Fitting Second-Order Finite Mixture Models to Data with Many Censored Values Using Maximum Likelihood Estimation

David E. Burmaster
Alceon Corporation
PO Box 382669 Harvard Square Station
Cambridge, MA 02238-2669
deb@Alceon.com

Andrew M. Wilson
Harvard School of Public Health
677 Huntington Avenue
Boston, MA 02115
awilson@hsph.harvard.edu

Key Words

variability, uncertainty, second-order random variables, finite mixture models, maximum likelihood estimation, Radon 222

Abstract

Finite mixture models, i.e., weighted averages of parametric distributions, provide a powerful way to extend parametric families of distributions to fit data sets not adequately fit by a single parametric distribution. First-order finite mixture models have been widely used in the physical, chemical, biological, and social sciences for over 100 years. Using maximum likelihood estimation, we demonstrate how a first-order finite mixture model can represent the large variability in data collected by the US Environmental Protection Agency for the concentration of Radon 222 in drinking water supplied from ground water, even when 28 percent of the data fall at or below the Minimum Reporting Level. Extending the use of maximum likelihood, we also illustrate how a second-order finite mixture model can separate and represent both the variability and the uncertainty in the data set.

1.0 Introduction to Finite Mixture Models

For over 100 years, finite mixture models have been used in the physical, chemical, biological, and social sciences (Everitt & Hand, 1981; McLachlan & Basford, 1988; Lindsay, 1995) as a way to analyze and model "... measurements from experimental units known to belong to one of a set of classes, but whose individual class-memberships are unavailable" (Titterington et al., 1985, page xi). Even though finite mixture models have a long and honorable history in many academic and practical fields, they have found little use in exposure or risk assessment to date.

Paraphrasing the formal definition from Titterington et al. (page 1), a finite mixture model for a random variable (or a random vector), X, takes values in a sample space and can be represented by a probability density function of the form:

$$p(x) \quad = \quad \pi_1 f_1(x) + \cdots + \pi_k f_k(x) \qquad \text{for x in the sample space} \qquad \text{Eqn 1}$$

with

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

and

$$\pi_1 + \cdots + \pi_k \quad = \quad 1$$

All $f_k(\bullet)$ are probability density functions (either continuous or discrete). Here, the π_k are called the mixing weights and the $f_k(x)$ are called the components (or the component distributions) of the finite mixture. In most situations, the $f_k(\bullet)$ have specified parametric forms:

$$p(x) \quad = \quad \pi_1 f_1(x \mid \theta_1) + \cdots + \pi_k f_k(x \mid \theta_k) \qquad \text{Eqn 2}$$

where θ_j denotes the vector of parameters in density $f_j(\bullet)$. Titterington et al. (1985) define finite mixture models with full mathematical rigor.

In practice, an analyst uses computer-aided visualization (D'Agostino & Stephens, 1986; Burmaster & Thompson, 1999) to investigate the suitability of different finite mixture models.

An analyst may investigate various mixture models under different circumstances:
(i) when the data are known to include measurements from two or more different subpopulations, e.g., males and females, indoor and outdoor air, or lakes, estuaries, and oceans; or (ii) when no single parametric distribution provides an adequate fit to the measurements. In some situations, the analyst may know the proper number of components to include in the mixture, but in many situations, the analyst does not know the proper number of components to include. In many cases, an analyst uses the method of maximum likelihood estimation (MLE) to fit different models to a data set and

then uses computer-aided visualization as a way to see if the mixture model fits the data better than a single parametric distribution. Since the goodness of fit of a finite mixture model never decreases as an analyst adds another component to a mixture, an analyst must always consider whether the increase in the goodness of fit afforded by the addition of another component is worth the effort. In the spirit of Occam's Razor, the analyst must trade-off the quality of the fit against the number of components (or the total number of parameters). Statisticians and econometricians sometimes penalize the overall complexity of a model by adding a term to the MLE objective function to favor parsimony (see discussion and references on page 9 et seq. in O'Ruanaidh & Fitzgerald, 1996).

2.0 The NIRS Data

To illustrate the use of first- and second-order finite mixture models [EndNote 1], we re-analyze a data set collected by the US Environmental Protection Agency (US EPA). As explained by Longtin (1988; 1990) and Barry and Brattin (1998), the National Inorganic and Radionuclides Survey (NIRS) selected a stratified random sample from the 47,700 community drinking water systems supplied from ground water as inventoried by the US Environmental Protection Agency (US EPA) in 1984. Of the 1,000 water service districts initially included in the NIRS survey, 990 responded (8 of which were later excluded for data quality reasons). Six water samples per system were analyzed to compute the mean concentration of Radon 222 in picocuries per liter (denoted [Rn222] in pCi/L) in the finished water in each distribution system. The resulting 982 data points in the final NIRS data set comprise a representative sample of water supply systems (but not a representative sample of human population) stratified by geographical region and other factors.

The NIRS data exhibit a large degree of variability. [EndNote 2] The measurements for [Rn222] were originally reported in the range from less than 100 pCi/L to greater than 25,000 pCi/L. (Longtin, 1990). Subsequently, the US EPA concluded that a single Minimum Reporting Level (MRL = 100 pCi/L) applies to all the data in the Survey (Barry, 1998). Since the US EPA decided that all of the concentrations originally reported at or below 100 piC/L are unreliable, we followed the Agency's lead by treating all values originally reported at or below the subsequent MRL as censored values, i.e., as nondetects with an effective detection limit of 100 pCi/L. Of the 982 data points in the data set, 275 (or 28 percent) of the data points were originally reported at or below the subsequent MRL. [EndNote 3]

After disaggregating the NIRS data, Barry and Brattin (1998) used probability plots, maximum likelihood estimation (MLE), and other statistical techniques to analyze and model various patterns in these left-censored data for 5 size categories and 8 regions (Table 2 therein). After fitting Weibull, Gamma, Pearson V, Pearson VI, LogWeibull, LogGamma, and other distributions to the data for each size category and each region, Barry and Brattin conclude (1998, page 598) that the distribution of [Rn222] in community ground water systems "is reasonably well described" by first-order LogNormal distributions (whose parameters depend on population served and/or geographic region). They continue, "However, despite the good overall [LogNormal] fit, some NIRS data ... exhibit systematic deviations from lognormality, displaying longer (heavier) than lognormal tails. Although these deviations generally occur beyond the 93rd - 95th percentiles, for situations in which high-end estimation is important, a better fitting long-tailed distribution or compound distribution might be better suited to these strata." Here, one may interpret "compound distribution" as a finite mixture model.

Because Barry and Brattin considered and dismissed several families of two-parameter distributions in their analyses, including the Weibull, Gamma, Pearson V, Pearson VI, LogWeibull, and LogGamma distributions, we explore a number of finite mixture models as a way to obtain a better parametric fit to of the NIRS data -- including the upper tail of the data. Because Barry and Brattin emphasize that each subpopulation or stratum has characteristics which differ from the others in size, region, and geology, we explore whether a two-component mixture model can fit the pooled data better than a single parametric distribution.

