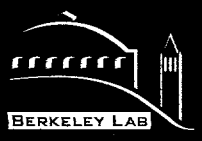
LBNL-43604



ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Development and Evaluation of Probability Density Functions for a Set of Human Exposure Factors

Randy L. Maddalena, Thomas E. McKone,
Agnes Bodnar, and Janet Jacobson

Environmental Energy Technologies Division

June 1999

LBNL-43604

Development and Evaluation of Probability Density Functions for a Set of Human Exposure Factors

Randy L. Maddalena
Environmental Energy Technologies Division

Thomas E. McKone
Environmental Energy Technologies Division
and
School of Public Health
University of California, Berkeley

Agnes Bodnar
Environmental Energy Technologies Division

Janet Jacobson
Earth Sciences Division

Funded by
U.S. Environmental Protection Agency
Office of Emergency and Remedial Response (OERR)

June 1999

University of California
Ernest Orlando Lawrence
Berkeley National Laboratory
Berkeley, California

This work was supported by the U.S. Environmental Protection Agency and carried out at Lawrence Berkeley National Laboratory through the U.S. Department of Energy under Contract Grant No. DE-AC03-76SF00098. Environmental Protection Agency funding was provided by the Office of Emergency and Remedial Response (OERR) through Interagency Agreement #DW-899-38067-01-0.

Contents

Executive Summary

The purpose of this report is to describe efforts carried out during 1998 and 1999 at the Lawrence Berkeley National Laboratory (LBNL) to assist the U. S. EPA in developing methods for constructing, evaluating, and scoring the robustness of probability distributions. The U.S. Environmental Protection Agency (EPA) Office of Emergency and Remedial Response (OERR) is in the process of updating its 1989 Risk Assessment Guidance for Superfund (RAGS) as part of the EPA Superfund reform activities. Volume III of RAGS, when completed in 1999 will provide guidance for conducting probabilistic risk assessments. This revised document will contain technical information including default probability density functions (PDFs) and methods used to develop and evaluate these PDFs.

Probabilistic risk assessment methods have emerged that promise to improve the way that uncertainty and variability are treated and communicated. The effectiveness of these methods is largely dependent on our ability to characterize the uncertainty and variability associated with exposure factors using probability distributions. A variety of methods are available for developing these distributions from raw data or from summary statistics. However, a framework for determining when a distribution is appropriate for a given assessment has not yet been established. In an effort to develop a practical and reliable method for evaluating the performance and appropriateness of PDFs, LBNL has collected and critically evaluated data for the following exposure factors:

- body weight
- exposure duration (amount of time living at a residence)
- exposure frequency (fraction of the day spent at the exposure location)
- total water intake and
- inhalation rates

For each of these exposure factors, available data was collected and used to develop evaluate, and score distributions. The most appropriate distribution for each subset was selected based on a combination of standard procedures and on a novel graphical method. Lessons learned during the data collection, evaluation and distribution development process were used to design a scoring system based on the quantity, quality and relevance of the data and on our ability to identify a parametric model (or other distributional form where appropriate) that adequately describes the data. A key contribution of this report is the development of a simple method for scoring the quality of distributions in the context of the cohort/population to which the distribution is to be applied.

Data Collection

Several studies have reported distributions for exposure factors that were developed from percentiles or statistical summaries of the data. Although this approach for developing distributions is statistically sound and economically feasible, it fails to provide adequate detail about influential factors, possible sub-populations within the sample and the power of the selected parametric distributions. Therefore, it was important for this work to use raw data from its original source whenever possible.

An exhaustive review of the available literature was performed and the best candidate sources of data were identified for each exposure factor. When several sources were equally suited for a given problem, a decision about which data set to use was based on 1) how well the sample survey represents the overall US population and demographic subsets of the population, 2) the number of samples in the data set, 3) the number of individual exposure factors included in the survey and how well the measurements or reported values represent those exposure factors and 4) the availability and usability of the data. Data sources that were not used in the initial development of distributions may be used at a later date to evaluate the performance of the recommended distributions against independent samples using cross-validation experiments. Information gained from cross validation and analysis of random samples selected from independent surveys will help to further characterize the power of the recommended distributions and to highlight those exposure factor distributions that are particularly robust across the population.

With exception of inhalation rates, results from high quality nationally representative surveys were available for all exposure factors. Although the body weight values were self-reported, we expect these values to be reasonably accurate. The reported current residence time (length of time that sample person has lived in their current home) was used as a surrogate value to estimate exposure duration. Without further study it is not apparent how the use of the surrogate data will affect the reliability of the estimated exposure duration. Exposure frequency values were both self-reported as well as surrogate in that short-term diary data was used to represent long term behavior.

Self-reported water intake values were obtained from the same nationally representative survey used for body weight. However, the relevant questions in the survey did not elicit appropriate information for direct estimate of tap water intake. The survey data included information on the amount of tap water consumed by the sample person as plain drinking water and the amount and type of food and beverage (coffee, tea, juice) ingested. To estimate the total amount of tap water ingestion one would need to estimate the amount of tap water used in the preparation of each food and beverage and then use the intake data to estimate indirect tap water ingestion. Developing and validating a database of the extrinsic water content of each food and beverage is beyond the scope of this project. We were able to extract information on total water

intake (extrinsic plus intrinsic) and this information was used to demonstrate the procedure for water intake distributions.

For inhalation rate, we selected a small data set from a well-designed experiment that directly measured inhalation rate for five distinct activity levels. The study included a number of demographic variables and the experimental design used a demographic composition (age, gender and race) that represented the population in California. Although the values from the study were directly measured, the data must be treated as a proxy for actual inhalation rates due to the complex relationship between inhalation rate and activity.

Data Analysis

Default values and distributions for exposure factors are commonly reported for specific age and gender classes. However, it is rare that these individual classes are tested to determine whether they are statistically different from one another given the inherent variability in the population and the quality of the data. We used Classification and Regression Tree (CART) data mining software to systematically identify the optimum number of statistically different subsets within the sample for each exposure factor. Results of the CART analysis are concise, easy to understand, and are appropriate for use as a decision making tool. The technique was developed almost 20 years ago and has been applied in many fields, including engineering, medicine, public health and economics. The CART analysis provided a statistically and scientifically defensible approach for identifying important variables in complex data sets. These variables are used to decompose the original data set into important demographic subsets. The classification provides an initial family of data sets for each exposure factor that are based solely and objectively on the data. Other demographic subsets of the population may need to be considered for subjective or policy reasons.

Our analysis of the body weight data indicated that age, gender and race were important variables for constructing demographic subsets of the population. Not surprisingly, the body weight of children less than 12 years is only dependent on age and the group was separated into four age groups. For individuals 12 years of age and older, both gender and age became important. An interesting finding was that race was an important factor regarding variability in weight among adult females. Black and American Indian females, Asian/Pacific Islander females and Caucasian females all had statistically different body weights. In addition, adult Asian/Pacific males had significantly lower body weights than other adult males.

For exposure duration the age of the person sampled and housing status (renter or owner) were important variables. For people over 67 years of age, the region that the individual lived in was also important. Younger people reported shorter current residence times and, for each age group, renters reported shorter residence time than people living in owner-occupied houses.

Exposure frequency was defined in this study as the fraction of the day that an individual spends indoors at home. This factor depended mainly on employment status-- people who were employed spend less time indoors at home than the unemployed or part time employed. Not surprising, the activity pattern for employed individuals changed significantly on weekends.

For water ingestion data, the reported intake values were normalized to body weight (liters per kg day) prior to the CART analysis to reduce the inter-individual variance. The total water intake (normalized to body weight) for children 10 years of age and younger is strongly age dependent--children drink less water per unit body weight as they get older. The analysis indicated that, on a body-weight basis, adults living in the Northeast and South consume less water than do those living in the Midwest and West. No clear explanation for this regional difference was found although the results were consistent with previous work (Ershow & Cantor, 1989). The analysis also indicated that race may be an important variable and that pregnant and lactating women consume more water (comparable to adult males) than women who are not pregnant.

Preliminary analysis of inhalation rate data indicated that variance could be reduced by normalizing the values to body weight. The normalized inhalation rate data for children was similar to that of water ingestion where intake per unit body weight decreased with increasing age. This change in intake with age (for young children) is likely due to changes in metabolism as the child grows (CARB, 1993). Although the dependence of intake on age was obvious, the overall sample size was not adequate to clearly establish differences due to gender or race. Thus, the data was simply separated into children younger than 11 years of age and adults at the five activity levels reported in the study.

For future applications, we propose a two-stage procedure to identify the optimal number of demographic regions for each exposure factor. In the first stage of the procedure, CART or a comparable approach is used to systematically split the data set into statistically different demographic regions and each split is ranked in order of importance. The second stage of this procedure will use sensitivity analysis techniques to collapse the family of distributions for a given exposure factor into the fewest number of significantly different demographic regions (age, gender, race, etc.).

Distribution Development

After splitting the data into individual sample sets for each exposure factor, standard parametric distributions (when appropriate) were identified and fit to the data. Methods for fitting parametric distributions to data are well established and statistical software is available with the capability of automating much of the work. Standard methods were used to narrow down the choice of parametric distributions for each data set. A graphical method was then developed and

used to identify the best parametric model. The graphical method uses a plot of the fitted parametric model and the empirical cumulative distribution along with an overlay of a plot of the vertical residuals (i.e., residuals in the estimated percentiles) between the parametric model and the data. A 95% confidence band is used with the residual plot to show where no further improvement in fit can be expected or justified.

In addition to providing a visual stopping rule, the residual method highlights the regions of the distribution where the selected parametric model provides the best fit and the poorest fit. This feature is important because we are often more interested in a specific region of the distribution and visualization can facilitate judgements about which distribution works better in the important region even if the goodness of fit score indicates otherwise. Final distributions are selected for their simplicity, their theoretical representation of the particular exposure factor and their overall fit to the data.

Lognormal and extreme value distributions provide the best fit for each of the body weight data sets except that of the youngest group for which a logistic distribution worked best. There was no apparent theoretical basis for choosing the lognormal over the extreme value so the decision was based solely on quality of fit. Due to the nature of the data used to estimate exposure duration, no distributions were fit for this exposure factor. Rather, the strengths and limitations of previously published distributions are identified. Truncated logistic distributions worked well for all of the exposure frequency data sets except for employed individuals during the week (not on the weekend). For these individuals a mixture model (logistic and uniform) was required. The mixture was likely due to the activity pattern of a subset of sample persons the day of the survey (most were at work but some may have been home sick or on vacation) but adequate information was not available to further decompose the data. Lognormal distributions worked well for all water intake and inhalation data sets.

Distribution Scoring

A key contribution of this report is the development of a simple method for scoring the quality of distributions in the context of the cohort/population to which the distribution is to be applied. The scoring system is not limited to measuring how well a model fits the data that was initially used to construct the distribution – this can be accomplished using standard goodness of fit procedures. Rather, the method developed in this report is designed to facilitate decisions about how well a distribution is expected to perform with a different but presumably similar data set or sample.

For example, a distribution of produce ingestion rates from a well-designed study performed in California may be deemed appropriate for a similar population in Minnesota. However, such a decision requires an intimate familiarity with the data (population) that was

used to construct the distribution, the ability of the distribution (parametric or otherwise) to describe the original data, and the degree of similarity between the original data set and the population to which the distribution is to be applied. The scoring system used here is designed to facilitate this decision making process. The system is based on a combination of quantitative and qualitative information for each distribution. The quantitative information includes factors such as sample size, confidence intervals about the distribution, sensitivity of the exposure equation to the particular exposure factor and graphical/analytical measures of how well the recommended PDF represents the available data. The qualitative information includes knowledge of how well the sample survey captures the demographics of the population and how well the sampled data represents the particular exposure factor.

The proposed scoring system uses a questionnaire type format designed to combine the quantitative and qualitative information into a single scenario-specific measure for the quality of a given parametric model (or other form of distribution). Although the final scores fall on a continuum from "not applicable" to "highly recommended", the continuum is partitioned into four regions defined as "highly recommended" for use (H), "medium quality" (M), "low quality" (L) and "not applicable" for use (NA). This partitioning is admittedly subjective but it provides a picture of performance that can be used to help facilitate decisions about the appropriateness of a distribution for a given application.

Using the scoring system developed in this report, all of the distributions for body weight were found to be either highly applicable or of medium quality. The medium scores were a result of small sample size for some of the demographic subsets used in the analysis. The exposure duration distributions scored a medium to low due to the high degree of qualitative uncertainty associated with the use of current residence time to estimate exposure duration. In addition, distributions for demographic regions of the population identified as important in this study have not been developed in the literature. Distributions for exposure frequency scored medium to low because of significant qualitative uncertainty about the relevance and representativeness of the short-term diary data used to approximate time spent indoors at home.

