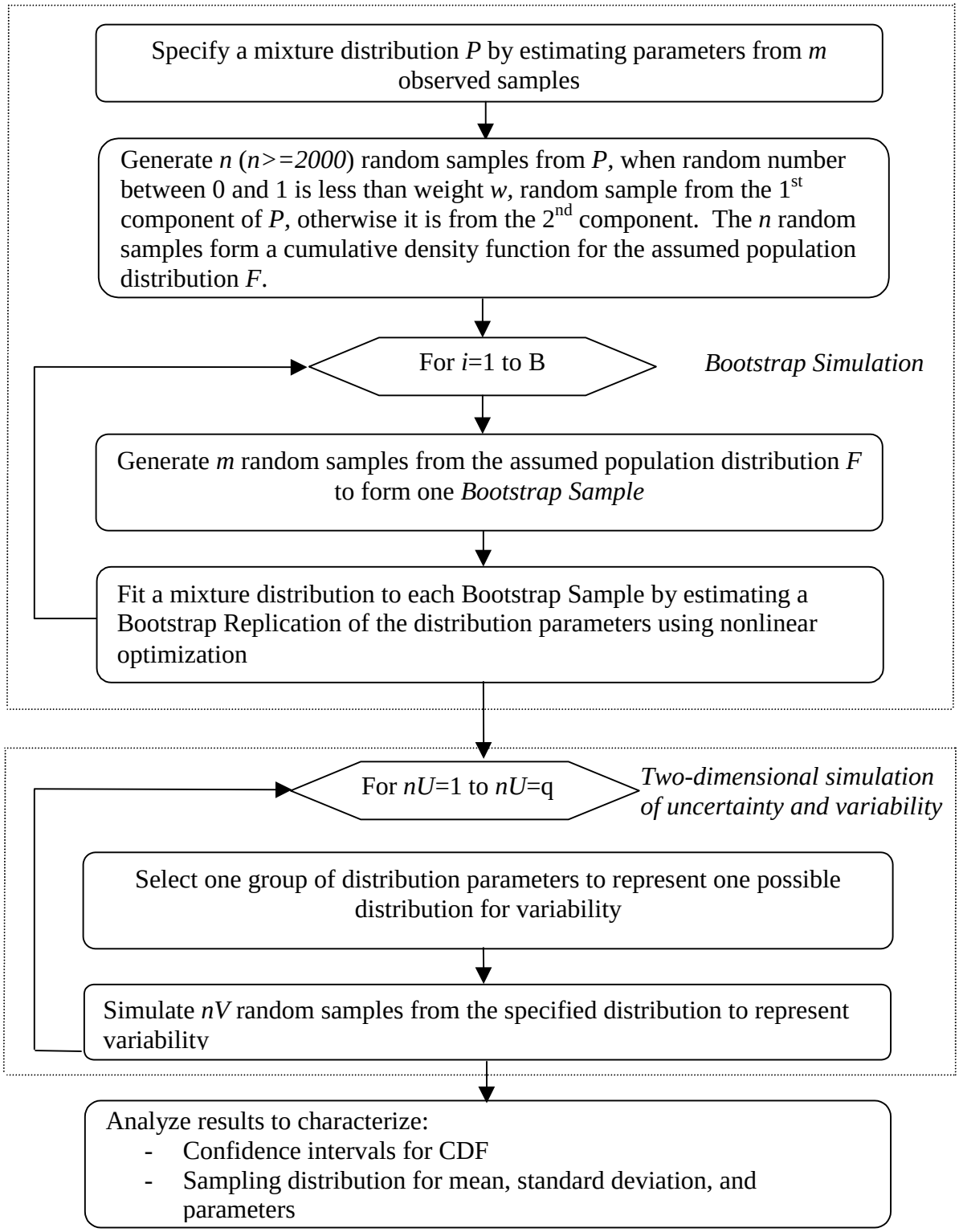
The optimization problem here is a multidimensional constrained one. A variety of methods are available to solve such problems. These include: the downhill simplex method; the direction-set method, of which Powell's method is the prototype;[22] the penalty function method; and others. In this paper, Powell's method is employed. This method is relatively easy to program and provides good results.