3.0 Two Models

Over the last decade, many people have observed that concentrations of many chemical species (both natural and anthropogenic) in different environmental media often follow LogNormal distributions (e.g., Ott, 1990; 1995). We too have found that a LogNormal distribution often provides a satisfactory fit to measurements of environmental concentrations, especially when the data come from a single environmental medium or a small region. Thus, following Barry and Brattin (1998) and using maximum likelihood methods, we first fit a single LogNormal distribution, called Model 1, to the 982 pooled NIRS data points without regard to size category, region, or other characteristics of the water supply:

Model 1: Eqn 3

$$X \sim \text{LogNormal}[\, x \mid \mu, \sigma \,]$$

For Model 1, we use the probability density function (PDF) and the cumulative distribution function (CDF) for the LogNormal distribution as follows (Aitchison & Brown, 1957; Crow & Shimizu, 1988):

$$\text{PDF}[\text{ LogNormal}[\, x \mid \mu, \sigma \,]] = \frac{1}{\sigma\; x \sqrt{2\,\pi}} \exp[\, -\frac{1}{2} \frac{(\ln[\, x \,] - \mu)^2}{\sigma^2} \,] \quad \text{Eqn 4}$$

$$\text{CDF}[\text{ LogNormal}[\, x \mid \mu, \sigma \,]] = \frac{1}{2} [\, 1 + \text{Erf}[\, \frac{\ln[\, x \,] - \mu}{\sigma \sqrt{2}} \,] \,] \quad \text{Eqn 5}$$

Here, exp[•] denotes the exponential function, ln[•] denotes the Napierian (or natural) logarithm function, and Erf[•] denotes the error function (Abramowitz & Stegun, 1964). Alternatively, we could have achieved the same numerical results by first taking the natural logarithms of the data and then using the PDF and CDF for the Normal distribution in all calculations.

When working with medium to large data sets (say, $n > 200$), we too have found that a LogNormal distribution sometimes provides a less than satisfactory fit to the data points above the 90th percentile of the empirical distribution, especially for pooled data including measurements from several known or suspected subpopulations. Thus, in the spirit of finite mixture models (and after extensive simulation modeling to gain intuition about alternative models with different component density distributions), we fit the following mixture model, called Model 2, to the 982 pooled NIRS data points, again without regard to size category, region, or other characteristics of the water supply:

Model 2: Eqn 6

$$X \sim \pi \;\text{LogNormal}[\, x \mid \mu_1, \sigma_1 \,] + (1 - \pi)\, \text{LogNormal}[\, x \mid \mu_2, \sigma_2 \,]$$

Model 2 has 5 parameters: π, μ_1, σ_1, μ_2, and σ_2. Model 2 comes from a family of finite mixture distributions. As appropriate, the models in this mixture family can be constrained (e.g., by forcing $\sigma_1 = \sigma_2$ in Model 2), generalized (e.g., by using a Gamma distribution or other parametric distribution as the second component), and/or extended (e.g., by adding additional component(s)).

4.0 Introduction to Maximum Likelihood Estimation

Sir Ronald A. Fisher developed the method of maximum likelihood estimation (MLE) as a powerful, general purpose method for fitting parametric distributions to data. The general idea is to choose the parameter values that will, when substituted into the model, make the probability of the sample larger than its probability using any other parameter values (paraphrased from Keeping, 1995). Fisher later generalized the idea to develop joint confidence regions for the parameters, an idea that was further generalized to the profile likelihood method for marginal distributions for parameters.

This section only summarizes some of the most basic ideas of MLE theory to convey the nature of our approach and to present the notation used in this paper. For a thorough and rigorous discussion of MLE, see Mood et al (1974), Appendix E in Lawless (1982), or Edwards (1992).

When using MLE, the analyst first computes the likelihood function and then the loglikelihood function, J, for a particular parametric model as a function of all the parameters in the model, θ, with the given data taken as fixed values. Using nonlinear optimization, the analyst then finds the maximum value of J[θ] and the corresponding values for all the parameters in the vector $\theta = \hat{\theta}$. For large sample sizes (and under certain regularity conditions met by the models considered here), the uncertainty in the parameters θ follows a multivariate Normal distribution (see, e.g., Evans et al, 1993) with (i) the mean vector equal to the optimal values of the parameters, $\hat{\theta}$, and (ii) the variance covariance matrix equal to (Appendix 1, Cox & Snell, 1989):

$$\mathrm{VarCov}[\theta] = [- \nabla_\theta[\nabla_\theta[J]]_{\hat{\theta}}]^{-1} \qquad \text{Eqn 7}$$

Here, ∇_θ is the gradient operator with respect to θ. Using matrix algebra, one may compute the correlation matrix for the parameters (denoted Corr[θ]) from the VarCov[θ] matrix (see, for example, Anderson, 1958).

Because the US EPA has concluded that <u>all</u> of the concentrations originally reported at or below the subsequent MRL (100 piC/L) are unreliable, we followed the Agency's lead by treating all values originally reported at or below the MRL as censored values, i.e., as nondetects with an effective detection limit of 100 pCi/L. With this treatment, the loglikelihood function for Model 1 is this function of its two parameters (see page 15 in Tanner, 1996; Burmaster & Thompson, 1998):

$$J_1[\mu, \sigma] \qquad \text{Eqn 8}$$

$$= \sum_{x_i \leq MRL} \ln[\text{ CDF[LogNormal[MRL } | \mu, \sigma]]]$$

$$+ \sum_{x_i > MRL} \ln[\text{ PDF[LogNormal[} x_i | \mu, \sigma]]]$$

Similarly, the loglikelihood function for Model 2 is this function of its 5 parameters:

$$J_2[\pi, \mu_1, \sigma_1, \mu_2, \sigma_2] \qquad \text{Eqn 9}$$

$$= \sum_{x_i \leq MRL} \ln\Big[\pi \text{ CDF[LogNormal[MRL } | \mu_1, \sigma_1]] + (1 - \pi) \text{ CDF[LogNormal[MRL } | \mu_2, \sigma_2]] \Big]$$

$$+ \sum_{x_i > MRL} \ln\Big[\pi \text{ PDF[LogNormal[} x_i | \mu_1, \sigma_1]] + (1 - \pi) \text{ PDF[LogNormal[} x_i | \mu_2, \sigma_2]] \Big]$$

5.0 Results

We used Mathematica® Version 3.0.1 for all computations and graphs here (Wolfram, 1996; Wickham-Jones, 1994). In the optimizations, we set Mathematica's AccuracyGoal and PrecisionGoal each to 14 decimal digits.

First, adapting techniques from exploratory data analysis (Tukey, 1977; Fowlkes, 1979) and from graphical analysis (D'Agostino & Stephens, 1986; Cleveland, 1993; Cleveland, 1994), we used standard and animated probability plots (Burmaster & Thompson, 1999) to see how well different finite mixture models fit the 982 pooled data points. During the exploration, we considered a variety of different parametric distributions as models for the data, but we found that LogNormal distribution(s) gave the most appealing visual fits. From a standard LogNormal probability plot, we estimated the two parameters of Model 1, and from a series of animated LogNormal probability plots, we estimated the 5 parameters of Model 2. In this step, we relied on visual appeal, not algebraic diagnostics, to compare different models.

Figure 1 shows an initial LogNormal probability plot for the 982 pooled data points (D'Agostino & Stephens, 1986; Burmaster & Hull, 1997). In it, the smaller dots depict the 275 censored values with [Rn222] ≤ MRL, and the larger dots depict the 707 noncensored values with [Rn222] > MRL. As a part of exploratory data analysis, we used ordinary least-squares regression to fit a straight line to the 707 data points with [Rn222] above the MRL. This straight line, shown in Figure 1, has an adjusted R-squared value equal to 0.983; it has an intercept equal to 5.2372 (an initial estimate of μ in Eqn 3 for Model 1) and a slope equal to 1.48968 (an initial estimate of σ in Eqn 3 for Model 1). [EndNote 4] In Figure 1, the data points with z-score > 1.30 (i.e., the points above the 90^{th} percentile of the data) depart from the straight line in a small but systematic way. For the censored pooled NIRS data, these results agree with Barry and Brattin (1998).

Figures 2A and 2B show the 707 noncensored data points plotted as an empirical cumulative distribution function (ECDF).