Although the sample size and data quality used to develop distributions for water intake were very good, the distributions score poorly on relevance. The lack of direct information about tap water intake (both direct and indirect) makes the high quality distributions developed for total water intake irrelevant to most exposure assessments. To make the distributions appropriate, one would either have to convert the value from total water intake to tap water intake using an appropriate metric or approximate the amount of tap water ingested with food and beverage then reconstruct the distributions accordingly. Due to the lack of relevance of the data used to construct the distributions, the water intake distributions score low. The inhalation rate distributions score medium because the data is of good quality and representative of the population but the sample size is small.

Findings and Recommendations

One of the main lessons from the LBNL project is that judging the quality of a distribution requires a clear and complete understanding of the data used to develop the distribution in question, the procedure used to construct the distribution and the population in the analysis objective. The scoring method described here provides a simple and reliable tool for developing that understanding. Findings and recommendations from this work are summarized below:

(i) To score the quality, reliability and relevance of a distribution, it is critical that the user have a clear and complete understanding of the data used to develop the distribution in question, the procedure used to construct the distribution and the population that the distribution will be used to represent. Whether this understanding comes from developers of the distributions, the user of the distributions or a combination of the two is not readily apparent.

(ii) When a large amount of data is available, CART is an efficient and effective tool for identifying the most appropriate way to split complex data along demographic lines. Further splits in the data may be necessary for political or policy reasons but that is beyond the scope of this report.

(iii) Systematic methods incorporating sensitivity/uncertainty analysis should be use to determine when and to what degree the demographic subsets of data identified by CART can be recombined to form the optimal number of members in the family of distributions for each exposure factor.

(iv) Future work should be directed towards better understanding how to fit truncated distributions and how truncated distribution influences the calculation of dose/risk.

(v) Although not included in the body of this report, we found that model-free methods show promise as a tool for learning more about the underlying shape of distributions but more work is needed to determine just how useful they might be.

(vi) Neither set of currently available exposure duration (ED) distributions include information on ethnicity or socio-economic status. Such distributions could be determined by applying the analytical/statistical procedures of either Israeli and Nelson (1992) or Price *et al.* (1998) to the 1995 AHS-N data or the Monte Carlo procedure of Johnson and Capel (1992). In addition, the information was split on variables that were not found to be important in this analysis (gender, multiple age groups) and the strong relationship between young adults and children living at home was not accounted for. A

reanalysis of the methods and data used to estimate ED and the construction of new distributions that can be tested using the methods introduced in section 3 is warranted.

(vii) All of the exposure factors included in this report can benefit from cross-validation experiments designed to test the performance of the parametric models (or other distributional forms) against independent data sets.

(viii) A better understanding of the relevance of short-term diary data for estimation of activity patterns and exposure frequency is warranted.

(ix) Direct measurements for inhalation rates are limited. Resources should be directed towards the collection of quality data that can provide a better understanding of the physiological differences and inter-individual variability.

(x) The relevance and reliability of nutrition studies for estimating water intake should be verified using a series of small-scale studies designed specifically to estimate the amount and source of water consumed by various demographic subsets of the population.

1.0 Introduction

The purpose of this report is to describe efforts carried out during 1998 and 1999 at the Lawrence Berkeley National Laboratory (LBNL) to assist the U. S. EPA in developing and ranking the robustness of a set of default probability distributions for exposure assessment factors. Among the current needs of the exposure-assessment community is the need to provide data for linking exposure, dose, and health information in ways that improve environmental surveillance, improve predictive models, and enhance risk assessment and risk management (NAS, 1994). The U.S. Environmental Protection Agency (EPA) Office of Emergency and Remedial Response (OERR) plays a lead role in developing national guidance and planning future activities that support the EPA Superfund Program. OERR is in the process of updating its 1989 Risk Assessment Guidance for Superfund (RAGS) as part of the EPA Superfund reform activities. Volume III of RAGS, when completed in 1999 will provide guidance for conducting probabilistic risk assessments. This revised document will contain technical information including probability density functions (PDFs) and methods used to develop and evaluate these PDFs. The PDFs provided in this EPA document are limited to those relating to exposure factors.

Exposure assessments use of a number of factors that are both variable and uncertain. As a result, the magnitude of these factors can not accurately be represented by a single value in a risk assessment, but must be characterized by a range of values reflecting both the population variability and the uncertainty that results from limited and imprecise data. Methods have emerged that promise to improve the way that uncertainty and variability in risk assessment are characterized and communicated. The effectiveness of these methods is largely dependant on our ability to characterize the uncertainty and variability associated with the individual factors that are used in the calculations. Arguably, the most powerful way to characterize and use stochastic inputs is through probability distributions.

LBNL has evaluated for EPA the quality, reliability and relevance of data and distributions for the following exposure factors:

body weight

exposure duration (amount of time living at a residence)

exposure frequency (fraction of the day spent at the exposure location)

inhalation rates

water intake rates

For each of these factors a family of PDFs has been developed and evaluated according to the quantity, quality and relevance of the data available to construct the PDFs and our ability to identify a parametric model that adequately describes the data.

1.1 The Use of Probability Distributions in Exposure Assessment

Estimating potential human exposures involves the use of large amounts of data coupled with the use of models. Because these data and models must be used to characterize individual behaviors, contaminant transport, human contact, and uptake among large and often heterogeneous populations, there can be large variabilities and uncertainties associated with exposure predictions.

One common approach to address variability and uncertainty in exposure and risk assessments is the practice of compounding upper bound estimates in order to make decisions based on a highly conservative estimate of exposure. Such compounding of upper bound estimates leaves the decision maker with no flexibility to address margins of error; to consider reducible versus irreducible uncertainty; to separate individual variability from true scientific uncertainty; or to consider benefits, costs, and comparable risks in the decision making process. Because the compounding of conservative estimates does not serve the exposure assessment process well, there has been a growing effort to include explicit variance propagation and uncertainty analyses into the risk assessment process.

For human populations, total exposure assessments that include time-activity patterns and micro-environmental data reveal that an exposure assessment is most valuable when it provides a comprehensive view of exposure pathways and identifies major sources of variability and uncertainty. Probability distributions are the most versatile and informative means for describing uncertainty and variability in model inputs.

1.2 Aims of this Study

The work reported here has two specific aims. The first aim is to develop PDFs that reflect variability and uncertainty in a set of exposure factors. The second aim is to develop and apply a system for scoring these and other distributions for use in risk assessments.

The first specific aim of this study is to contribute to and evaluate two types of PDFs for use in exposure assessments. One type reflects variability and uncertainty on a national scale, and the second type reflects variability and uncertainty in specific subsets of the population. In carrying out this aim, we assess the extent to which the data available to construct the

distributions support dividing the population into subgroups. In constructing these distributions we also identify gaps in the available data that, if filled, could help to better elucidate important subgroups within the general population. As noted above, exposure factors included in this study are body weight, exposure duration grouped by categories (e.g., home owners vs. renters, rural vs. urban residents), exposure frequency (days per year at the exposure location), inhalation rates, and water intake rates.

The second specific aim of this study is to develop and apply a system for scoring or ranking the resulting PDFs for each of the above-identified exposure factors and recommend ways to improve quality, reliability and relevance of the distributions where necessary. Some of the more common exposure factors included in this study (body weight, water intake) have extensive nationally representative data sets from which distribution scores can be developed and tested. However, other exposure factors must be derived from proxy measures (exposure duration) or developed from small sets of data (inhalation rate). The purpose of the scoring methodology is to provide a means for characterizing the overall value of a distribution for use in risk assessments.

1.3 Overview of the report

The remainder of this report consists of eight sections. In the next section, Section 2, we provide a background on the development and ranking of PDFs in exposure and risk assessment. In Section 3, we describe the methods used in the study. Included here are the methods used to identify and assemble data sources; methods for identifying the subgroup structure of large data sets and decomposing the data by demographic factors; methods for distribution development including statistical models and data simulation using non-parametric methods; and definition of a system for scoring the PDFs and the data sets used to develop the distributions. In Sections 4 through 8, we describe the development and evaluation of PDFs for, respectively, body weight, exposure duration, exposure frequency, breathing rates, and water intake. Section 9 provides summary discussion and recommendations for this effort.

2.0 Background

An important step in the process of conducting an uncertainty analysis is the construction of a PDF or CDF for each model input. The purpose of these distributional representations is to define the range of values that an input can take on and to assign a probability of obtaining any particular value within that range. One step in the process of constructing a probability distribution is to define the range and moments of the input data. Once this is done we can use various subjective, graphical, and statistical methods to select an appropriate distribution to represent inputs and to judge the effectiveness with which the distribution fits the data. These issues are addressed in this section.

2.1 The Need for Distributional Inputs to Models

Exposure models are used to describe the relative magnitude and variation in human contact with environmental contaminants. An important attribute of exposure models is the ability to account for factors that control variation in human contact, i.e. age, gender, location, activity patterns, etc. Uncertainties limit the ability of models to characterize these relationships. Uncertainty in model predictions arises from a number of sources, including specification of the problem; formulation of the conceptual model; formulation of the computational model; estimation of input values; and calculation, interpretation, and documentation of the results. Of these, only uncertainties due to estimation of input values can be quantified in a straightforward manner using variance propagation techniques. Uncertainties that arise from mis-specification of the problem and model formulation are clearly important but fall outside of the scope of this report.

Single value inputs to models fail to express exposure variation and the uncertainty that arises from the use of incomplete and proxy data. Such issues can be addressed in part with the use of probability distributions as inputs to models. The value of information derived from an uncertainty analysis is very much dependent on the care given to the process of constructing the input parameter distributions.

The data, scenarios, and/or models used to represent human exposures to environmental contaminants include at least five important relationships:

(i) The magnitude of the source medium concentration, that is, the level of contaminant in the air, water, soil, and food with which the population has contact;

(ii) the contaminant concentration ratio, which defines how much a source-medium concentration changes as a result of transfers, transformation, partitioning, dilution etc. before human contact;

(iii) the level of human contact, which describes (often on a body-weight basis) the frequency (days per year or hours per day) and magnitude (m^3/day, L/day or kg/day) of human contact with a potentially contaminated exposure medium;

(iv) the frequency and duration of potential contact for the population of interest as it relates to the fraction of lifetime during which an individual is potentially exposed; and

(v) the averaging time for the type of health effects under consideration, i.e. is the appropriate averaging time the cumulative duration of exposure (as is typical for cancer and chronic diseases) or some relatively short time period (as is the case for acute effects).

2.2. Variability and Uncertainty

One of the issues in uncertainty analysis that must be confronted is how to distinguish between the relative contribution of variability (i.e., heterogeneity) versus true uncertainty (measurement precision) to the characterization of predicted outcome. Variability refers to quantities that are distributed empirically–such factors as rainfall, soil characteristics, weather patterns and human characteristics that come about through processes that we expect to be stochastic because they reflect actual variations in nature. These processes are inherently random or variable and cannot be represented by a single value, so that we can only determine their moments (mean, variance, skewness, percentiles, etc.) with precision. In contrast, true uncertainty or model-specification error (e.g., statistical estimation error) refers to an input that, in theory, has a single value that can not be known with precision due to measurement or estimation error.

In many situations, an exposure model is used to characterize the relative magnitude and importance of parameter uncertainty (lack of information) versus parameter variability (inter-individual variation). It is important to distinguish between the two forms of variance because each plays a unique role in decision making. Uncertainty can be reduced by further experimentation while variability can only be better characterized (i.e., uncertainty about the variability can be reduced through further experimentation).

To fully express the combined impact of uncertainty and variability, it is sometimes necessary to carry out a two-dimensional Monte Carlo simulation consisting of an inner set of calculations embedded within an outer set. Bogen and Spear (1987) first described this approach. In the first phase, a single realization is obtained from the distribution of each uncertain parameter, followed by repeated sampling from the variable parameters. This process is repeated until a large number (~ 500) of uncertain parameter value sets are taken in the outer phase and a larger number (~ 1000 or more) of the variable parameter values are selected. The simulation

results are plotted as either a two-dimensional surface or as a family of variability curves at different levels of uncertainty. Typical results for this type of simulation are shown in Figure 2.1.