2.3 Quantification of Uncertainty in Statistics Based Upon Mixture Distributions

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.[23] 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 [24] developed a bootstrap procedure to calculate the upper and lower confidence bounds for the mean of a log-normal distribution based on complete samples. Freedman and Peters [25] 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 [25]. 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.[12] 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

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



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.

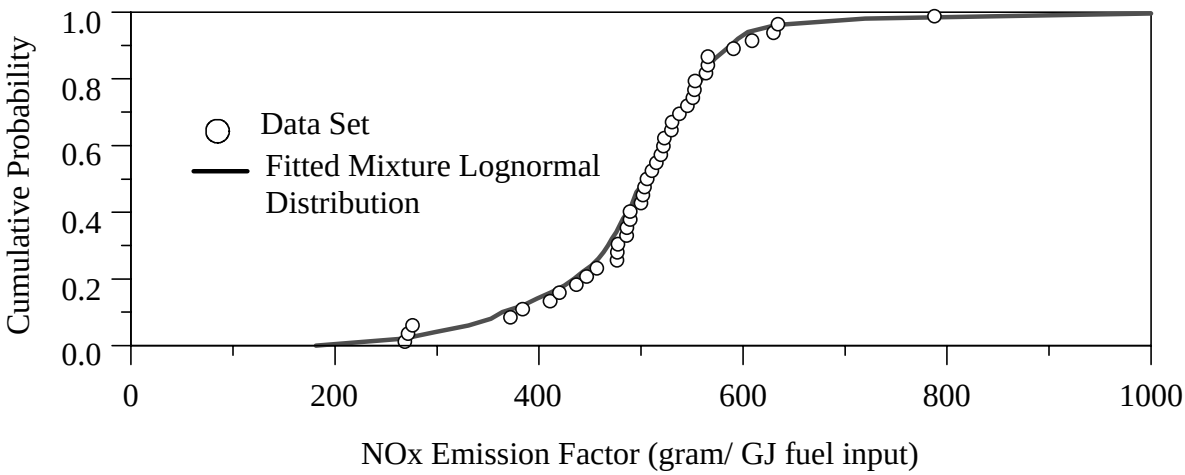
While there are standard numerical methods for drawing random samples from single component parametric distributions (e.g., see Cullen and Frey [12] for an overview), the methods for drawing random samples from mixture distributions are more complicated in the context of bootstrap simulation. Although 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 BC_a.[26] The percentile method is possibly the most frequently used in practice though the theoretical justification for this method is the weakest;[27] 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 BC_a intervals are comparable in that both have been demonstrated theoretically to be "second-order correct" for one-sided intervals in some relatively simple situations.[27] In this paper, for simplicity and because it is the most widely used method in practice, the percentile method is used.

Figure 2. Mixture lognormal distribution fitted to six month average NO_x emission factor data for T/LNC1 technology group (n=41).



3.0 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.[28] This scenario was chosen because: (1) this dataset can not be fit well by any single distribution; (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).

3.1 Parameter Estimation for the Fitted Distribution

A mixture distribution with two lognormal components was fit to the case study dataset. In addition, some single component distributions, such as Normal, Lognormal and Weibull, were fit to the data for comparison. MLE was used to estimate parameters in all cases. 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 2 through 5 for the two-component Lognormal mixture distribution, the single component Normal distribution, the single component Lognormal distribution, and the single component Weibull distribution, respectively.

Figure 3. Normal distribution fitted to six month average NO_x emission factor for T/LNC1 technology group (n=41).