- Figure 2A shows the single best-fit cumulative distribution function for Model 1 fit to the 982 pooled data points by MLE. Figure 2A shows the variability in Model 1 as a single CDF fit to the pooled data. From the optimization of Eqn 8, $J[\hat{\mu}, \hat{\sigma}] = -5{,}684.02$. Table I shows the optimal values of $\hat{\mu}$ and $\hat{\sigma}$ as a column vector, $\hat{\theta}$. These refined MLE estimates of μ and σ differ slightly from the ones obtained from the regression on the LogNormal probability plot shown in Figure 1. These results also agree with Barry and Brattin (1998).

- In contrast, Figure 2B shows 200 realizations of the variability and the uncertainty in the CDF for Model 1 fit to the pooled data by MLE. In the figure, each CDF represents the variability in the data, and the multiplicity of the 200 CDFs overlapping in a tight band represents the uncertainty in the fitted second-order distribution. Table I shows the numerical values in the $\mathrm{VarCov}[\hat{\theta}]$ matrix, along with the $\mathrm{Corr}[\hat{\theta}]$ matrix derived from it. [EndNote 5] In Figure 2B, the uncertainty band does not fully envelope the data points in the ECDF, thereby exhibiting a systematic deviation from Model 1.

From Figures 1, 2A and 2B, we see that Model 1 provides a first approximation to the data throughout the noncensored range, even though we see that neither the first-order version nor the second-order version of Model 1 fits the data perfectly in any figure. Remembering George Box's dictum (Box, 1979), "All models are wrong, but some are useful," we accept Model 1 as a first-approximation for predicting [Rn222] in ground water drinking water supplies. Even though Model 1 is the best-fitting two-parameter distribution that we could find, its systematic deviations from the pooled data motivated and informed our search for an even better fit.

Figures 3A and 3B also show the 707 noncensored data points plotted as an empirical cumulative distribution function (ECDF).

- Figure 3A shows the single best-fit CDF for Model 2 fit to the pooled data by MLE. Figure 3A shows the variability in Model 2 as fit to the NIRS data as a single CDF. From the optimization of Eqn 9, J[$\hat{\pi}$, $\hat{\mu}_1$, $\hat{\sigma}_1$, $\hat{\mu}_2$, $\hat{\sigma}_2$] = -5,662.93. Table II shows the optimal values of $\hat{\pi}$, $\hat{\mu}_1$, $\hat{\sigma}_1$, $\hat{\mu}_2$, and $\hat{\sigma}_2$ as a column vector, $\hat{\theta}$. These optimal values differ from our initial estimates from the animated probability plots by only a few percent.

- In contrast, Figure 3B shows 200 realizations of the variability and the uncertainty in the cumulative distribution function for Model 2 fit to the pooled data by MLE. In the figure, each CDF represents the variability in the data, and the multiplicity of the 200 CDFs overlapping in a tight band represents the uncertainty in the fitted second-order distribution. Table II shows the numerical values in the VarCov[$\hat{\theta}$] matrix, along with the Corr[$\hat{\theta}$] matrix derived from it. [EndNote 6] In Figure 3B, the uncertainty band fully envelopes the data points in the ECDF.

Comparing Figures 3A and 3B to Figures 2A and 2B, we see that Model 2 provides a better fit to the data throughout the noncensored range than does Model 1. From the values for $\hat{\pi}$ and $\sigma[\hat{\pi}]$ in Table II, we infer that $\pi \neq 1$, a significant result that implies that both component distributions are needed. We did not use a likelihood ratio test Model 2 against Model 1 due to boundary problems discussed in the literature (Roeder, 1994; and references therein).

Figure 4A shows the CDF for Model 2 as a black line and the CDFs for its two (pure) components as gray lines. Here, the CDF for Model 2 is a weighted average of the CDFs for the two components, i.e., 91.4 percent of component 1 and 8.6 percent of component 2. In Figure 4A, one can clearly see how the upper tail of Model 2 differs from the upper tail of its first component. Figure 4B shows the PDF for Model 2 as a black line and the PDFs for its two (pure) components as gray lines. In this figure, one cannot discern a difference between the upper tail of Model 2 and the upper tail of its first component.

Figure 5A shows the CDF for Model 1 as a dashed line and the CDF for Model 2 as a solid line, while Figure 5B shows the PDF for Model 1 as a dashed line and the PDF for Model 2 as a solid line. These two graphs highlight the subtle differences between Model 1 and Model 2. The CDFs in Figure 5A show the small differences between Model 1 and Model 2 more clearly than do the PDFs in Figure 5B. By looking at Figure 5A, one can see that Model 2 has a slightly longer (or heavier) upper tail than does Model 1, a difference that one might miss or dismiss if looking just at Figure 5B.

At this point, we knew that Model 2 fits the pooled data better than does Model 1 over the full range of the measurements. However, we did not have a theory to supply the proper value for k, the number of components in Eqns 1 and 2. While we knew that adding a third component to Model 2 would improve the fit to the pooled data, we did not know how to estimate k. Haughton (1997) and Everitt and Hand (1981) discuss the difficulty of estimating k. In written comments on an earlier paper, William A. Huber (1998) showed us a graphical way to approach this thorny issue for mixtures of Normal or LogNormal distributions. Basically, Huber analyzed the log-transformed pooled NIRS data using a mixture detection plot, a technique developed by Roeder (1994). Huber's graphical results, similar to Figure 1h in Roeder (1994), confirm the presence of two Normal components, one stronger than the other, in the log-transformed data. His results also suggest the presence of one or more other components of negligible contribution to the mixture.

6.0 Using Model 2 in a Simulation

An analyst may generate random variates from Model 2 as follows: For the outer loop (for uncertainty in the fitted distribution), draw realizations from a 5-dimensional Normal distribution with mean $\hat{\theta}$ and variance-covariance matrix VarCov[$\hat{\theta}$] matrix shown in Table II. Remembering to keep $0 \le \pi \le 1$, $\sigma_1 > 0$, and $\sigma_2 > 0$, discard any realizations that contain infeasible values of π, σ_1, or σ_2. For the inner loop (for the variability in the fitted distribution), obtain a uniform random variate between 0 and 1. If it is less than π, draw the value randomly from the first LogNormal component of Model 2 with parameters μ_1 and σ_1; if it is greater than or equal to π, instead draw the value randomly from the second LogNormal component of Model 2 with parameters μ_2 and σ_2.

7.0 Conclusions and Discussion

We conclude that Model 1 provides an approximation to the pooled NIRS data and that Model 2 provides a much better fit to the pooled data. In the context of a particular analysis, the exposure assessor must then decide whether the better fit provided by Model 2 makes it worth the extra effort relative to the first approximation provided by Model 1.

An adage from R.W. Hamming in the early days of computer science states "The purpose of computation is insight." This re-analysis of the pooled NIRS data suggests a new approach to understanding the concentration of Radon-222 in drinking water derived from ground water. The results here suggest that there may be geophysical or geochemical explanations for the two component distributions made visible in Model 2. If one looked, would one find that water samples from systems with lower [Rn222] come from overburden aquifers? Do water samples from systems with higher [Rn222] come from bedrock aquifers in fractured granite? Are ~91 percent of the water supplies aerated or otherwise treated to reduce [Rn222]? Are there other geophysical, geochemical, or treatment processes that create the statistical patterns seen in the two components in Model 2? We do not know the answers to these questions, but we think that the questions suggest new ways to view the pooled NIRS data.

Even though finite mixture models have a long and honorable history in many academic and practical fields, they have found little use in exposure or risk assessment to date. When fit by maximum likelihood estimation, finite mixture models have many strengths:

- They often achieve high-fidelity, parsimonious fits to data when no standard parametric distribution provides an adequate fit. Considering the results here, Model 2 provides an excellent fit to the NIRS data in compact and efficient form. This strength flows primarily from the finite mixture model itself.

- When developed as second-order distributions, they separate and quantify the variability and the uncertainty present in the data. This strength and the remaining ones flow primarily from the MLE technique.

- They work with combinations of distributions from many parametric families; each component may come from a different parametric family.