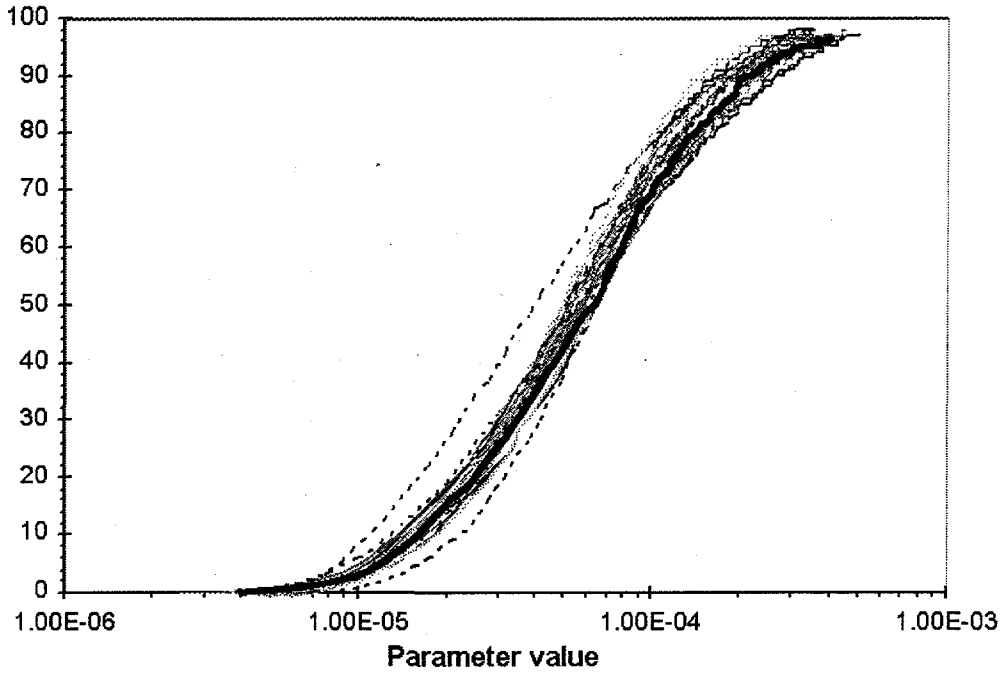


Figure 2.1. A set of cumulative probability plots that reflect both variability and uncertainty. Each curve expresses a realization of a variability distribution at a different level of uncertainty.

2.3 The Link Between PDFs and Data

When constructing input distributions for an uncertainty analysis, it is often useful to present the range of values in terms of a standard probability distribution. It is important that the selected distribution be matched to the range and moments of any available data. It is often appropriate to simply use the raw data or a custom distribution. Other more commonly used standard probability distributions include the normal distribution, the lognormal distribution, the uniform distribution, the log-uniform distribution, and the triangular distribution.

Probability distributions are typically displayed as probability density functions (PDFs) or as cumulative distribution functions (CDFs). For a continuous distribution, the PDF is a smooth function, f(x), which represents the probability that a parameter x has a value between x –dx and x +dx, where dx is an infinitely small interval. In a CDF, F(x) represents the probability that a parameter, x, has a value less than or equal to any value, x. Figure 2.2a shows a PDF for a continuous distribution and Figure 2.2b shows the corresponding CDF for this distribution.

These two diagrams illustrate the PDF and CDF that would be used to represent a normal distribution with mean value, 1, and standard deviation, 0.3. Continuous distributions, like the samples in Figure 2.2, may be separated into three categories: (1) Those that represent only variability, (2) those that represent only uncertainty and (3) those that represent both variability and uncertainty.

There are a number of methods for constructing a parametric PDF to fit observations (data). These include moments matching, graphical methods, goodness of fit tests and Bayesian methods, among others (Cullen and Frey, 1999; D'Agostino and Stephens, 1986). When these methods are applied, one obtains a distribution that has in some way been optimized to fit the available observations. Often, more than one parametric model can be satisfactorily fit to a data set. The function used to optimize model fit rarely provides decisive information about which of two models is better. Thus, the final choice of the best model for describing a given data set is often subjective or based, at least in part, on theory, convention and/or convenience.

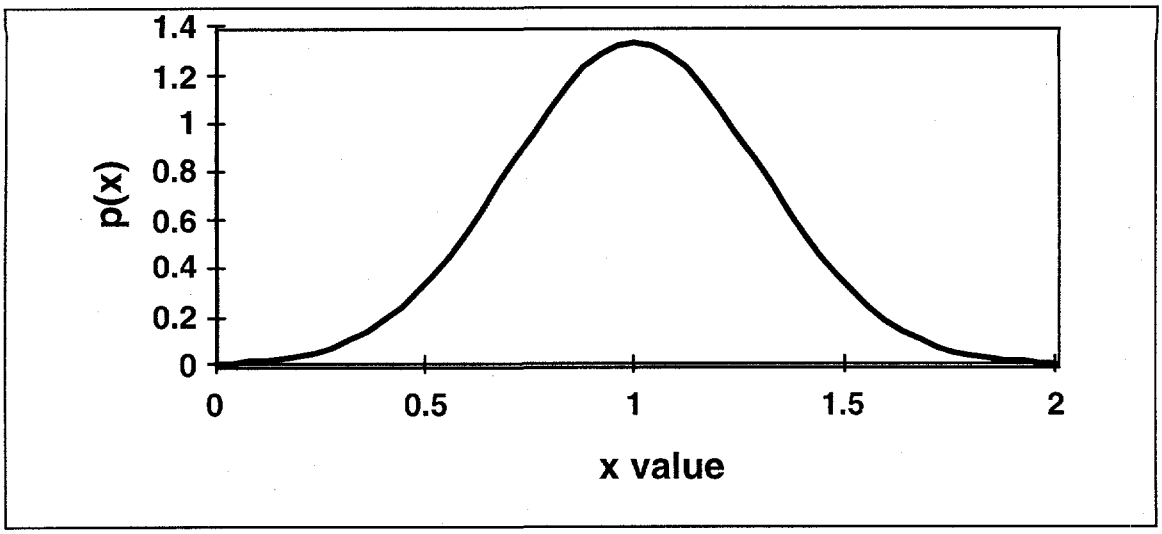


Figure 2.2a. Probability density function for a continuous distribution.

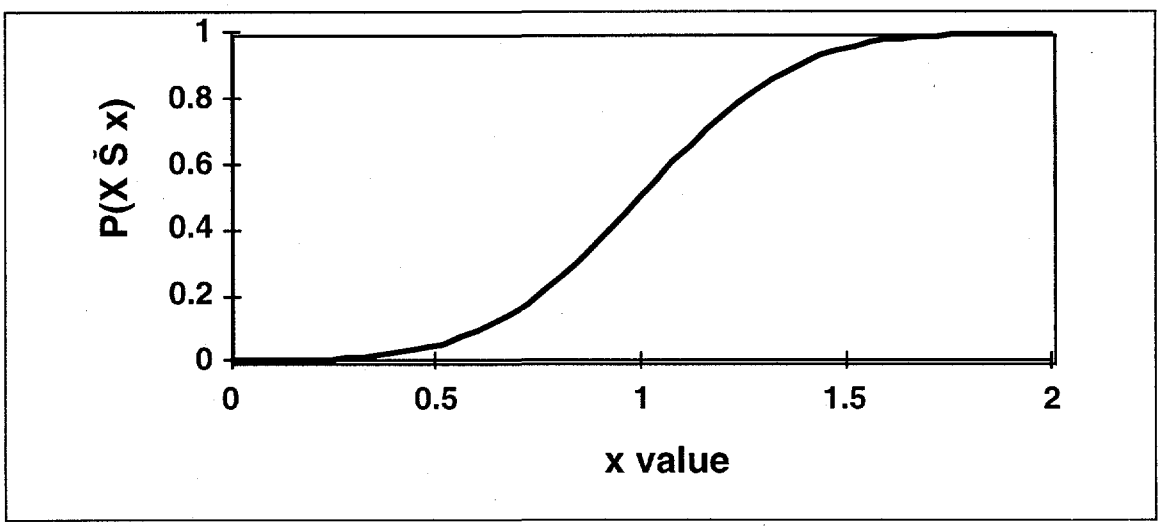


Figure 2.2b. Cumulative density function for a continuous Normal distribution with mean of 1 and standard deviation of 0.3.

2.4 A Tiered Approach to Uncertainty/Variability Analysis

As was noted earlier, it has been determined that compounding of upper bound estimates is no longer considered an appropriate approach to exposure and risk assessment. A more reasonable approach is one that provides the decision maker with flexibility to address margins of error; to consider reducible versus irreducible uncertainty; to separate individual variability from true scientific uncertainty; and to consider benefits, costs, and comparable risks in the decision making process. In order to make an exposure assessment consistent with such an

approach, it should have both sensitivity and uncertainty analyses incorporated directly into an iterative process by which premises lead to measurements, measurements lead to models, models lead to better premises, and better premises lead to additional, but better-informed measurements, and so on. In 1996, the U.S. EPA Risk Assessment Forum held a workshop on Monte Carlo Analysis. Among the many useful discussions at this meeting was a call for a "tiered" approach for probabilistic analysis, which is iterative and progressively more complex. The need for formal uncertainty analysis and a tiered approach will require the development by the exposure assessment community of new methods and will put greater demands on the number and types of exposure measurements that must be made. In such an approach at least three tiers are needed. These are described below.

First, the variance associated with all input values should be clearly stated and the impact of these variances on the final estimates of risk assessed. At a minimum, this can be done by listing the estimation error or the experimental variance associated with the parameters when these values or their estimation equations are defined. It would help to define and reduce uncertainties if a clear summary and justification of the assumptions used for each aspect of a model are provided. In addition, it should be stated whether these assumptions are likely to result in representative values or conservative (upper bound) estimates.

Second, a sensitivity analysis should be used to assess how model predictions are impacted by model reliability and data precision. The goal of a sensitivity analysis is to rank the input parameters on the basis of their contribution to variance in the output.

Third, variance propagation methods (including but not necessarily limited to Monte-Carlo methods) should be used to carefully map how the overall precision of risk estimates is tied to the variability and uncertainty associated with the models, inputs, and scenarios.

2.5 The Need for PDF Scores

The quality or validity of the PDFs that are used in a probabilistic exposure analysis directly influences the reliability of the predictions or decisions that are based on the exposure analysis outcome. Many times default distributions are prescribed. In such situations, there is a risk that policy guidelines can be looked on as fact. Default values need to be clearly represented as to their quality, reliability and relevance for various exposure scenarios.

One begins the process of constructing a distribution function for a given parameter by assembling values from the literature or from personal knowledge. These values should be consistent with the model and its particular application. The values will vary as a result of measurement error, spatial and temporal variability, extrapolation of data from one situation to another, lack of knowledge, etc. The process of constructing a distribution from limited and

imprecise data can be highly subjective. Because the uncertainty analyst must often apply judgment to this process, there is a need for expertise, wisdom and an intimate relationship with the data. The process becomes more objective as the amount of data for a given parameter increases. However, a large set of data does not necessarily imply the existence of a suitable distribution function nor does it imply that the data is directly relevant to the problem or question.

As was noted above, PDFs are developed from data sets and there are a number of methods for making the best fit of a distribution to the data. When these methods are applied, one obtains a distribution that provides an optimum fit to the available data. However, once this process is completed, the resulting distribution does not provide the user of that distribution with a quantitative measure of how well the distribution replicates either the underlying data or the true variability of the exposure factor being represented. What is needed to address these issues is some measure of the quality of the distribution as it relates to the subject factor.

In order to define a score we consider those factors that would increase one's confidence about a given distribution. These factors include quantity of data available, sources and measurements techniques used to collect data, data quality, relevance and representativeness of the data, mixtures of distributions with contamination of the data, and correlation among inputs. These issues must be explicitly addressed to develop a score for any distribution. By developing a series of questions in a score sheet or questionnaire, one can systematically incorporate all of the confidence-building-factors into a single measure of quality, reliability and relevance.

3.0 Methods

This section provides an overview and explanation of the methods and procedures used to collect and analyze existing data and to select, fit and score the parametric distributions for each of the five exposure factors included in this study. This effort began with an exhaustive literature search to identify candidate data sets for each exposure factor. The best data set for each factor was then identified based on a simple set of decision rules. Next, each data set was processed to remove extraneous information, data were converted to appropriate units and formatted for factor analysis. The factor analysis used an objective and systematic data mining technique to identify the population descriptors that have the greatest contribution to central tendency and variance. Individual data sets were constructed for each of the important factor classes and statistical/graphical techniques were used to identify and parameterize the best distribution for each class.

After the appropriate distribution for each of the individual classes within each exposure factor is constructed, a series of Monte Carlo analyses will be designed and performed using the basic exposure equation. Results from these analyses will help characterize how sensitive the calculation of dose is to differences in the class specific distributions. By comparing the outcome distribution from each combination of class specific distributions we proposed to collapse the class distributions into the final set of recommended distributions for each exposure factor. In this way, one can highlight what actually influences the variability of each exposure factor (age, gender, region, etc.) then test how this variability contributes to the overall estimate of exposure. Thus, one can provide an objective recommendation for the appropriate number of classes for each exposure factor and the appropriate level of complexity for the class specific distributions.

Finally, we measure both the analytical and subjective power of each distribution and summarize it with a simple scoring system. The 4score is designed to incorporate information about how well the data represents the exposure factor of interest (quality, quantity and relevance of data) and how well the parametric model represents the data. The scoring procedure is designed to provide an indication of how well the recommended distributions are expected to work for a subset of the population or sample that was not included in the original data set.