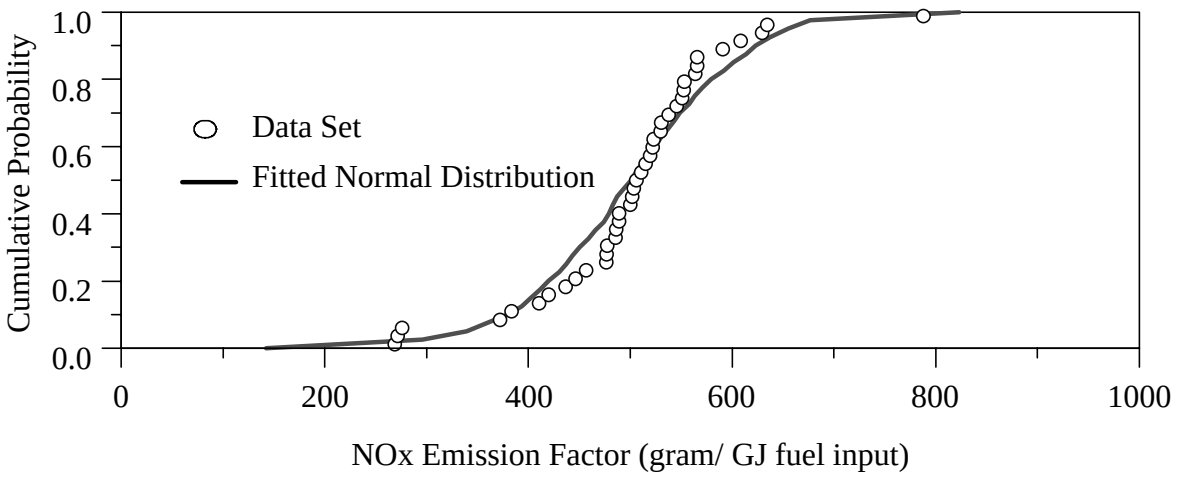


Figure 4. Lognormal distribution fitted to six month average NO_x emission factor data for T/LNC1 technology group (n=41).

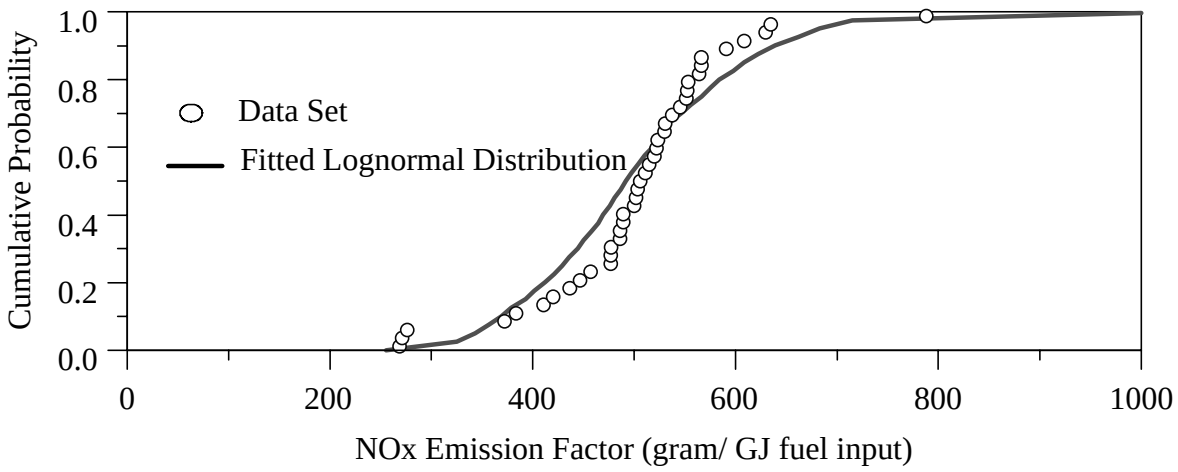


Figure 5. Weibull distribution fitted to six month average NO_x emission factor data for T/LNC1 technology group (n=41).