- They work with censored and/or binned data, e.g., measurements reported at or below a Minimum Reporting Level, even when each datum has a different MRL.

- They produce joint confidence regions with the proper correlations among the parameters being estimated.

- As the number of data points grows large, the uncertainty in the finite mixture model converges asymptotically to Normal theory and produces joint confidence regions as ellipses.

- They produce results that are easily visualized and used in "two-dimensional" Monte Carlo simulations.

The MLE method also has practical limitations when applied to finite mixture models because the fitting procedures require much more effort and care than does fitting by the method of moments or by regression. Even skilled analysts can make mistakes in the optimizations. There are several common pitfalls: (i) The mixture model may not be identifiable. (See, for example, Chapter 7 in Lancaster, 1990). (ii) The optimization routine in the computer software may terminate before reaching the true optimum. In this case, the analyst will then mistakenly report suboptimal results which may deviate significantly from the correct ones. (iii) Without excellent initial estimates for the parameters in Eqn 2, the analyst may find and erroneously report one or more local maxima of the loglikelihood function instead of the global maximum. (Actually, there are two symmetric and identical global maxima of the loglikelihood function for a two-component mixture.) (iv) The optimization routines may converge to a boundary of the feasible space ($\pi = 0$, $\pi = 1$, $\sigma_1 = 0$, or $\sigma_2 = 0$). Along these boundaries, the variance-covariance matrix may not exist or other problems may arise. (v) The MLE problem may be ill-conditioned, a situation that may lead to numerical instabilities and related problems. Simply because an analyst can find a software package that will fit first-order mixture models to data sets [EndNote 7] does not eliminate the need for careful work and for review by colleagues.

Finally, even though finite mixture models have a long and honorable history in many academic and practical fields, they have found little use in exposure or risk assessment to date. And even though the MLE method has a long and honorable history as a powerful way to fit finite mixture models to data, statisticians have developed even more powerful methods in the last 25 years. For many practical data sets, statisticians now routinely use the Expectation-Maximization (EM) algorithm -- often in combination with the method of partial likelihood -- to fit finite mixture models to data. Lawless (1982), Titterington et al (1985), Lancaster (1990), and Schafer (1997) discuss these more powerful methods in depth.

EndNotes

1. For a full discussion of second-order random variables, see Burmaster and Wilson (1996) and the references therein. For a full discussion on how to fit first- and second-order LogNormal distributions to data using MLE, see Burmaster and Thompson (1998) and the references therein.

2. In this paper, variability represents true heterogeneity in the (statistical) sample or population which cannot be reduced through further measurement or study (although such heterogeneity may be disaggregated into different components associated with different subgroups in the population). Variability is a fundamental property of Nature and of the population. Variability in a population is best analyzed and modeled in terms of a probability distribution, often a first-order parametric distribution with constant parameters. As distinguished from variability, uncertainty represents ignorance -- or lack of perfect knowledge -- about a phenomenon for a population as a whole which may sometimes be reduced through further measurement or study. Uncertainty is a property of the analyst performing the assessment. Uncertainty about the variability in a population can be analyzed and modeled in terms of a probability distribution, usually a second-order parametric distribution with nonconstant (distributional) parameters.

3. One may download a list of the 982 data points in the NIRS Survey as a part of a PDF file for this article. Go to the publications section at this URL: www.Alceon.com

4. Throughout, we report the results by showing digits beyond the ones usually considered the "significant digits" for two reasons. First, for values that have explicit standard deviations or standard errors associated with them, we present the extra digits because the standard deviations or standard errors directly show the tolerances properly associated with the results. Second, for the values in the VarCov[•] and Corr[•] matrices in a later section, we present the extra digits to prevent round-off errors in the matrix algebra from making the matrices inconsistent.

5. In Table I for Model 1, the Corr matrix differs from the Identity matrix usually found when working with a single LogNormal distribution and all noncensored values because the censoring induces nonzero cross terms between μ and σ.

6. In Table II for Model 2, there is no reason to expect that the Corr matrix should equal the Identity matrix.

7. For information on commercial software packages that use MLE to estimate the parameters of finite mixture models, see, for example, Haughton (1997) and Mcdonald (1998). In addition, colleagues tell us that these commercial programs will fit finite mixture models to data: S-Plus® (www.StatSci.com), Systat® (www.SPSS.com), and SAS® (www.SAS.com). Other commercial software programs may also fit mixture models to data. We do not know if any of these programs work with data sets containing censored values.

Acknowledgments

Many people generously gave us help during this research. We thank each and all. We thank Louis Anthony "Tony" Cox, Jr. and two anonymous reviewers for excellent suggestions to improve the presentation. William A. Huber and Edmund A.C. Crouch independently re-did all the calculations and identified suboptimal results for Model 2 in an earlier manuscript. William A. Huber calculated and graphed the mixture detection plot (Roeder, 1994). William A. Huber, Robert T. Clemen, and Scott D. Ferson read an earlier manuscript and gave us many comments and suggestions for improvements. Timothy M. Barry provided the data and Ronald J. Bosch gave us help and encouragement throughout this research.

Alceon Corporation funded this research.

Trademarks

Mathematica® is a registered trademark of Wolfram Research, Inc: in Champaign, IL, www.Wolfram.com

Alceon® is a registered trademark of Alceon Corporation in Cambridge, MA, www.Alceon.com

References

Aitchison & Brown, 1957
Aitchison, J. and J.A.C. Brown, 1957, The Lognormal Distribution, Cambridge University Press, Cambridge, UK

Abramowitz & Stegun, 1964
Abramowitz, M. and I.A. Stegun, Eds, 1964, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, Applied Mathematics Series Number 55, Issued June 1964, Tenth Printing with corrections in December 1972, US Government Printing Office, Washington, DC

Anderson, 1958
Anderson, T.W., 1958, An Introduction to Multivariate Statistical Analysis, Wiley, New York, NY

Barry, 1998
Barry, T.M., 1998, Email message dated 28 January 1998 sent to D.E. Burmaster

Barry & Brattin, 1998
Barry, T.M. and W.J. Brattin, 1998, Distribution of Radon-222 in Community Groundwater Systems: Analysis of Type I Left-Censored Data with Single Censoring Point, Human and Ecological Risk Assessment, Volume 4, Number 2, pp 579 - 603, April 1998

Box, 1979
Box, G.E.P., 1979, Robustness is the Strategy of Scientific Model Building, in Robustness in Statistics, R.L Launer and G.N. Wilkinson, eds., Academic Press, New York, NY

Burmaster & Thompson, 1999
Burmaster, D.E. and K.M. Thompson, 1999, Using Animated Probability Plots to Explore the Suitability of Mixture Models with Two Component Distributions, Risk Analysis, revised 22 February 1999, available as a PDF file from http://www.Alceon.com

Burmaster & Thompson, 1998
Burmaster, D.E. and K.M. Thompson, 1998, Fitting Second-Order Parametric Distributions to Data Using Maximum Likelihood Estimation, Human and Ecological Risk Assessment, Volume 4, Number 2, pp 319 - 339, April 1998

Burmaster & Hull, 1997
Burmaster, D.E. and D.A. Hull, 1997, Using LogNormal Distributions and LogNormal Probability Plots in Probabilistic Risk Assessments, Human and Ecological Risk Assessment, Volume 3, Number 2, pp 235 - 255

Burmaster & Wilson, 1996
Burmaster, D.E. and A.M. Wilson, 1996, An Introduction to Second-Order Random Variables in Human Health Risk Assessment, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 892 - 919

Cleveland, 1993
Cleveland, W.S., 1993, Visualizing Data, AT&T Bell Laboratories, Hobart Press, Summit, NJ

Cleveland, 1994
Cleveland, W.S., 1994, The Elements of Graphing Data, AT&T Bell Laboratories, Hobart Press, Summit, NJ