3.1 Literature Review and Summary of Data Sources

Several studies have previously presented distributions for exposure factors that were based on the reported percentiles or statistical summaries from sample surveys (Burmaster, 1998; ODEQ 1998; RTI, 1998). Although this approach for developing distributions is statistically sound and economically feasible, it fails to provide adequate detail about influential factors, possible sub-populations within the sample and the power of the selected parametric

distributions. Therefore, it was important for this work to use raw data from its original source whenever possible.

An exhaustive review of the available literature was performed to identify the best candidate sources of data for each exposure factor. Often there were several data sources that were equally well suited for a given problem and a decision about which data set to use was based on convenience (ease of access and use or multiple exposure factors included). The remaining data sources can be used at a later date to measure how well the recommended distributions predict independent samples using cross-validation experiments. Information gained from cross validation and analysis of random samples selected from independent surveys will help to further characterize the power of the recommended distributions and to highlight those exposure factor distributions that are particularly robust to changes in demographics across the population.

The original source of candidate data sets for this study are given in Appendix 1, Tables A.1 through A.5 along with a brief description of each data set, where the data has been used or referenced and the contact person(s) or data source (if known). The data sets are listed in order of relevance to this study in each exposure factor table. The specific rational for selecting a data set for a given exposure factor is provided in the sections related to each factor. The basic decision points that were used to identify the best candidate data sets include:

1) how well the sample survey represents the US population,

2) how well the survey accounts for demographic regions within the population,

3) how well the measured or reported value represent the exposure factor (measured values are better than self reported values and self reported are better than surrogate values),

4) how large the sample size is,

5) how easy the data set is to obtain and work with,

6) the number of exposure factors and independent variables included in the survey (this is both for convenience and for investigating potential correlation within the population at a later date)

3.2 Measuring the Importance of Demographic Factors

A number of the distributions and parameters associated with exposure factors have been related to age and gender classes (Oregon DEQ 1998, Burmaster, 1998; BOC, 1995). However, it

is rarely tested whether or not these individual classes are indeed statistically different from one another given the inherent variability in the population and the quality of the data. We use Classification and Regression Tree (CART) data mining software to analyze the full data sets and systematically identifying the optimum way to decompose the data for each exposure factor.

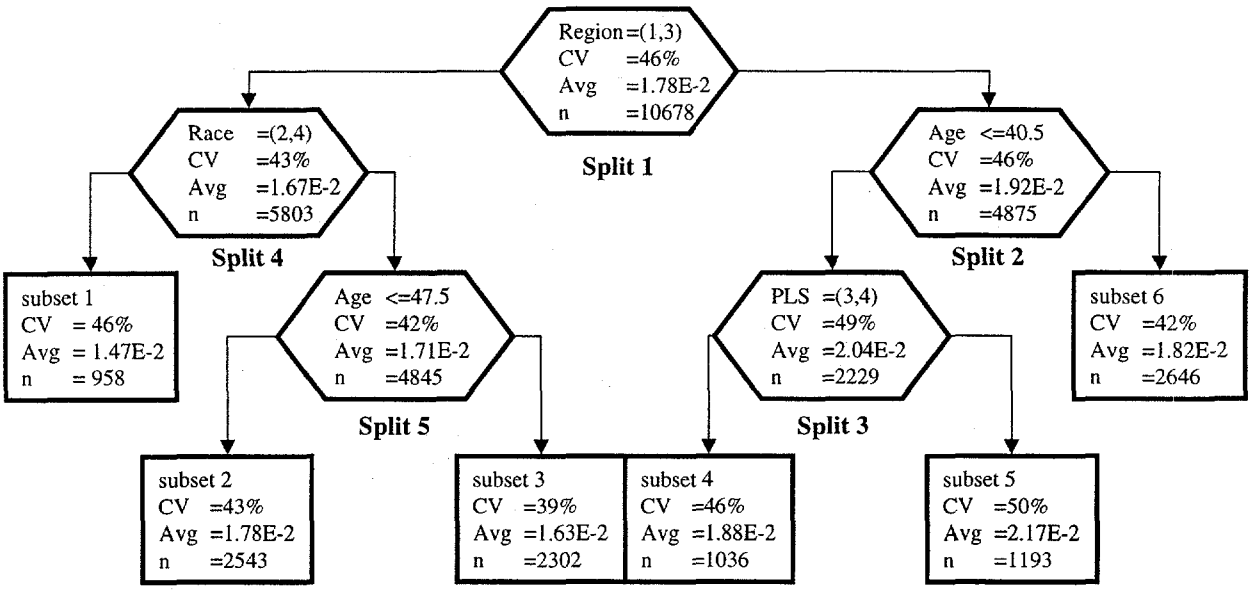
CART uses binary recursive partitioning to develop a classification or regression tree (Brieman et al., 1984). The CART software (Steinberg and Colla, 1997) provides a non-analytic, computationally intensive procedure that uses a set of rules for determining what factor and what value of that factor should be used to split the original data set into subsets. Each new subset is then analyzed and split until either the sample size reaches a lower limit or the cost (added complexity) of an additional split exceeds what would be gained in the form of reduced variance and spread between resulting subsets. Each split is chosen to maximize the statistical difference or separation between two new sub-groups created from the original data. The results of the CART analysis are concise, easy to understand, and are appropriate for use as a decision making tool. For an explanation of how a classification and regression tree is read, see Example 1. The technique was developed almost 20 years ago and has been applied in many fields, including engineering, medicine, public health and economics (Eisenberg et al., 1998; Eisenberg and McKone, 1998; Eisenberg et al, 1998; Pilote, et al., 1996; Tronstad, 1995; Spear et al., 1994; Spear et al., 1991). Detailed explanation of the methodology used in CART is available elsewhere (Breiman et al., 1984; Breiman, 1992)

CART is easy to use for identifying subsets or reducing variance in large complex samples. CART was developed to systematically decompose complex data sets into statistically different subgroups or samples. Subgroups that are *not* statistically different from the other members of the sample may still need to be treated differently based on political or policy reasons. However, after using CART to analyze the data, we can present a systematic and scientifically defensible process for decomposing the data and make a clear judgement about how appropriate or necessary the resulting subsets of the population are. Each time we split out a new subset of the population based on demographics, and construct a new distribution, we increase the complexity of the exposure analysis. This in turn increases both the likelihood of an error in the analysis and the complexity of the regulatory review process. Thus, it is beneficial to have the fewest justifiable number of classes for each exposure factor while still capturing the important differences within the population.

A two-stage procedure is used to identify the optimal number of demographic regions for each exposure factor. In the first stage of the procedure, CART is used to systematically split the data set into statistically different demographic regions and each split is ranked in order of importance. The most important split produces the greatest difference in the two resulting regions of the data. The second most important split in combination with the first split produces the greatest difference between three demographic regions and so on. The second stage of this

procedure, which comes after the distributional analysis, uses sensitivity analysis techniques to collapse the set of distributions for a given exposure factor into the fewest number of significantly different demographic regions (age, gender, race, etc.).

Example Box 1:



Legend

CV = percent coefficient of variation
Avg = average
n = sample size

Variables include:

Region (1=northeast, 2=midwest, 3=south and 4=west),
Race (1=white, 2=black, 3=Asian/Pacific, 4=native American and 5=other),
Age (reported in years) and
PLS (1=pregnant, 2=lactating, 3=pregnant and lactating, 4=not pregnant nor lactating and 5=not female 10-55).

This example demonstrates how CART can be used to systematically decompose a large complex data set into demographic regions of sampling space. The data is total water intake from the USDA's Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey (CSFII/DHKS) 1994-96. Total water intake includes both extrinsic and intrinsic water. Extrinsic water includes tap water and water added to food and beverage and intrinsic water includes all water naturally occurring in food. Total water intake is normalized to body weight ($\ell\ kg^{-1}\ d^{-1}$) prior to the analysis and only individuals older than 10 years of age are included. For individuals 10 and under, the single most important factor is age.

The first five splits in the data set are illustrated above. The output from the analysis is read as a binary decision tree. A logical statement is given at each decision point (hexagon). The

data for which the statement is true move to the left to form a new data subset. When the statement is false, the data moves to the right.

In the example, the first split occurs on "Region = 1,3" which means that all sample persons living in the northeast and south go to the subset to the left (average = 1.67E-2 ℓ kg^{-1} d^{-1}) and sample persons in the midwest and west are included in the other subset (average = 1.92E-2 ℓ kg^{-1} d^{-1}). The second split occurs on age < 40.5 for sample persons in the midwest and west. Sample persons 41 years and older go into subset 6 (average = 1.82E-2 ℓ kg^{-1} d^{-1}) and those 40 years and younger are again split on PLS. The split on PLS brings up a cautionary note. The selected variable seems to indicate that pregnancy/lactating status is the next most important variable. However, closer inspection shows that all non-pregnant females 40 years and younger living in the midwest and west are assigned to subset 4 (average = 1.88E-2 ℓ kg^{-1} d^{-1}) and pregnant or lactating women go to subset 5 along with all males (average = 2.17E-2 ℓ kg^{-1} d^{-1}). Pregnant and lactating women have water intake rates similar to that of men when normalized to body weight.

Next, the sample persons in the south and northeast are split on "Race = 2,4" moving all blacks and native Americans into subset 1 (average = 1.47E-2 ℓ kg^{-1} d^{-1}). The remainder of the data set is split again on "Age<=47.5" creating subset 2 (average = 1.78E-2 ℓ kg^{-1} d^{-1}) and subset 3 average = 1.63E-2 ℓ kg^{-1} d^{-1}).

The splitting order also indicates the relative importance of each variable in decomposing the data. For the example, for individuals older than 10 years of age, the region of the country is the most important variable followed in order by age, gender and race.

We used CART to analyze the original data sets and identify the importance of various demographic descriptors within the population. An advantage of the CART approach is that a wide range of factors can be included in the initial analysis whether these factors are continuous (age), binary, (gender) or categorical (race, region). Selection of the initial set of demographic test factors is limited only by the number of factors included in the data set.

Another important advantage of using CART to decompose the data into unique demographic regions of sample space is that we can avoid the thorny issue of developing distributions based on stratified surveys that require sample weighting in order to represent the population of interest. Typically, data from national surveys are weighted to correct for bias that may result from non-respondents or intentional over-sampling of specific members of the population (e.g., young children). Methods are readily available for calculating the population mean and percentiles from weighted data (Snedecor and Cochran, 1989, pp. 431-456) however, these methods are not easily applied to the development of distributions. Using the CART analysis, we can identify the subsets of the population that are different. The regions of the sample space that aren't different can then be combined without introducing bias into the

distribution. i.e., if the data does not indicate that men and women consume water at a different rate then weighting the sample will not influence the characteristics of the resulting distribution.

3.3 Constructing Distributions

Once the data are split into individual sample sets, the next step is to fit parametric distributions to the data for use in the probabilistic analyses. Methods for fitting parametric distributions to data are well established and statistical software is available with the capability of automating much of the work. Even though fitting parametric distributions to data using the method of moments, "goodness of fit" tests or maximum likelihood estimation (MLE) techniques is relatively easy with modern software, the challenge remains to determine when the model fits well enough or when one distribution is more appropriate than another. Analytical methods often lack the power to choose between two competing distributions (D'Agostino and Stephens, 1986). Therefore, we rely at least in part on subjective or visual methods for choosing the best parametric model for a given exposure factor.

The current study uses a combination of graphical and standard goodness-of-fit techniques to identify the best candidate distributions for each data set. By plotting the empirical cumulative distribution function (ECDF) of the data along with the cumulative distribution function (CDF) of several candidate parametric distributions (parameterized using MLE or other goodness-of-fit function) the quality of the fit can be visualized. We enhance this visualization process by plotting the relative deviation of the predicted percentiles for each CDF (residuals). This helps to highlight the regions of the ECDF where the parametric CDFs provide the best fit and where they provide the poorest fit. This feature is important because we are often more interested in a specific region of the distribution and the visualization process can facilitate judgements about which distribution works better in the important region even if the goodness of fit score indicates otherwise. A risk assessment is often focused on the upper region of the distribution of risk. How the exposure factor influences this region depends on where the factor fits into the exposure algorithm (see for example Eq. 3). For example, *Body Weight* is used in the denominator of the exposure equation and as a result, the lower region of the BW distribution produces the higher estimate of exposure. By contrast, the *Intake rate* and the *Exposure Frequency* are both in the numerator so that the upper region of the distribution for these factors produces the higher estimate of exposure.