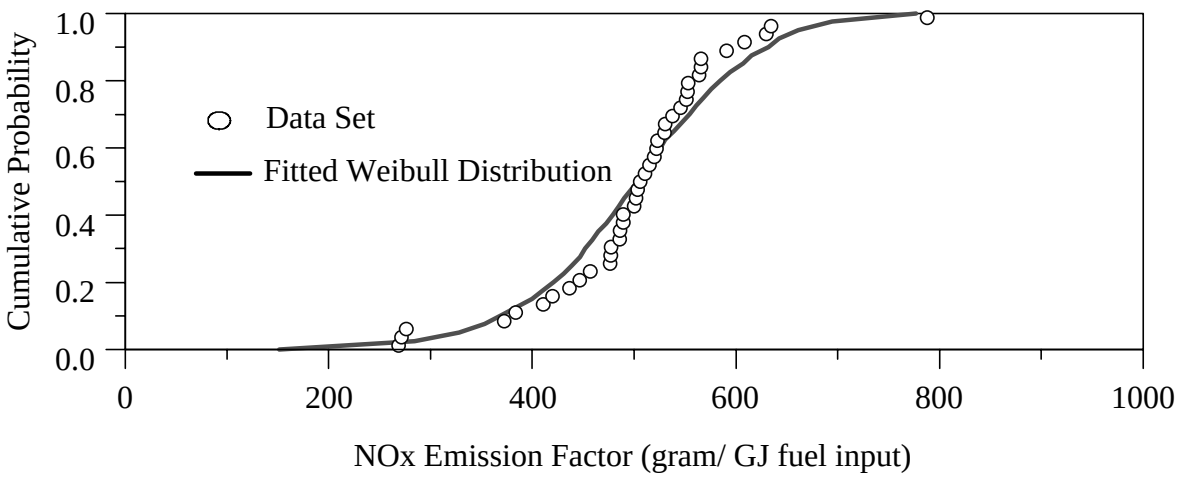
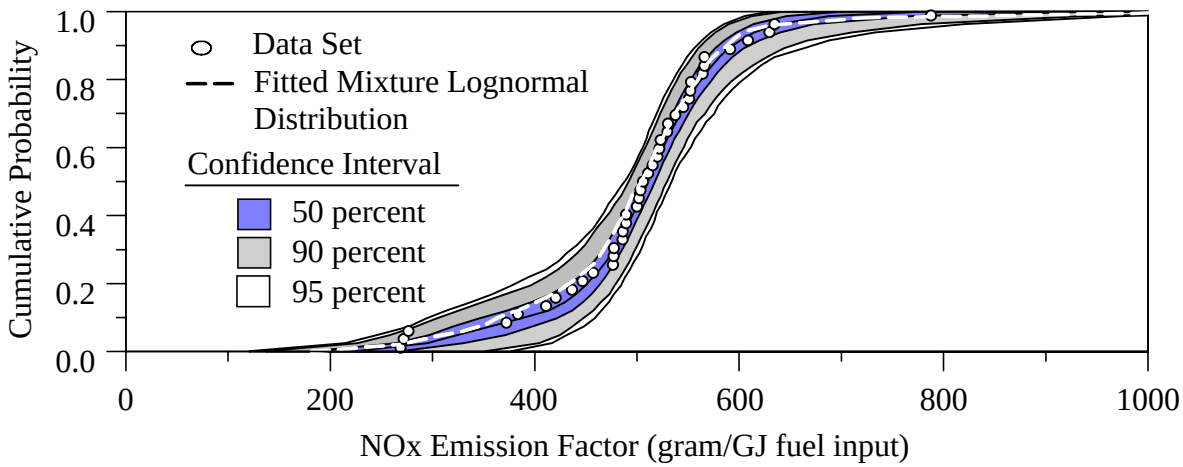


Figure 6. Probability band for fitted mixture lognormal distribution.