Cox & Snell, 1989
Cox, D.R. and E.J. Snell, 1989, Analysis of Binary Data, Second Edition, Chapman & Hall, London, UK

Crow & Shimizu, 1988
Crow, E.L. and K. Shimizu, Eds., 1988, Lognormal Distributions, Theory and Applications, Marcel Dekker, New York, NY

D'Agostino & Stephens, 1986
D'Agostino, R.B. and M.A. Stephens, 1986, Goodness-of-Fit Techniques, Marcel Dekker, New York, NY

Edwards, 1992
Edwards, A.W.F., 1992, Likelihood, John Hopkins University Press, Baltimore, MD

Evans et al, 1993
Evans, M., N. Hastings, and B. Peacock, 1993, Statistical Distributions, Second Edition, John Wiley & Sons, New York, NY

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

Fowlkes, 1979
Fowlkes, E.B., 1979, Some Methods for Studying the Mixture of Two Normal (Lognormal) Distributions, Journal of the American Statistical Association, Volume 74, Number 367, pp 561 - 575

Haughton, 1997
Haughton, D., 1997, Packages for Estimating Finite Mixtures: A Review, The American Statistician, Volume 51, Number 2, pp 194 - 205, May 1997

Huber, 1998
Huber, W.A., 1998, Technical Comments dated 8 September 1998 sent to D.E. Burmaster

Keeping, 1995
Keeping, E.S., 1995, Introduction to Statistical Inference, Dover, New York, NY

Lancaster, 1990
Lancaster, T., 1990 The Econometric Analysis of Transition Data, Cambridge University Press, Cambridge, UK

Lawless, 1982
Lawless, J.F., 1982, Statistical Models and Methods for Lifetime Data, Wiley & Sons, New York, NY

Lindsay, 1995
Lindsay, B.G., 1995, Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5, IMS, Hayward, CA

Longtin, 1990
Longtin, J.P., 1990, Occurrence of Radionuclides in Drinking Water, In Radon, Radium, and Uranium in Drinking Water edited by C.R. Cothern and P.A. Rebers, Lewis Pubishers, Ann Arbor, MI

Longtin, 1988
Longtin, J.P., 1988, Occurrence of Radon, Radium, and Uranium in Groundwater, Journal of the American Water Works Association, pp 84 - 93, July 1988

Mcdonald, 1998
Mcdonald, P., 1998, Fitting Mixture Distributions, Software and Applications, Hamilton, Ontario, Canada, http://icarus.math.mcmaster.ca/peter/mix/mix.html

McLachlan & Basford, 1988
McLachlan, G.J. and K.E. Basford, 1988, Mixture Models: Inference and Applications to Clustering, New York, NY

Mood et al, 1974
Mood, A.M., F.A. Graybill, and D.C. Boes, 1974, Introduction to the Theory of Statistics, Third Edition, McGraw Hill, New York, NY

O'Ruanaidh & Fitzgerald, 1996
O'Ruanaidh, J.K. and W.J. Fitzgerald, 1996, Numerical Bayesian Methods Applied to Signal Processing, Springer Verlag, New York, NY

Ott, 1995
Ott, W.R., 1995, Environmental Statistics and Data Analysis, Lewis Publishers, Boca Raton, FL

Ott, 1990
Ott, W.R., 1990, A Physical Explanation of the Lognormality of Pollutant Concentrations, Journal of the Air and Waste Management Association, Volume 40, pp 1378 et seq.

Roeder, 1994
Roeder, K., 1994, A Graphical Technique for Determining the Number of Components in a Mixture of Normals, Journal of the American Statistical Association, Volume 89, Number 426, pp 487 - 495

Schafer, 1997
Schafer, J.L., 1997, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, UK

Tanner, 1996
Tanner, M.A., 1996, Tools for Statistical Inference, Springer-Verlag, New York, NY

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

Tukey, 1977
Tukey, J.W., 1977, Exploratory Data Analysis, Addison-Wesley, Reading, MA

Wickham-Jones, 1994
Wickham-Jones, T., 1994, Mathematica Graphics, Techniques & Applications, Springer-Verlag, Telos, Santa Clara, CA

Wolfram, 1996
Wolfram, S., 1996, The Mathematica® Book, Third Edition, Wolfram Media, Champaign, IL

$$\hat{\theta} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} = \begin{bmatrix} 5.30930 \\ 1.40572 \end{bmatrix} \quad ; \quad \sigma[\hat{\theta}] = \begin{bmatrix} 0.047565 \\ 0.039646 \end{bmatrix}$$

$$\text{VarCov}[\hat{\theta}] = \begin{bmatrix} 0.002262 & -0.000369 \\ -0.000369 & 0.001572 \end{bmatrix}$$

$$\text{Corr}[\hat{\theta}] = \begin{bmatrix} 1. & -0.195662 \\ -0.195662 & 1. \end{bmatrix}$$

Table I: Parameters for Model 1 Fit by MLE

$$\hat{\theta} = \begin{bmatrix} \hat{\pi} \\ \hat{\mu}_1 \\ \hat{\sigma}_1 \\ \hat{\mu}_2 \\ \hat{\sigma}_2 \end{bmatrix} = \begin{bmatrix} 0.91423 \\ 5.13085 \\ 1.06025 \\ 7.92975 \\ 0.97975 \end{bmatrix} \quad ; \quad \sigma[\hat{\theta}] = \begin{bmatrix} 0.043907 \\ 0.078083 \\ 0.076203 \\ 0.584203 \\ 0.263189 \end{bmatrix}$$

$$\text{VarCov}[\hat{\theta}] = \begin{bmatrix} 0.001928 & 0.002932 & 0.002831 & 0.024429 & -0.010156 \\ 0.002932 & 0.006097 & 0.004285 & 0.038484 & -0.015812 \\ 0.002831 & 0.004285 & 0.005807 & 0.035712 & -0.013950 \\ 0.024429 & 0.038484 & 0.035712 & 0.341293 & -0.141251 \\ -0.010156 & -0.015812 & -0.013950 & -0.141251 & 0.069268 \end{bmatrix}$$

$$\text{Corr}[\hat{\theta}] = \begin{bmatrix} 1. & 0.855141 & 0.846049 & 0.952370 & -0.878898 \\ 0.855141 & 1. & 0.720167 & 0.843642 & -0.769407 \\ 0.846049 & 0.720167 & 1. & 0.802205 & -0.695549 \\ 0.952370 & 0.843642 & 0.802205 & 1. & -0.918672 \\ -0.878898 & -0.769407 & -0.695549 & -0.918672 & 1. \end{bmatrix}$$

Table II: Parameters for Model 2 Fit by MLE

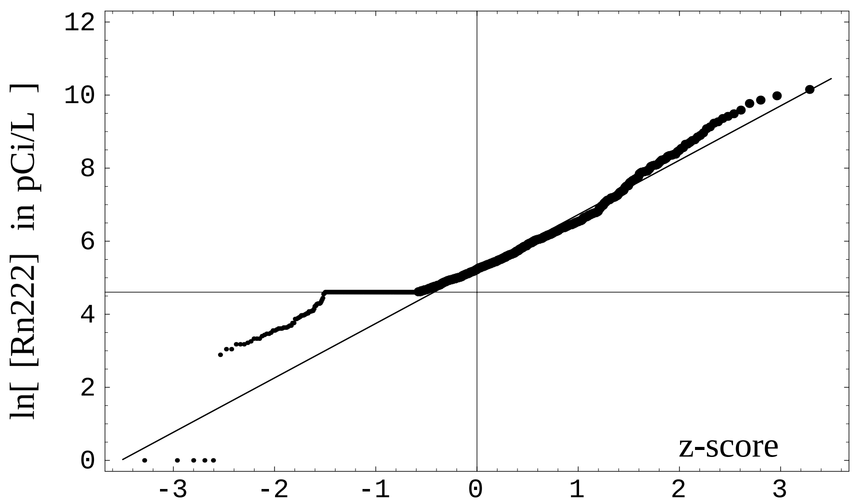

Figure 1: LogNormal probability plot showing the 707 values > 100 pCi/L as larger dots and the 275 values ≤ 100 pCi/L as smaller dots. The US EPA has concluded that all the data originally reported at or below 100 pCi/l are unreliable. See the text for discussion.