In finding the best distribution, we start with a standard set of parametric distributions. Each exposure factor has a different set of candidate distributions. These distributions are selected for their simplicity and their theoretical representation of the particular data type. For example, when exposure frequency is specified as a "fraction of the day" it results in a distribution that is bounded by 0 and 1. Candidates for this case include log uniform, log triangular, Beta or truncated continuous distributions. We limit the use of truncated distributions

whenever possible because the process of truncation fundamentally changes both the distribution and the procedures used to parameterize the distribution.

3.3.1 Statistical Methods (Parametric)

Well established methods are currently available for fitting parametric distribution to data. In addition to extensive reviews and discussion of the available methods, [D'Agostino and Stephens, 1986, Cullen and Frey, 1999] there are several software packages that do a good job of implementing the methods. Therefore, we do not go into great detail about the standard methods. Most of the distributional analyses in this study were performed in a spread sheet program. Some of the initial visualization of the data sets were performed using the Minitab data analysis software or using C++ subroutines developed for a specific task (see Section 5.1.3).

3.3.2 Model-Free and Graphical Methods

The simplest graphical methods for distributional analysis are the direct comparison of distribution functions plotted on the same chart. This comparison can be facilitated using a variety of plots such as the P-P, Q-Q plot or some form of linear transformation using standardized Z-scores (D'Agostino and Stephens, 1986). We prefer a simple plot of the residuals between the ECDF and the parametric CDF as given in Equation 3.1 and illustrated in Figure 3.1:

$$R_i = (P_i - P_{ik}) \tag{3.1}$$

where: R_i = residual between the i^{th} percentile of the ECDF and the i^{th} percentile of k^{th} parametric CDF

P_i = i^{th} percentile of the ECDF

P_{ik} = i^{th} percentile of the k^{th} parametric CDF

Several residual plots can be included on offset horizontal axis on a single chart along with the associated distributions making it easy to locate the regions a distribution where the model provides the best fit. Figure 3.1 demonstrates the use of the residual where raw data was drawn directly from a standard normal distribution (mean=0, s=1) and the residual plot calculated as the difference between the percentiles of the raw data and those of the parent distribution. The figure demonstrates that even when the data is drawn from a known distribution, a certain degree of scatter can be expected due to the random nature of the sampling process. As the sample size gets smaller, the range of scatter in the residuals increases. Using residuals to visualize fit is useful in that it provides information on the density of data in the different regions of the curve as well as an indication of bias or trend in the parametric model.

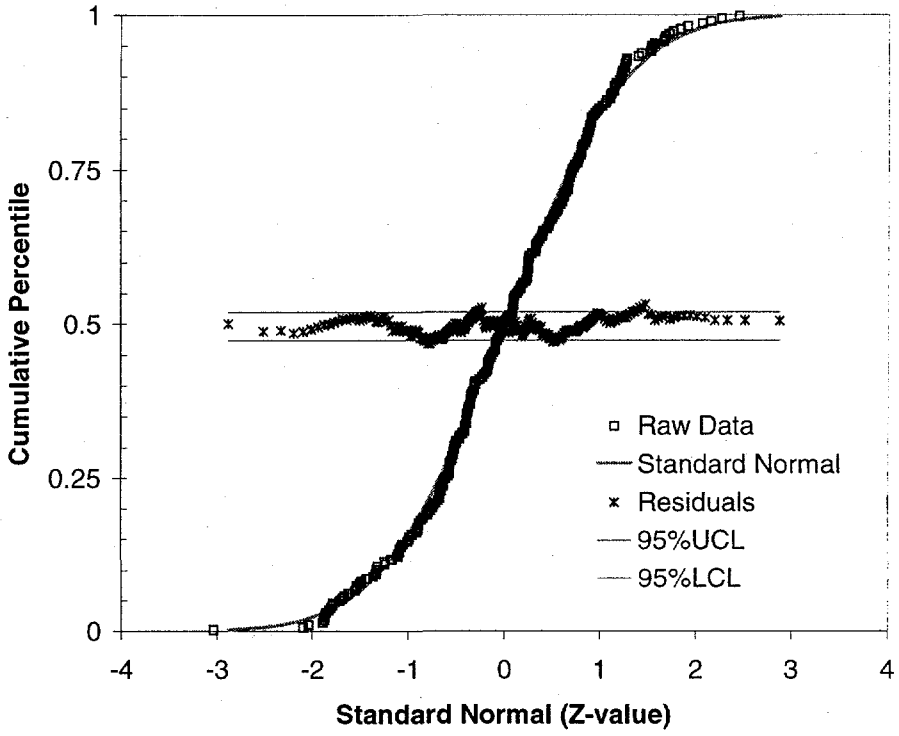


Figure 3.1: Illustration of the residuals for visualizing the fit of a parametric distribution to an empirical distribution. The raw data was drawn directly from a standard normal distribution (mean=0, s=1). The residuals not only give an indication of where the parametric distribution fits best, it shows the density of the data at the tails and any bias or trend in the model.

The 95% confidence bounds in Figure 3 were developed by repeatedly generating random samples from know distributions and calculating the resulting residuals. This was done for a wide range of sample sizes and standard deviations (or scale values in the case of Beta distribution). Confidence bounds in the vertical direction (uncertainty about the percentiles rather than un(uncertainty about the percentiles rather than uncertainty about the quantiles) were found to be insensitive to changes in the spread of the data. For sample sizes greater than 50, the confidence bounds on the residuals are inversely proportional to the square root of the sample size. The relationship between sample size and the 95% confidence bound is given in Table 3.1 for three distributions.

Table 3.1: Confidence Interval Estimation Equations

Distribution type	Confidence interval	r^2	comments
Normal	CI=0.837/sqrt(n)	0.9997	Linear for n>=50
Lognormal	CI=0.815/sqrt(n)	0.9872	Linear for n>=50
Beta	CI=0.942/sqrt(n)	0.9966	Linear for n>=30

A sample size of 100 drawn from a lognormal and Beta distribution would yield 8% and 9% confidence interval, respectively. Given the similarity across differing parametric distributions, we settle on a single average predictor for the confidence interval (*CI*) for all distributions as given in Equation 3.2:

$$CI = \frac{0.86}{\sqrt{n}} \tag{3.2}$$

Although the confidence interval in Figure 3.1 is illustrated as parallel lines, the actual confidence interval about the percentiles is greatest at the mean and less in the tails. For convenience, parallel lines are used to approximate the confidence interval for visualizing model fit.

Model-free methods are also available for visualizing data and investigating the underlying composition of the data set – whether the distribution is a composite or mixture model and what parametric distribution or combination of distributions best represent the data (Tarter, 1991). The model-free methods show promise as a tool for learning more about the underlying shape of distributions but more work is needed to determine just how useful they might be.

3.3 Stochastic Analysis and Distribution Class Reduction

Once each class is assigned an appropriate distribution, a next step will be to collapse the set of distributions for each exposure factor into the minimum set of stochastically important distribution. Although we introduce this step here, the actual work and results will be provided in a separate report.

A Monte Carlo sensitivity analysis technique will be used with each distribution to determine the exposure factor's influence on the calculation of dose or risk. The central tendency

and variance (uncertainty and variability) in each exposure factor will be combined through a dose equation in a predictable multiplicative manner (Rai and Krewski, 1998). Analytically, the contribution to variance in the dose estimate is easily calculated. However, when considering the overall distribution (including the tails) it is often easier to simply run a Monte Carlo analysis and generate an outcome distribution along with correlation between inputs and outputs to assess the importance of a given model input. We propose to use the latter approach, along with Equation 3.3, to test the individual exposure factor classes and determine which demographic exposure factor subsets can be recombined without significant loss of information.

$$ADD_i = \frac{C_i \times IR_i \times EF \times ED}{BW \times AT} \tag{3.3}$$

where: ADD_i = average daily dose received through ingestion of the i^{th} contaminated media (mg Kg^{-1} d^{-1}),

C_i = contaminant concentration in media i, (mg m^{-3}),

IR_i = intake rate of the i^{th} media, ($m^3 d^{-1}$),

EF = exposure frequency, (unitless fraction)

ED = exposure duration, (y)

BW = body weight, (Kg) and

AT = averaging time, (y)

A spread sheet program will be developed and used to systematically test each combination of input distributions in a Monte Carlo analysis. By comparing the outcome distributions we can collapse the set of input distributions to the appropriate number of unique demographic classes for each exposure factor. When testing a set of distributions for an exposure factor, all other inputs to the exposure equation will be assigned their most precise parametric model from their respective distribution sets. For example, when testing the water intake distributions in Example 1, all inputs are assigned distributions that minimize the level of outcome variance. Thus, if no difference in the distributions of exposure model outcomes is detected for two adjacent exposure factor distributions (subsets 2 & 3 or subsets 4 & 5 in Example 1) then those subsets can be collapsed or recombined into a single subset and a new model fit to the data. The analysis should include variance in the concentration term (assumed to be log normal with a coefficient of variation of 10%) but uncertainty and variability in the toxicity data will be excluded. As a result, the stochastic analysis of the distributions is expected to be conservative in that additional variance will be introduced into the system by the toxicity data and as a result the variance threshold (the level below which one cannot tell the difference between two distributions) will increase.

3.3 Scoring the PDFs

The main contribution of this report is the development of a simple method for scoring the quality of distributions in the context of the cohort/population to which the distribution is applied. After the final set of exposure factor data sets are identified along with their respective distributions, we introduce a scoring system based on a combination of quantitative and qualitative information for each distribution. The quantitative information includes items such as sample size, confidence intervals about the distribution, sensitivity of the exposure equation to the particular exposure factor and graphical/analytical measures of how well the recommended PDF represents the available data. The qualitative information will include an assessment of how well the sample survey captures the demographics of the population and how well the sampled data represents the particular exposure factor.

The challenge in developing high quality, reliable and relevant distributions is that we often have an incomplete picture of the population (independent studies, small sample sizes). Without a statistically representative sample of the population, it is difficult to know the amount of variation between and among subsets of the population. This limits our ability to know *a-priori*, how well a distribution for one demographic part of the population will represent another without some minimal sampling of the new population. For example, we can use national census data to construct a single distribution of water consumption that will encompass all members of the population in all regions of the country with a high degree of certainty (assuming the data from the census is representative). Statistically different subsets of the population can then be identified and the initial PDF can be decomposed into a mixture of class-specific PDFs each representing a unique subset of the population (gender, age, ethnicity and region) as described in section 3.2. We can continue to decompose the sample as long as the variance reduction, sample size and distance between the new distributions warrant a separate model.

However, if we were to start with a data set consisting of only a subset of the population, we may be able to construct a highly representative PDF for the existing data set but the distribution will not necessarily encompass the entire population or other demographic regions of the population. Expanding the distribution to include other members of the population will rely on qualitative information and insight gained from the better-characterized exposure factors. In this case, the size of the confidence bands about the distribution will be influenced by quantitative information such as sample size and qualitative information such as how representative the data set is and uncertainty about the selected parametric model.

The scoring system introduced in this report is a questionnaire designed to combine quantitative and qualitative information about the data and models into a single scenario-specific measure for the quality of a given parametric model (or other form of distribution). Although the final scores fall on a continuum from **not applicable** to **highly recommended,** the continuum is partitioned into four basic regions defined as Highly recommended for use (H), Medium (M),

Low (L) and Not Applicable for use (NA). The questions are designed to elicit information about:

1) The quantity of data used to construct distributions,
2) Relevance of the data (actual measurement, self-reported or surrogate value),
3) Analytical goodness of fit for standard distributions,
4) Theoretical basis for standard distributions,
5) Visual performance of the model across the range of data including the percentiles of greatest interest to the particular analysis objective,
6) Extent to which variability and uncertainty can be represented, is the amount of measurement or reporting error known, and
7) Ability of the recommended distribution to forecast samples from independent but related surveys and/or data sets.

Although the final form of the questionnaire and scoring system should come through peer reviewed literature and/or from extensive open debate among experts from a wide range of disciplines, an initial format is developed from the above list of criteria and presented in Example 2. At this stage, the lines between each score are assigned somewhat arbitrarily. As we gain experience with a variety of distributions and data sets, the lines separating HA, M, L and NA will likely converge on the most effective location. The questions may also evolve as we gain additional experience and insight.

Some of the criteria in the questionnaire are quantitative where the value given is dependent on an actual measurement of sample size or fit. For other criteria such as data quality the score falls on a continuum from very poor to very good. To assign a score to these criteria, the user must become familiar with the data. The more intimate a person is with a given data set, the more qualified that person is to judge the quality of the data for a given task. Information on how the data was collected and the precision of the measuring device used to collect the data should also be considered when judging data quality. There are several sources of guidance for judging data quality and the reader is referred to these papers and books for further discussion (Cullen and Frey 1999, page 162, Thompson, 1999). Overall, in order to judge the quality of a distribution, it is critically important that the user have a clear and complete understanding of the data used to develop the distribution in question, the procedure used to construct the distribution and the population in the analysis objective. The questionnaire is designed to help lead the user towards the necessary level of understanding.