3.1 Parameter Estimation of Fitted Distribution

Figure 2 displays a good agreement between the fitted mixture distribution and the observed data set. In contrast, Figures 3, 4, and 5 display significant disagreements between the fitted Normal, Lognormal, and Weibull distributions, respectively, and the observed data set. All three single component distributions tend to underestimate the data in the lower range of cumulative probability (approximately from a cumulative probability of 0.2 to 0.5) and to overestimate the data in an upper range of cumulative probability (e.g., from a cumulative probability of 0.6 to at least 0.95). For example, in the case of the Normal distribution, the 25th percentile is underestimated to be 440 g/GJ compared to an observed value of 470 g/GJ, and the 85th percentile is overestimated to be 600 g/GJ compared to an observed value of 550 g/GJ. Similarly discrepancies exist with the Lognormal and Weibull distributions. Based upon visual inspection of the fitted distributions, it appears that the mixture distribution is a substantially better fit compared to the single component distributions.

3.2 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 6 through 9 for the two-component Lognormal mixture distribution, the single component Normal distribution, the single component Lognormal distribution, and the single component Weibull distribution, respectively. 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

Figure 7. Probability band for fitted normal distribution.

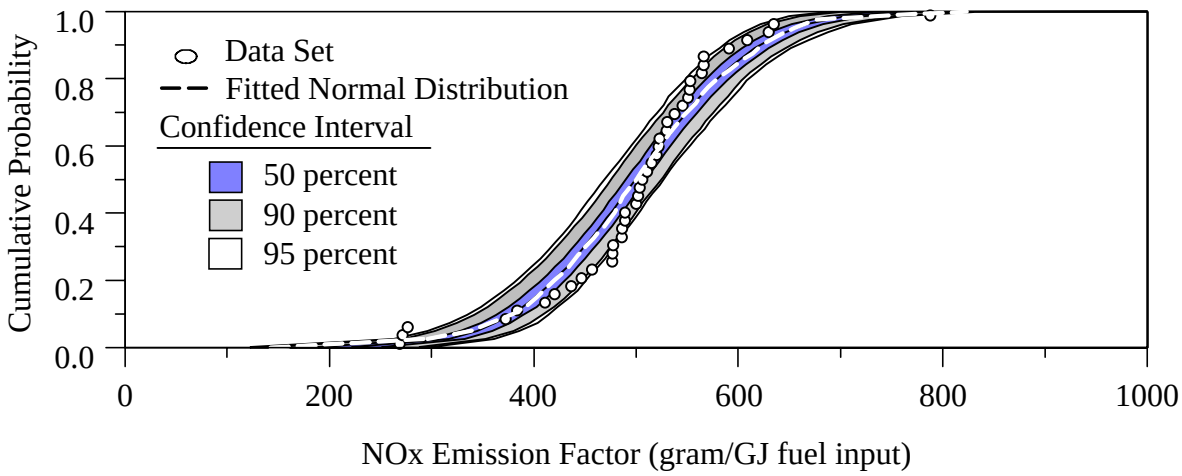


Figure 8. Probability band for fitted lognormal distribution

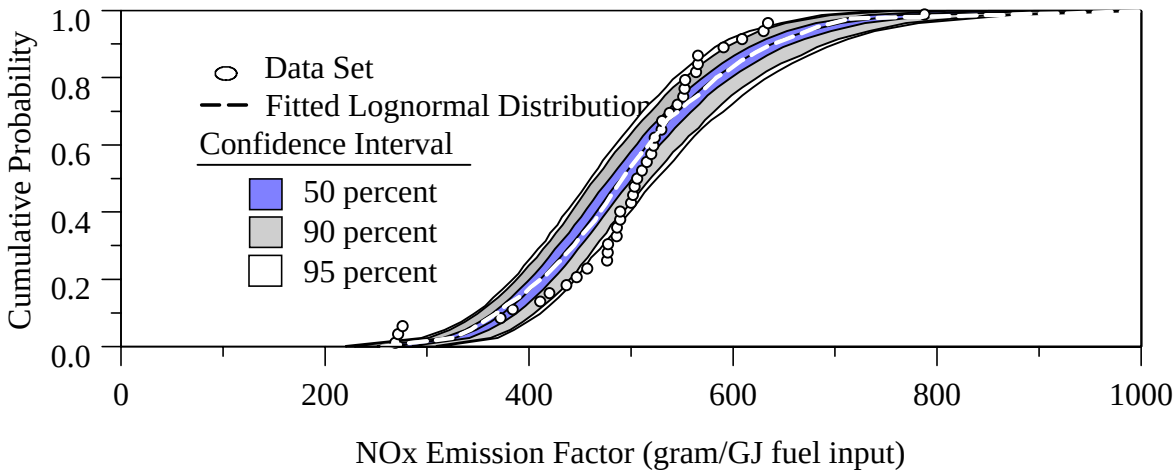
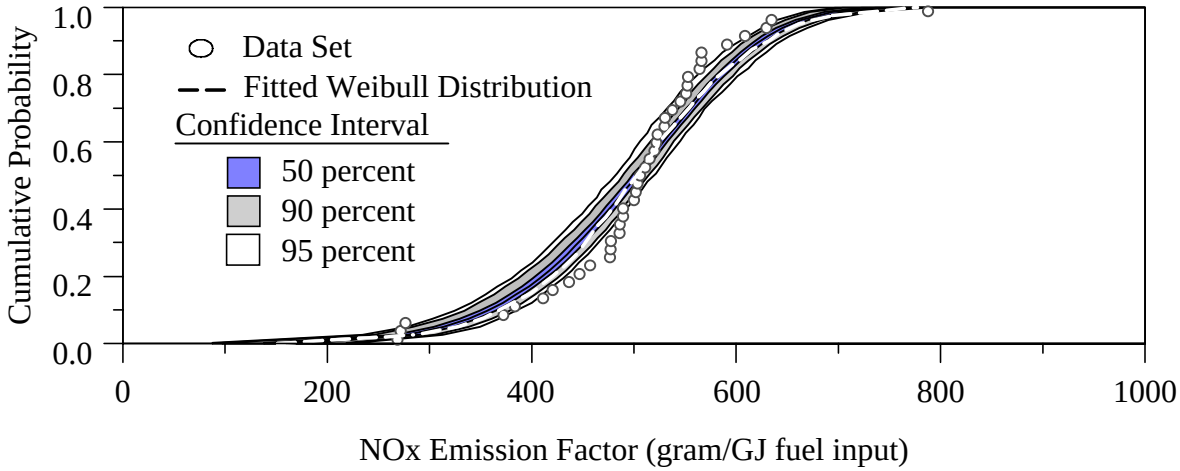


Figure 9. Probability band for fitted normal distribution



data are outside of the 50 percent confidence interval. This indicates that the data are highly consistent with the assumed mixture distribution. In contrast, 15 percent of the observed data fall outside the 95 percent confidence interval of the fitted Normal distribution, 22 percent of the observed data fall outside the 95 percent confidence interval of the fitted Lognormal distribution, and 49 percent of the observed data fall outside the 95 percent confidence interval of the fitted Weibull distribution. The large percentages of data falling outside of the 95 percent confidence interval in the cases of the Normal, Lognormal, and Weibull distributions suggest that these three single parametric distributions are not good assumptions regarding the unknown population distribution, in that they are inconsistent with the observed data.

Figures 6 through 9 also illustrate that there is better agreement in the tails of the mixture distribution with the observed data than with single component distributions. This agreement is important as uncertainty is often large in the tails of the distribution and failure to properly quantify uncertainty in the tails can potentially lead to biased results.