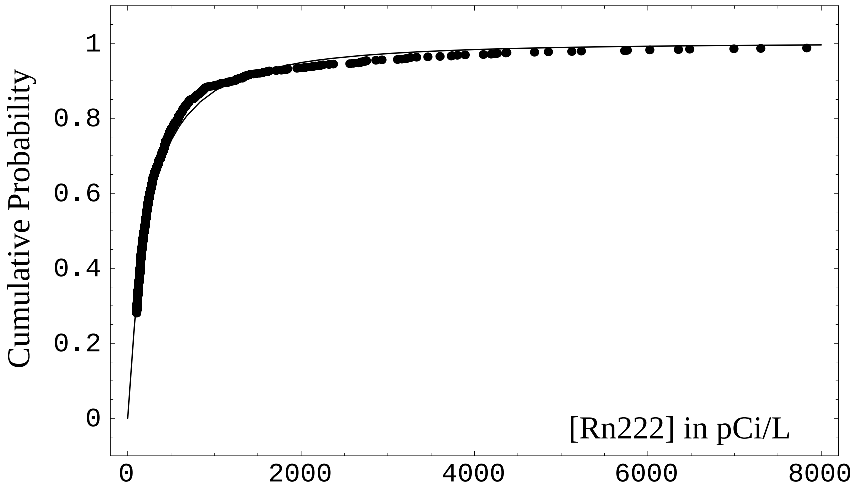

Figure 2A: CDF for Model 1 fit by MLE,
with dots showing the 707 values > 100 pCi/L

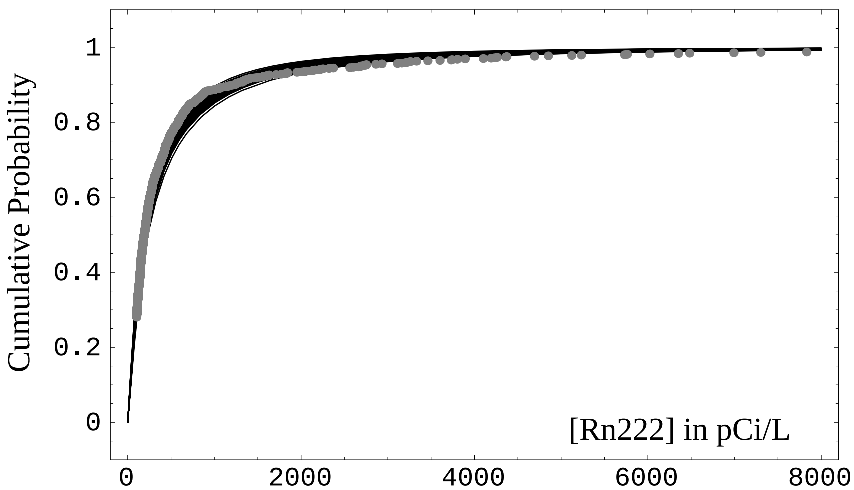

Figure 2B: Multiple CDFs for the uncertainty in Model 1,
with dots showing the 707 values > 100 pCi/L

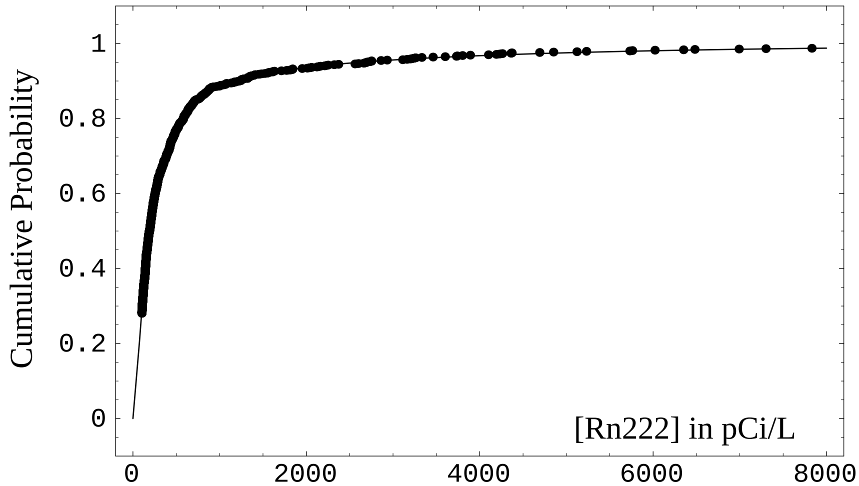

Figure 3A: CDF for Model 2 fit by MLE,
with dots showing the 707 values > 100 pCi/L

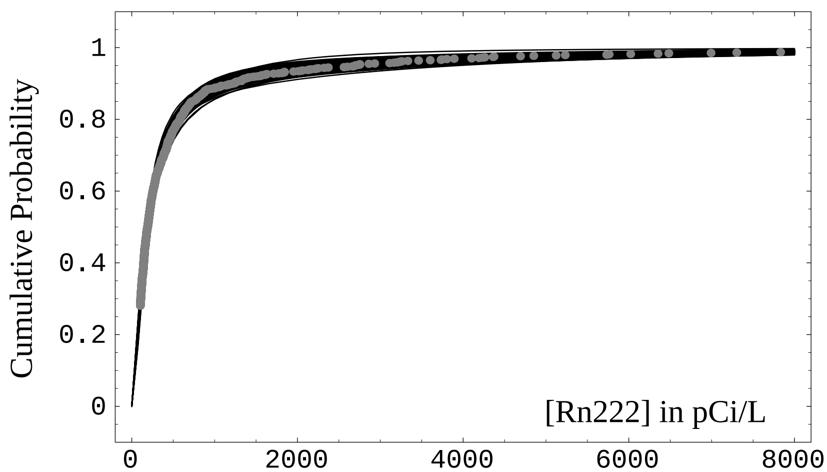

Figure 3B: Multiple CDFs for the uncertainty in Model 2,
with dots showing the 707 values > 100 pCi/L

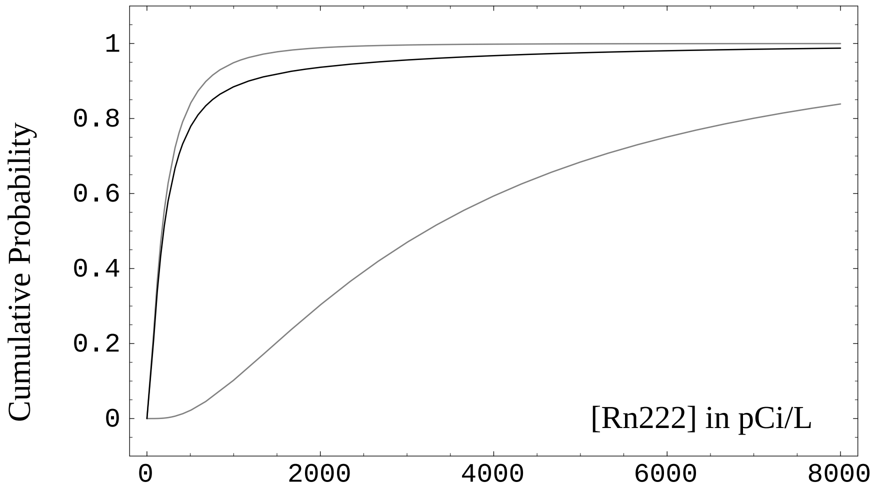

Figure 4A: CDF for Model 2 fit by MLE (black line), flanked by the CDFs for the pure components (gray lines)