Example Box 2

For each of the following criteria, enter a number from 0 to 3 in the box to the right. For some questions the values will be arrived at quantitatively and for others the values will be assigned on a low to high scale. After filling in each of the sections, add up each score and refer to the bar at the bottom of the page to locate the score.

Sample size

For ($n \leq 10$); enter 0
For ($10 < n \leq 50$); enter 1
For ($50 < n \leq 250$); enter 2
For ($250 < n$); enter 3 _____

Data relevance

For irrelevant data; enter 0
For surrogate values; enter 1
For self-reported values; enter 2
For actual measurements; enter 3 _____

Data Quality

Score data quality from 0-3 (low to high) _____

Theoretical basis for distribution

Score theoretical basis from 0 to 3 (low to high) _____

Analytical goodness of fit

For KS or AD in 50%; enter 1
For KS or AD in 75%; enter 2
For KS or AD in 95%; enter 3 _____

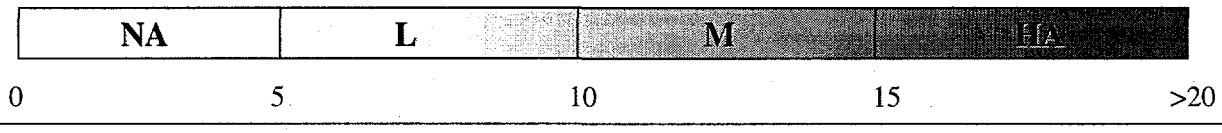
Visual performance

Poor fit across range; enter 0
High scatter but low bias; enter 1
Low scatter low bias in region of interest; enter 2
Low scatter and bias across range of data; enter 3 _____

Model performance in cross-validation

Enter 0 if no cross-validation has been performed
Score model performance from 1-3 (poor to good) for each independent cross-validation experiment _____

Add the values in the right hand column and locate the score on the following bar.



3.4 Method summary

The overall method, starting from the point where the appropriate data set has been identified and carrying through the robustness score is summarized in Table 3.2 (preprocessing phase) and 3.3 (data processing phase).

Table 3.2: Steps in the preprocessing phase of the project

Pre-processing phase
1. Acquire and install data base and necessary software
2. Identify factors of interest that are available (both dependent and independent variables)
3. Extract data files and save as an Excel file(s)
4. Combine files that were extracted separately from the same database (body weight and water intake for example). Use the sample identification number when merging files.
5. Remove files with incomplete data. Keep record of files that are removed to adjust the weighting factor if necessary.
6. Convert units (oz. – grams, lb – kg, cups – grams, months to fraction of year …).
7. Combine variables where feasible, (convert months to season) Age typically has column for years and for months – combine in a single column for years by converting the months into fraction of a year (months/12).
8. Convert files to format for analysis in CART
9. Go to Data Processing Phase

Table 3.3: Summary of procedure for processing data

Data Processing Phase
1. Construct histograms and baseline distribution for complete data set (using Minitab, SAS or Excel) prior to decomposition.
2. Perform CART analysis to identify subsets of the data
3. Save the subsets as independent files and construct empirical distributions and histograms for each
4. Use T-test and/or Anderson Darling test to assess differences and confirm extent of curve decomposition
5. Save final data subsets as Excel files
6. Construct distributions for each subset identified in CART analysis
7. Run Monte Carlo analysis and collapsed classes where warranted into more general distributions for each exposure factor
8. Determine confidence bounds and two dimensional distributions for each final distribution if sample size is less than 100
9. Calculate confidence bounds for each distribution (if significant)
10. Extract uncertainty (normal distribution about mean based on sample size, Chi-square distribution about standard deviation based on sample size) where uncertainty contributes significantly to variance
11. Test power of parametric model to predict independent or randomly drawn data sets. (Is residual error within the bounds of uncertainty?)
12. Combine quantitative and qualitative information to calculate final robustness score
13. Compile information and score the final distributions

4.0 Development of PDFs for Body Weight

In this section we provide details on the development and analysis of PDFs for body weight. Body weight is one of the most extensive and representative data sets of all the exposure factors. Measured and self-reported body weights are provided along with demographics in several nationally representative surveys. In addition, anthropometric data for children have been extensively reviewed and include studied using both longitudinal and cross sectional analyses. Results from these studies have been used to develop standard growth charts for use by medical practitioners when assessing the growth of children during physical examinations (Tanner et al., 1965; Hamill et al., 1977; Hamill et al., 1979). Updated growth charts should be available sometime in 1999 (http://www.cdc.gov/nchswww/about/major/nhanes/hanesrev.htm).

Age and gender have been used to characterize the distribution of body weight in the population (Burmaster and Crouch, 1997, USEPA, 1990; USEPA, 1997). By referring to the original data we can identify the optimal demographic classification for this exposure factor. When fitting distributions to the BW data it is important to consider what part of the distribution is most influential to the dose calculation. BW is in the denominator of the calculation and as a result, small values of BW (lower tail of the distribution) produces the largest estimate of dose and subsequently the highest risk. Thus, for this exposure factor, we are most interested in fitting the lower region of the data. However, it should be noted that all of the data should be considered as relevant and we only use the lower tail when trying to decide between two parametric distributions that each score well with standard goodness of fit tests.

4.1 Sources of Data

In addition to numerous smaller studies, two extensive nationally representative surveys were available for assessing BW. These include the National Health and Nutrition Examination Survey, III 1988-94 (DHHS, 1997 revised; NHANES III) and the 1994-96 Continuing Survey of Food Intake by Individuals (CSFII). Both surveys provide complete representation of the population on both a national and regional level and include several potentially important variables. Even though both surveys include information on body weight and water intake the water intake information in the CSFII survey is more quantitative. For this reason, we chose to use the CSFII database for the initial development and save the NHANESIII database for cross validation and future work on robustness scoring.

It was learned late in the study that the two surveys also differ in that the CSFII has self-reported body weights while the NHANESIII includes actual measurements. Even though the actual measurements are better than self-reported values, we continued to use the CSFII data

because the benefit of having water intake data along with body weight exceeded the potential cost of having self-reported body weight.

The CSFII was conducted by the Food Surveys Research Group, Beltsville Human Nutrition Research Center, Agricultural Research Service (USDA, 1998). The survey includes a nationally represented population of noninstitutionalized (non-military, and not living in group quarters) households, excluding the homeless (and reservations). Low-income housing units were over-sampled but weighting factors were provided that correct the sample composition to reflect the composition of the US population based on the 1990 Census. For more information on the CSFII/DHKS 1994-96 Survey Methodology, see Tipett and Cypel (eds.) 1997 on Disk 1 in \dor9496\dor9496.pdf. Two 24-hour recall records were used as the collection method for dietary data.

The self-reported body weight [lbs] is given for each "Sample Person" in the particular record type (rt25.dat file on CD-ROM) for 16,103 individuals. Also provided in the data set are:

- age in yrs (0-90 yrs) and in months (if <1 yr)
- breast feeding status
- pregnant/lactating status
- race (white, black, Asian/Pacific Islander, American Indian/Alaskan native, or other race)
- origin (Mexican, Puerto Rican, Cuban, or other Hispanic)
- percentage of poverty level (the household income for the previous calendar year expressed as a percentage of the Federal poverty thresholds (Baugher and Lamison-White, 1996) adjusted for inflation)
- region

  Northeast = Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont;

  Midwest = Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin,

  South = Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia; and

West = Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming), gender, urbanization (MSA, central city, MSA, outside central city, or Non-MSA),

- and year of survey.

Further information can be found at
http://www.barc.usda.gov/bhnrc/foodsurvey/home.htm

4.2 Terminology and Definitions

- Body weight: Total mass of individual
- Longitudinal analysis: measurements collected from a single individual over time.
- Cross-sectional analysis: measurements collected across a population (both the NHANESIII and the CSFII provide cross-sectional analyses)

4.3 Data Classification and Distribution Analysis

The original data set from record type 25 of the CSFII was modified prior to analysis. The modification included the removal of all pregnant, lactating or pregnant and lactating women from file because of the strong dependence of BW on term of pregnancy. The two columns reporting age in years and age in months (when years < 0) were combined into a single column of age in years (fraction of year used when age<0). The pctpov (percent of poverty) variable was converted to categorical data such that "under poverty line"=1, "100-200%of poverty" = 2 and "greater than 200 % of poverty" = 3 (resulting frequency: 1=2673, 2=3667 and 3=9159). The independent demographic variables that were included in the CART analysis were age, gender, race, ethnicity, region, urban (whether individual lived in rural, urban or metropolitan area) and percent of poverty. A total of 15502 sample persons were included in the final data set for BW analysis.

The CART analysis was set up using the default options. The analysis was set for regression tree with v-fold cross validation (n=10) and the minimum cost tree was generated using the least squares method. The results are presented for ages 12 years and above in the tree diagram in Figure 4.1. Ages less than 12 were not included in the figure to reduce the complexity of the figure. Information not included in Figure 4.1 is split 1 (at age <= 11.5), split 3 (age <= 6.5), split 7 (age <= 2.5) and split 10 (age <= 9.5). These splits produce sub-regions of the population grouped by age from 0 to 2 years, 3 to 6 years, 7 to9 years and 10 to 11 years. There is no measurable difference in gender or race below age 12. For an explanation on how to

read the regression tree output, see Example 1 in section 3. The compositions of the demographic sub-regions of the data are summarized in Table 4.1.

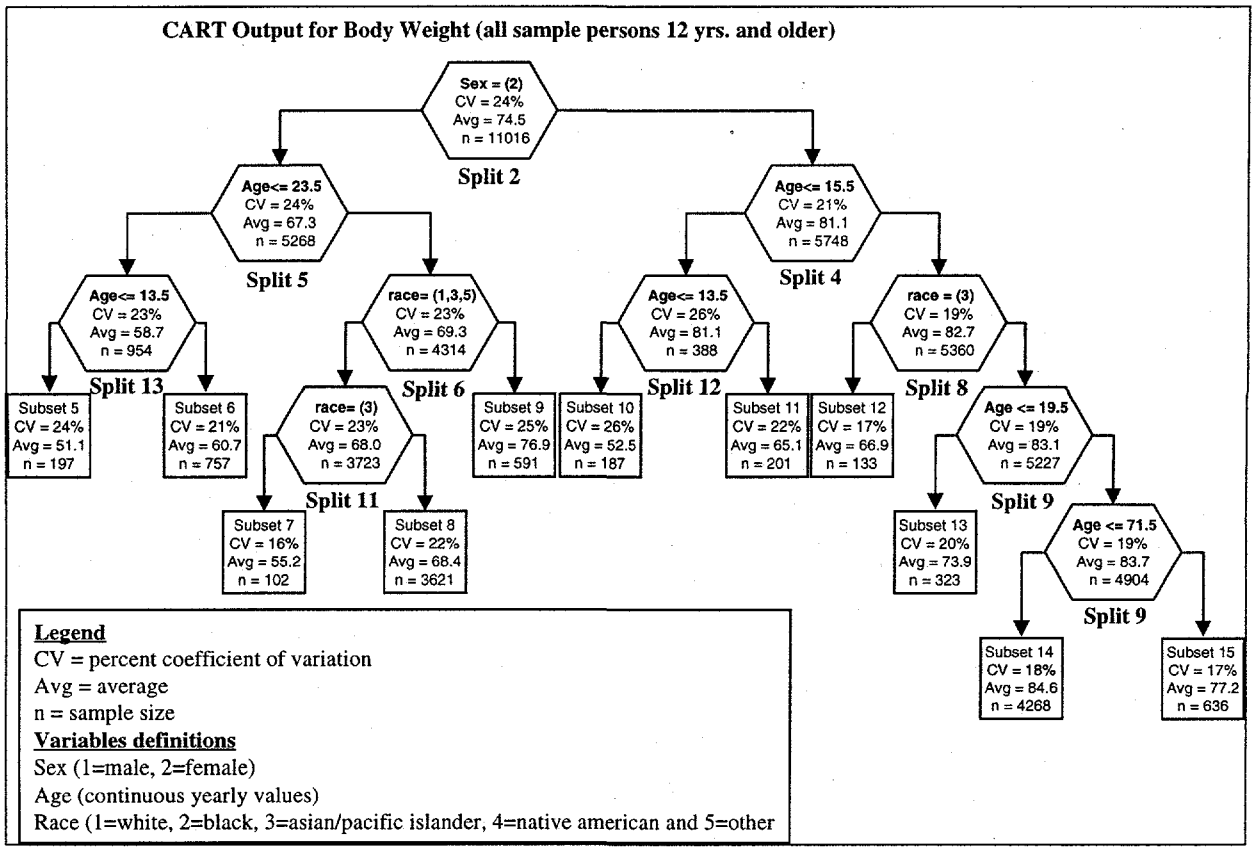


Figure 4.1: Classification and regression tree showing the decomposition of the original data set for body weight into demographic sub-regions. The tree begins at the second data split. The first split was on age<=11.5 years. As a result, the data in figure 4.1 are only for ages 12 years and older. See text for further explanation.