The Kolmogorov-Smirnov (K-S) test was used to evaluate the goodness of fit of the different hypothesized distributions with respect to the selected case study datasets in this paper. K-S test statistics for the different fitted distribution are: 0.06 for the mixture Lognormal distribution, 0.15 for the Normal distribution, 0.19 for the Lognormal distribution and 0.15 for the Weibull distribution. At a significant level of 0.05, the critical value of K-S test statistic is 0.21 (n=41). Thus, all of the hypothesized distributions in this paper can pass the K-S test. However, the mixture lognormal distribution has the smallest discrepancy in the estimated cumulative probability, as indicated by the smallest value of the K-S statistic.

A limitation of goodness-of-fit tests is that they often do not have statistical power when used in situation with small sample size, such as in this case. However, in evaluating both the graphical representation of the fitted distributions and the values of the K-S statistic, it appears that the mixture distribution is a better fit than the single component distributions.

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 1 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.

Table 1. 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

3.3 Uncertainty in the Mean NO_x Emission Factor

Table 2 summarizes results for uncertainty in the mean for the four fitted distributions considered in the case study. As shown in the table, there is some variation in the estimates of 95 percent confidence interval of the mean depending upon which distribution was fit to the observed data. The mixture distribution is the most accurate of the four cases shown. The estimated mean values are similar for all four cases; however, the confidence intervals for the mean differ among the four cases. In particular, the 95 percent confidence interval for the mean 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. 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.

Table 2. Comparison of Selected Statistics of the 95 Percent Confidence Interval for the Mean Based Upon a Mixture Distribution and Three Single Component Parametric Distributions.

Distribution Types	Mean	Lower 2.5th Confidence Level	Upper 97.5th Confidence Level	Relative Uncertainty in Mean (-) %	(+) %
Lognormal - Lognormal	503.4	470.9	535.6	-6.5	6.4
Normal	505.4	468.7	541.0	-7.3	7.0
Lognormal	503.0	461.2	538.4	-8.3	7.0
Weibull	500.0	457.0	541.9	-8.6	8.4

* Negative Random Error= (2.5th Percentile –Mean)/Mean
Positive Random Error=(97.5th Percentile –Mean)/Mean

4.0 DISCUSSION

In the case study, a mixture lognormal distribution with two components was used to demonstrate the methodology of quantification of variability and uncertainty using mixture distributions. However, the methodology introduced in this paper can be easily extended to other mixture models or more components. The parameter estimation method, based upon nonlinear optimization, can be extended to include more components. Similarly, the numerical simulation method for characterizing variability and uncertainty can be extended to more components. 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. 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.

The use of five parameters instead of two parameters in the case of the two-component mixture versus the single component distributions has an impact on the confidence interval of the fitted distribution. In particularly, the weighting parameter is a source of uncertainty as reflected in the confidence interval for the mixture distribution. 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. 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. On the other hand, the mixture distribution produced an excellent fit to the data, such that nearly all of the data were enclosed by the 50 percent confidence band of the fitted distribution. Thus, there is a clear trade-off between improved fit between the central tendency of the fitted distribution versus the observed data and the width of the confidence interval of the fitted distribution, at least for some statistics, as the number of parameters increases. As Leoroux [29] 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. Thus, it is important to have as many components in the mixture as needed to obtain a reasonable fit to the data, but not to have too many.

Even though there is a "bulge" in the confidence interval of the fitted mixture distribution associated with the weight parameter, the confidence interval can be relatively narrow for other portions of the cumulative distribution and for some statistics. In the case study, the narrowest confidence interval for the mean was obtained from the mixture distribution. This result indicates that, in this case, the mixture distribution yielded the most statistically efficient estimate of the sampling distribution of the mean. The results would vary in other cases. For example, if the weight parameter led to an inflection point at a location similar to the mean, the confidence interval for the mean could be comparatively wide.