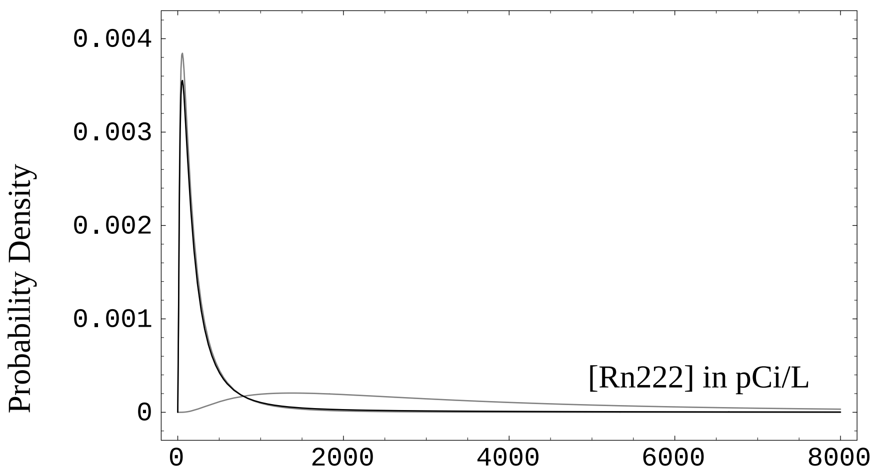

Figure 4B: PDF for Model 2 fit by MLE (black line) and the PDFs for the pure components (gray lines)

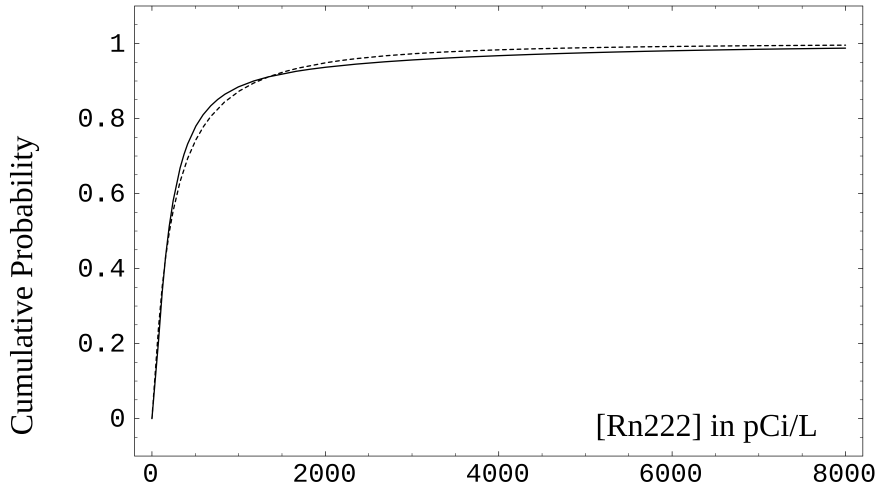

Figure 5A: CDF for Model 1 (dashed line), and
CDF for Model 2 (solid line), each fit by MLE

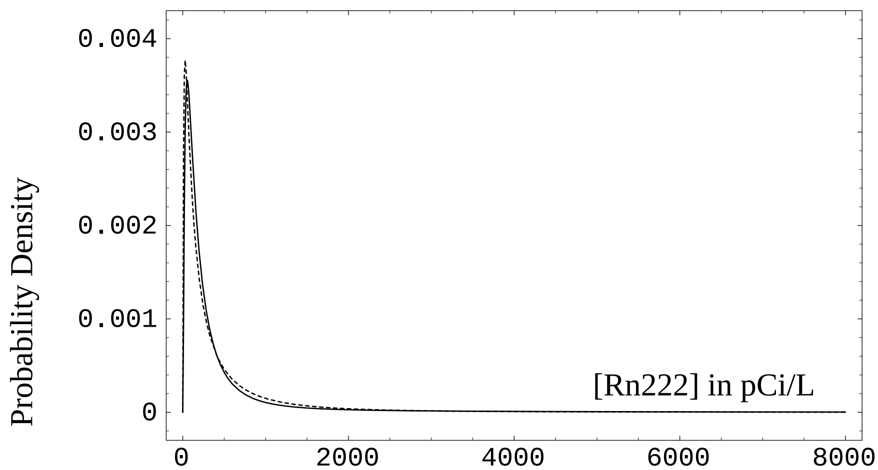

Figure 5B: PDF for Model 1 (dashed line), and
PDF for Model 2 (solid line), each fit by MLE

NIRS Radon 222 Data

Lrg Sys Pop	Lrg Sys pCi/L	Med Sys Pop	Med Sys pCi/L	Sml Sys Pop	Sml Sys pCi/L	VSml Sys Pop	VSml Sys pCi/L	VVSml Sys Pop	VVSml Sys pCi/L
15320	24	6177	24	842	18	455	1	41	1
12810	30	3400	28	750	21	150	1	48	1
70661	32	7800	56	752	28	464	21	42	1
26000	53	9000	100	600	28	225	24	70	26
12500	53	3583	100	1114	33	400	25	53	31
50000	100	4200	100	1290	35	448	32	92	36
126000	100	3500	100	680	35	150	37	50	37
15528	100	5764	100	857	37	400	40	43	38
16250	100	8400	100	743	38	451	43	43	39
14500	100	8778	100	2073	38	120	55	95	40
12643	100	4629	100	2700	48	380	59	40	43
10099	110	6500	100	638	49	156	59	29	48
13000	129	3500	100	1231	50	125	59	45	51
21300	140	3800	100	793	53	270	73	65	54
16000	149	7520	100	700	60	163	100	75	56
35000	160	3688	100	1400	68	450	100	100	60
12700	161	6000	100	1300	73	210	100	60	63
24800	164	10000	100	1250	73	350	100	75	69
20000	172	7350	100	592	74	256	100	100	72
70000	182	4500	104	1862	81	380	100	78	77
12000	222	7000	104	800	95	110	100	88	85
14000	240	4100	108	1480	97	440	100	100	100
17204	240	5500	111	900	100	150	100	32	100
12953	311	9357	112	1200	100	118	100	65	100
13750	468	5750	112	1180	100	203	100	26	100
14205	520	8800	115	521	100	270	100	28	100
10500	558	3862	120	975	100	400	100	30	100
13400	646	9005	122	760	100	429	100	80	100
		5100	130	2983	100	183	100	48	100
		5500	147	2831	100	455	100	55	100
		6500	159	1683	100	240	100	50	100
		9670	167	2150	100	350	100	95	100
		4481	168	3000	100	112	100	28	100
		7800	178	1350	100	220	100	35	100
		4506	194	750	100	375	100	73	100
		5461	199	3106	100	416	100	59	100
		4870	207	800	100	280	100	31	100
		9430	223	1000	100	120	100	99	100
		4060	242	720	100	382	100	100	100
		4300	246	803	100	120	100	30	100
		8000	249	760	100	164	100	30	100
		4860	290	700	100	369	100	56	100
		8356	291	2560	100	138	100	84	100
		4203	307	860	100	107	100	100	100
		3360	308	1545	100	350	100	40	100
		6000	372	2870	100	400	100	31	100
		6350	403	700	100	200	100	60	100
		7272	421	2300	100	272	100	42	100
		6000	522	1950	100	440	100	80	100
		4400	541	550	100	350	100	80	100
		5250	583	1440	100	130	100	40	100
		4500	650	1377	100	300	100	50	100
		8383	796	700	100	312	100	28	100
				600	100	500	100	36	100
				1050	100	293	100	33	100
				1868	100	440	100	50	100
				1578	100	143	100	51	100
				2643	100	171	100	70	100
				1800	100	180	100	44	100