Table 4.1: Composition of final BW nodes

Sub-region	Characteristics	n	Ave.	CV
Root	Full data set for U.S. population	15502	59	50%
1(2)	All persons 1 and 2 years of age	1718	12	29%
2(1)	All persons 3 to 6 years of age	1610	19	24%
3(4)	All persons 7 to 9 years of age	672	30	27%
4(3)	All persons 10 and 11 years of age	486	40	26%
5(6)	Females 12 and 13 years of age	197	51	24%
6(5)	Females 14 to 23 years of age	757	61	21%
7(8)	Asian/Pacific females 24 years and older	102	55	16%
8(7)	Caucasian females 24 years and older	3621	68	22%
9	Black and American Indian females 24 years and older	591	77	25%
10(11)	Males 12 and 13 years of age	187	53	26%
11(10)	Males 14 and 15 years of age	201	65	22%
12	Asian/Pacific males 16 years and older	133	67	17%
13(14,15)	Males 15 to 19 years of age (non-Asian/Pacific males)	323	74	20%
14(13,15)	Males 20 to 71 years of age (non-Asian/Pacific males)	4268	85	18%
15(13,14)	Males 72 years and older (non-Asian/Pacific males)	636	77	17%

The sub-region number refers to the terminal region in the CART tree illustrated in figure 4.1 Terminal node 1-4 are not included in Figure 4.1. The number in parenthesis indicates the "sister" node. Sister nodes are generated by splitting a single node and as such can be recombined if it is determined that the difference between the two nodes is not significant. The sub-region number does not reflect the order of importance. For data splitting order, refer to figure 4.1.

It is not surprising that body weight is strongly dependent upon age from birth through the teen years. The results in Table 4.1 show a strong dependence on BW for all children under 12 years (regions 1-4). For adolescents and adults (age 12 and up) gender becomes an important variable. Females are separated by age from 12 to 24 (regions 5-6) and by race for women 24 years and older (regions 7,8 and 9). Men are subdivided by age from 12 to 19 years (nodes 10, 11 and 13) and above 72 (regions 14 and 15) except that Asians and Pacific Islanders are split out of the data set for men older than 15 years (regions 12). It is interesting to note the significant difference in body weight for Asian/Pacific Islanders for adult men and women. The average body weight for this demographic sub-population of adult women is over 10 Kg less than the general population and 20 Kg less than that of Black and American Indian women. We also note that the average weight for adult males excluding Asian/Pacific Islanders is 85 Kg.

4.4 Presentation of Distributions

The output from the CART analysis was used to construct individual data sets for each node in Table 4.1. ECDFs for each resulting sample are illustrated in Figures 4.2 through 4.4. Figure 4.2 includes all children younger than 11.5 years (race, gender, region, ...) subdivided into 4 age categories. Figure 4.3 includes all females over 12 years of age and Figure 4.4

includes all males over 12 years of age. The figures further illustrate the separation between the different categories identified by the CART analysis.

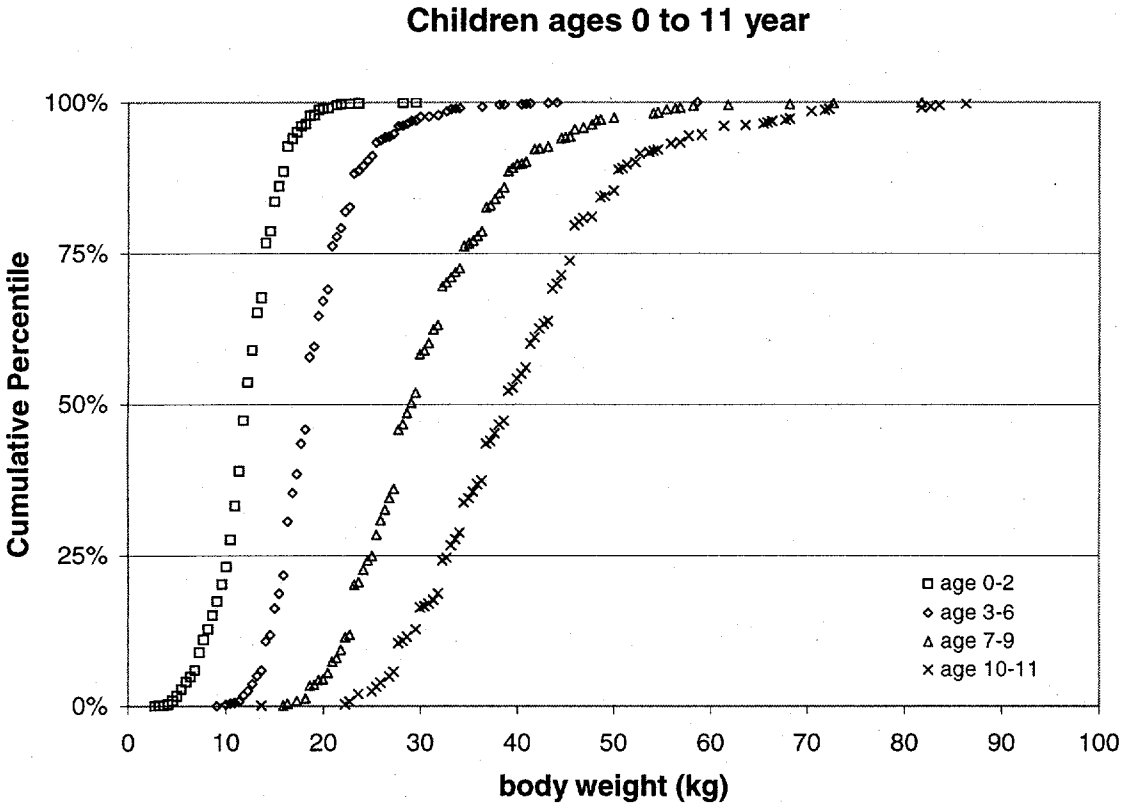


Figure 4.2: Age dependent empirical cumulative distribution functions for the body weight (kg) of all children under 12 years of age separated into four age groups. There was no significant difference in gender, race or other measured demographic variable for children under 12.

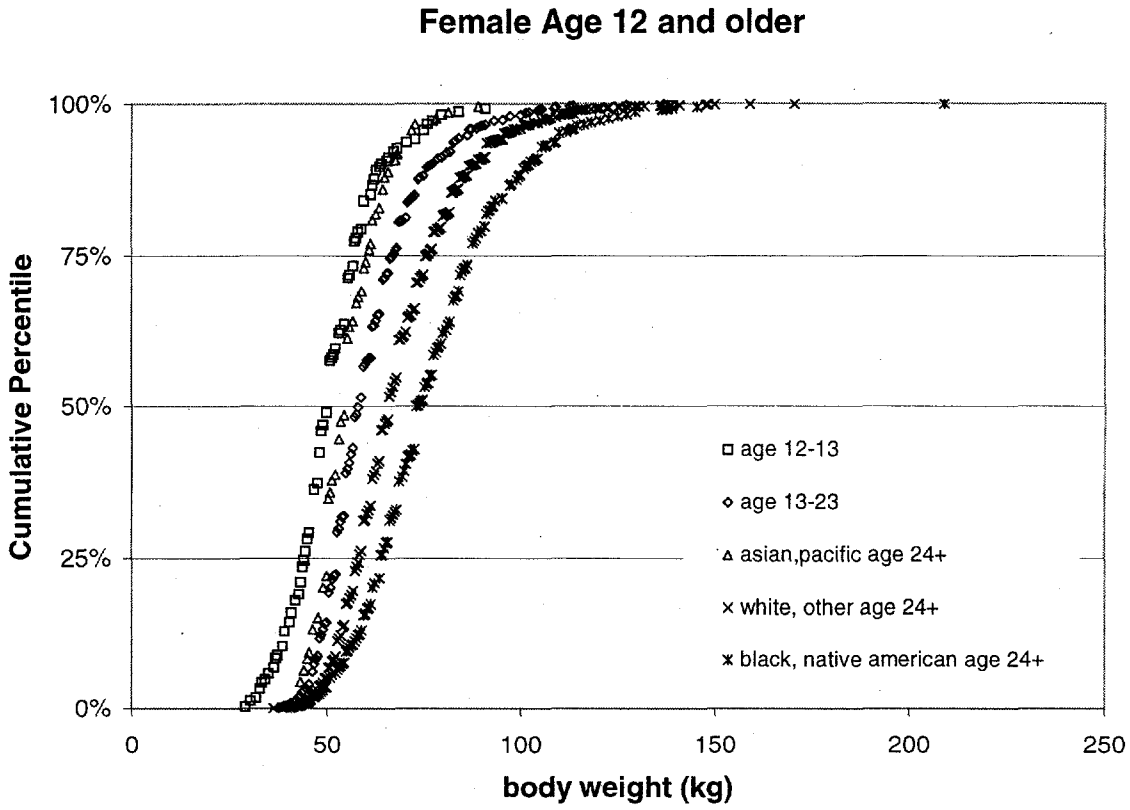


Figure 4.3: Empirical cumulative distribution functions for the body weight (kg) of females 12 years and older. In addition to the obvious age dependence, body weight for females is also dependent upon race. The "Black, Native American age 24+" category had a strong outlier but we were not able to justify removing the value from the data set.

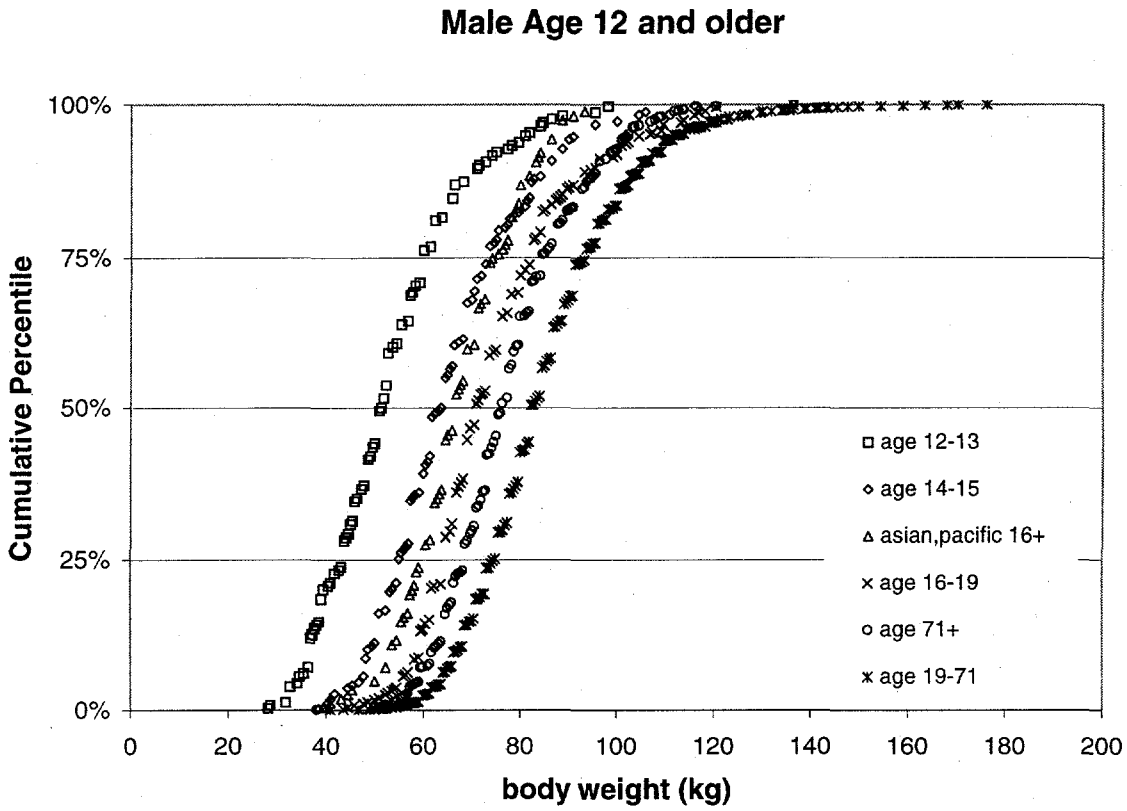


Figure 4.4: Empirical cumulative distribution functions for the body weight (kg) of males 12 years and older. The body weight of males is also somewhat dependent on race where the body weight of adolescent and adult Asian/Pacific Islander males is significantly different than that of the other members of the population.