5.0 CONCLUSION

This paper demonstrates a method for improving the characterization of variability and uncertainty in environmental data and models based upon the use of mixture distributions. A method for fitting mixture distributions to data using MLE and nonlinear optimization was successfully applied to a case study for an actual data set regarding NO_x emissions for a selected type of coal-fired power plant. Numerical methods based upon bootstrap simulation were successfully applied to characterize uncertainty in statistics estimated from the fitted distribution. The results from the use of a mixture distribution were compared to those from more commonly used single component distributions. While there is some effect on uncertainty estimates associated with the increased number of parameters of the mixture, overall the fit to the observed data was substantially better. In the example case study, the confidence interval for the mean based upon the mixture was narrower than that obtained based upon other distributions, indicating that mixtures can lead to more efficient estimates of uncertainty.

As noted in the discussion, the results from the case study are not generalizable to all data sets. For example, it is possible to have situations in which the uncertainty in the mean could be larger based upon a mixture than based upon a single component distribution. However, a key benefit of mixture distributions is greater fidelity to the original data set with respect to the shape of the assumed population distribution, even though the range of uncertainty can be influenced by the presence of additional parameters.

As part of future work, the method for characterizing variability and uncertainty will be evaluated in a series of systematic case studies, in which the effects of sample size, weights, and amount of separation between the components will be evaluated. In the longer term, other methods for fitting mixture distributions to data, and for simulating uncertainty, could be evaluated or developed.

In summary, 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 this work, 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.

6.0 ACKNOWLEDGMENTS

This work was supported by the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) grants program via grants R826766 and R826790. Although the research described in this paper has been funded wholly or in part by the U.S. EPA, this paper has not been subject to any EPA review. Therefore, the paper does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

7.0 REFERENCES

1. EPA, *Compilation of Air Pollutant Emission Factors*, AP-42 5th Ed. and Supplements, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1985.

2. 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.

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

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, RAGS 3A - Process for Conducting Probabilistic Risk Assessment, Draft, U.S. Environmental Protection Agency, Washington, DC, December 1999.

7. 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): 762-797

8. 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.

9. 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, June 2001 (in press).

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 Mangement Association, Orlando, FL, June 2001 (in press)

11. Radian, *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, 1996.

12. 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.

13. Morgan, M.G., and M. Henrion; *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press: New York,1999.

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

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

16. Burmaster, D.E. and K.M. Thompson; Fitted second-order parametric distributions to data using maximum likelihhod estimation, *Human and Ecological Risk Assessment* **1998**, 3(2): 235-255

17. Frey, H.C., *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, 1992.

18. Titterington, D.M, A.F.M, Smith, and U.E., Makov; *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons: New York, NY,1985.

19. Redner, R.A., Walker, H.F.; "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review* **1984**, 26:195-239

20. Everitt, B.S. and D. J. Hand; *Finite Mixture Distributions*, Chapman & Hall, London, UK,1981.

21. 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* **1997**, 39:1-38

22. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery B.P.; *Numerical Recipes in FORTRAN*, Cambridge University Press, New York, NY, 1992.

23. Efron,B., Tibshirani,R.J.; *an Introduction to the Bootstrap*, Chapman & Hall, London, UK,1993.

24. Angus, J.E., "Bootstrap one-sided confidence intervals for the log-normal mean," *Statistician* **1994**, 43:395-401.

25. Freedman, D.A., and Peters, S.C.; " Bootstrapping a regression equation: some empirical results," *J. of the American Statistical Association* **1984**, 79:97-106.

26. Burr, D., "A comparison of certain confidence intervals in the Cox model," *J. of the American Statistical Association* **1994**, 89:1290-1302.

27. Hall, P., "Theretical comparison of Bootstrap confidence intervals," *The Annals of Statistics* **1988**, 16:927-985.

28. 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, June 2001 (in press).

29. Leroux, B., "Constant estimation of a mixing distribution," *Annuals of Statistics* **1992**, 20:1350-1360.