1227	100	256	100	100	100
965	100	500	100	27	100
699	100	294	100	90	100
860	100	182	100	100	100
780	100	500	100	33	100
1200	100	210	100	40	100
1122	100	200	100	59	100
1480	100	184	100	40	100
750	100	227	100	63	100
3063	100	116	100	60	100
1050	100	350	100	36	100
1405	100	285	100	50	100
980	100	146	100	40	100
550	100	454	100	74	100
1100	100	500	100	70	100
921	100	300	100	30	101
2250	100	399	100	100	102
1500	100	189	101	81	102
1360	100	150	101	65	102
1500	100	160	102	40	104
1400	100	220	104	58	104
2000	100	320	104	71	106
1500	100	250	105	54	106
650	100	240	106	40	107
786	100	269	106	52	107
2493	100	483	106	36	109
1512	100	273	106	26	110
1500	100	258	107	50	111
640	100	400	108	25	112
1320	100	106	109	42	117
1033	100	140	109	100	117
625	100	106	111	35	117
1000	100	350	111	100	118
1279	100	285	113	35	120
800	100	355	114	45	121
871	100	120	114	38	121
2100	101	360	115	100	122
838	101	258	115	40	126
750	102	264	116	25	129
875	105	201	117	56	129
1644	106	135	117	35	132
570	107	380	121	68	138
1382	112	450	121	30	139
1000	114	180	121	87	140
690	116	130	121	33	141
600	116	250	122	28	142
552	118	218	123	82	143
550	118	270	124	50	143
2500	118	105	125	54	144
1293	119	250	125	100	144
693	120	385	127	84	144
628	127	275	131	30	146
2449	127	132	131	27	147
564	130	400	133	60	148
639	132	170	134	99	149
1018	132	180	135	36	155
900	133	295	136	55	156
3000	133	500	137	100	157
800	135	164	138	96	157
675	136	343	140	57	157
776	136	144	140	45	161
900	137	362	140	30	162

508	138	140	141	89	164
1610	138	316	141	84	165
1761	139	200	143	50	165
2000	141	200	143	25	166
3015	142	240	144	65	166
1596	144	283	144	30	168
950	147	250	146	30	172
2000	149	283	147	72	173
868	149	141	147	96	175
900	149	409	148	75	177
2600	150	486	150	74	180
600	150	225	151	60	180
994	153	225	151	80	189
650	159	110	151	60	192
2330	160	407	152	68	193
966	165	208	152	60	194
3000	165	400	153	100	194
2200	165	500	153	55	195
782	168	212	159	96	195
1000	172	320	162	96	198
1700	173	104	162	50	199
825	174	425	169	36	199
1500	178	160	170	85	201
2150	181	305	173	50	202
905	186	209	175	90	205
2350	190	350	176	70	206
1200	198	119	176	65	208
850	200	180	176	35	210
580	200	450	176	60	212
1500	201	162	177	33	212
563	202	124	180	50	213
974	203	132	182	73	213
2500	211	260	184	49	213
600	214	130	185	60	220
1000	215	135	185	38	220
839	216	185	186	89	222
700	216	200	186	65	226
1000	217	240	189	90	229
2505	222	240	192	88	229
1200	223	200	194	40	230
890	227	102	195	25	230
1150	237	120	203	100	232
2580	239	295	203	75	232
835	243	115	205	54	244
590	244	190	206	60	244
1050	250	266	207	90	253
813	257	140	209	70	257
3000	258	160	210	100	258
2025	260	207	216	80	260
695	278	346	217	42	267
1000	279	330	218	40	268
1250	279	280	218	77	271
2325	280	500	223	76	272
2950	289	220	224	27	272
800	300	110	226	46	274
1960	312	300	227	26	275
2450	316	159	229	55	276
1300	317	108	230	40	281
1080	336	130	230	57	283
750	346	150	231	30	283
2111	355	259	235	50	285
2000	376	150	236	88	287

2506	381	270	237	60	301
2274	382	200	239	60	311
611	412	250	243	30	311
1750	413	133	245	52	311
1000	414	377	247	60	322
3300	417	200	249	71	324
828	420	133	249	90	328
2400	426	200	251	25	328
1360	429	144	253	70	330
1971	431	350	254	90	334
1200	459	235	256	80	336
2500	459	350	260	66	340
600	464	118	260	50	346
1072	467	180	266	70	348
555	479	130	269	100	349
705	480	115	269	70	351
2700	481	250	270	90	352
2300	483	110	271	48	353
1900	497	132	280	50	353
3200	503	300	282	45	381
1065	514	324	282	99	386
840	514	470	284	36	387
583	521	224	287	75	388
635	528	110	290	71	398
790	531	105	291	25	399
1400	533	252	293	25	402
520	540	375	294	73	413
530	550	393	300	45	419
1900	564	105	301	35	421
1495	575	166	306	35	422
1100	586	340	309	40	424
1000	602	120	326	30	426
1200	604	148	327	50	426
550	614	220	332	58	428
1875	620	145	336	80	432
555	648	110	341	45	435
648	702	150	352	80	435
1053	710	263	354	42	440
700	774	180	355	30	450
800	832	357	357	60	451
1095	1170	133	361	25	455
525	1250	220	367	100	464
680	1350	183	371	50	467
2891	1390	152	373	50	486
1925	1610	174	376	50	487
760	3220	216	378	100	497
753	5120	120	382	98	505
1600	6480	140	386	60	525
623	7830	177	386	25	525
		165	389	56	532
		130	397	41	561
		300	401	56	567
		200	402	78	571
		305	412	65	577
		450	424	28	579
		397	432	29	581
		500	435	40	584
		148	436	75	587
		114	438	60	588
		146	445	93	599
		350	452	80	624
		450	455	90	627

450	465	60	631
214	475	37	633
120	478	100	636
330	479	68	669
105	484	60	670
216	486	51	674
120	495	54	688
135	496	70	691
128	504	32	700
122	519	75	705
120	537	25	710
356	553	75	751
115	580	40	758
355	586	95	767
387	598	85	785
115	604	52	786
440	608	30	821
190	615	70	826
140	627	100	829
230	631	48	846
325	632	45	860
115	635	35	865
132	651	57	874
170	655	27	878
290	656	28	879
115	668	42	880
180	680	50	907
172	683	100	962
300	684	65	999
147	689	30	1040
340	704	64	1060
205	719	30	1070
160	723	65	1070
200	736	60	1080
500	766	96	1130
275	779	45	1150
145	783	30	1170
103	800	96	1240
150	802	60	1240
175	817	42	1250
233	842	65	1260
280	853	90	1320
157	854	75	1320
120	871	60	1340
145	892	50	1350
133	896	59	1370
150	903	68	1400
363	929	60	1410
102	997	30	1450
240	1010	28	1480
125	1200	98	1520
186	1220	35	1560
140	1280	50	1630
224	1320	99	1710
126	1330	75	1830
156	1550	88	1840
328	1570	53	1840
130	1610	35	1950
390	1770	50	2060
400	1810	25	2120
160	2010	45	2160
250	2040	92	2250

252	2140	80	2320
152	2210	40	2370
300	2230	75	2560
130	2600	68	2660
328	2690	84	2680
220	2700	35	2730
264	2930	100	2750
111	3110	84	2750
300	3240	60	2860
208	3260	96	3160
425	3330	66	3200
300	3600	48	3460
190	3730	72	3740
300	3890	42	3800
110	4230	36	4100
444	4370	90	4190
255	5230	90	4260
275	5730	72	4360
156	6990	69	4690
300	7300	63	4850
208	9230	43	5760
280	10600	63	6020
145	11600	30	6350
132	12300	42	8780
200	13200	56	10200
353	14600	76	17500
340	19200	31	21600
		25	25700