The data used to construct the ECDFs in Figures 4.2 to 4.5 are placed into individual data sets and the method described in Section 3 is used to select the best parametric distribution for each set. Figure 4.6 shows a typical illustration of the results of the distributional analysis. Although both parametric distributions in Figure 4.6 (normal and logistic) do a good job describing the data, the logistic is selected because it likely arises from the rapid growth that occurs during the first two years of life (Johnson, 1995). A second example illustrating the body weight of Black and American Indian adult women is given in Figure 4.7. In this case both the lognormal and the extreme value distributions adequately fit the data. Typically if more than one distribution provides an adequate fit to the data, we selected the simpler or more common distribution for use. The three-parameter Gamma distribution and the three-parameter lognormal distribution fit the data better than the two-parameter distributions. However, the increased complexity of the distribution was not warranted given the inherent noise in the data. The results for the remaining demographic subsets of the data are illustrated in Figures 4.8 to 4.20 and all of the categories are summarized in Table 4.2.

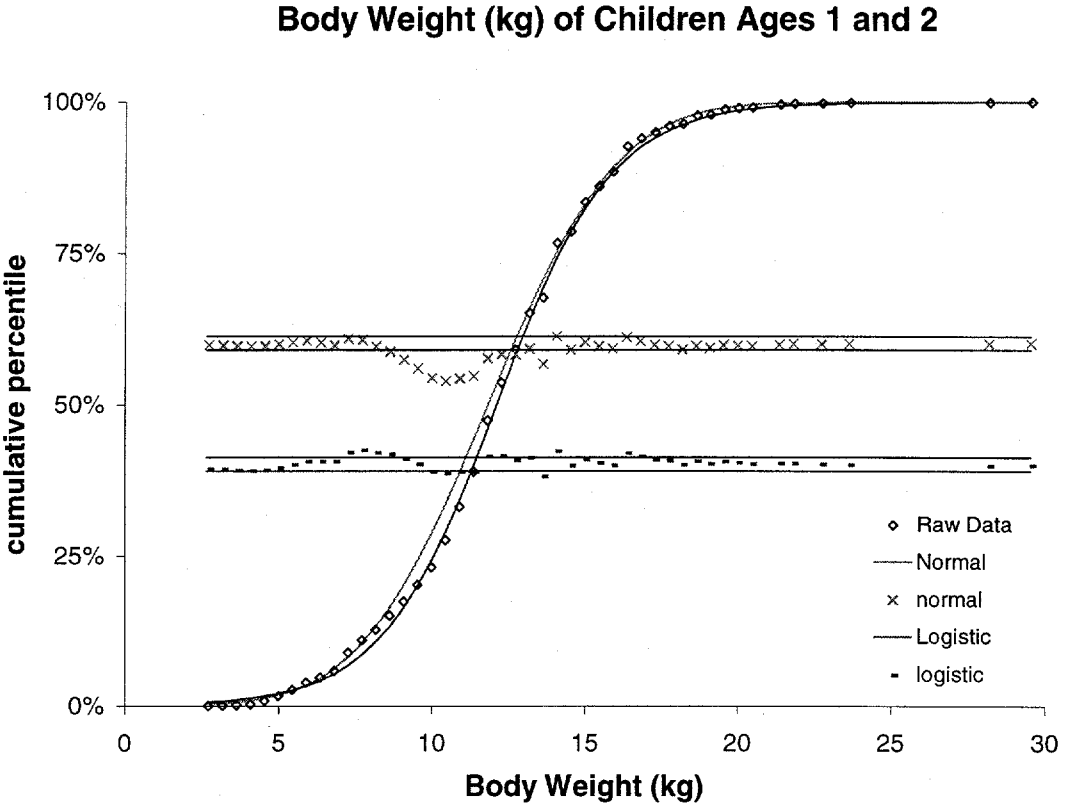


Figure 4.6: Body weight distribution for children ages 1 and 2. This is a typical overlay of the parametric distributions used to model body weight. It shows that each of the distributions do a good job mapping the actual data. The logistic distribution is selected here because it fits well in the tails of the data. The logistic probably arises from the rapid growth during the first two years. The straight lines are the 95% confidence interval for the residuals.

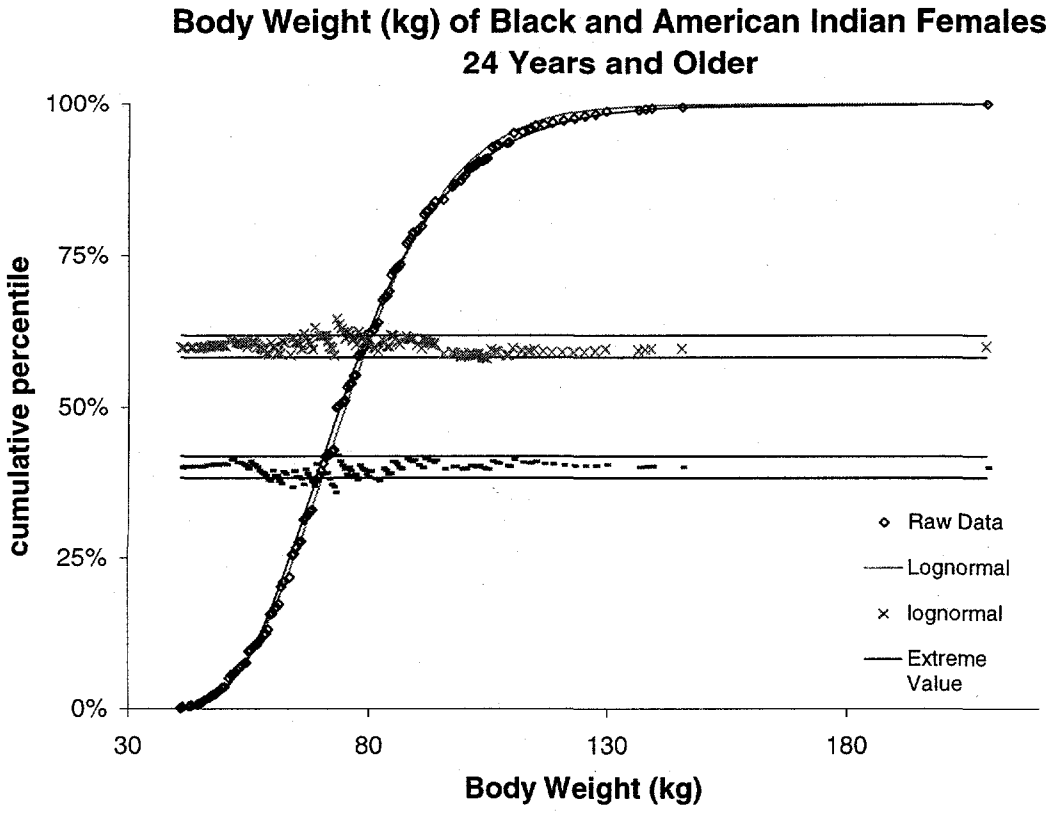


Figure 4.7: Body weight of Black and American Indian adult females showing the fit of the lognormal and extreme value distributions. Both distributions do a good job of fitting the data. There is an apparent outlier in the data (upper tail) but no justification for removing the point could be made.

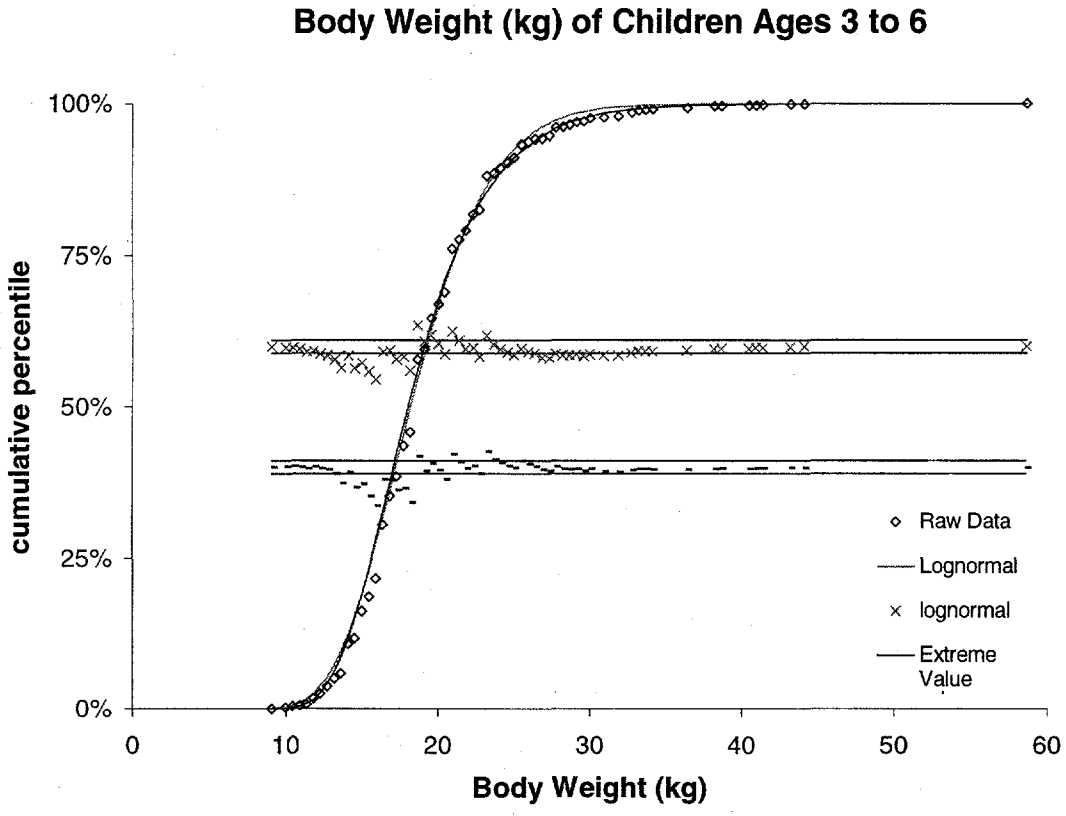


Figure 4.8: Body weight of children ages 3 to 6 years showing the fit of the lognormal and extreme value distributions.

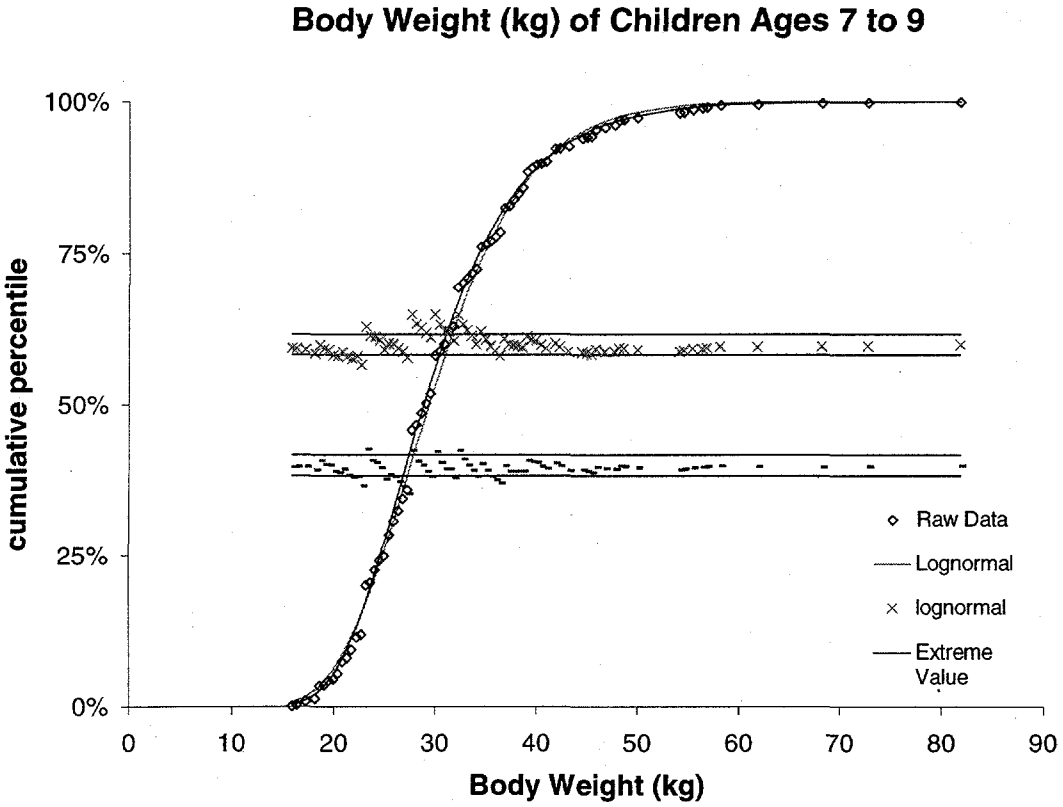


Figure 4.9: Body weight of children ages 7 to 9 years showing the performance of the lognormal and extreme value models.

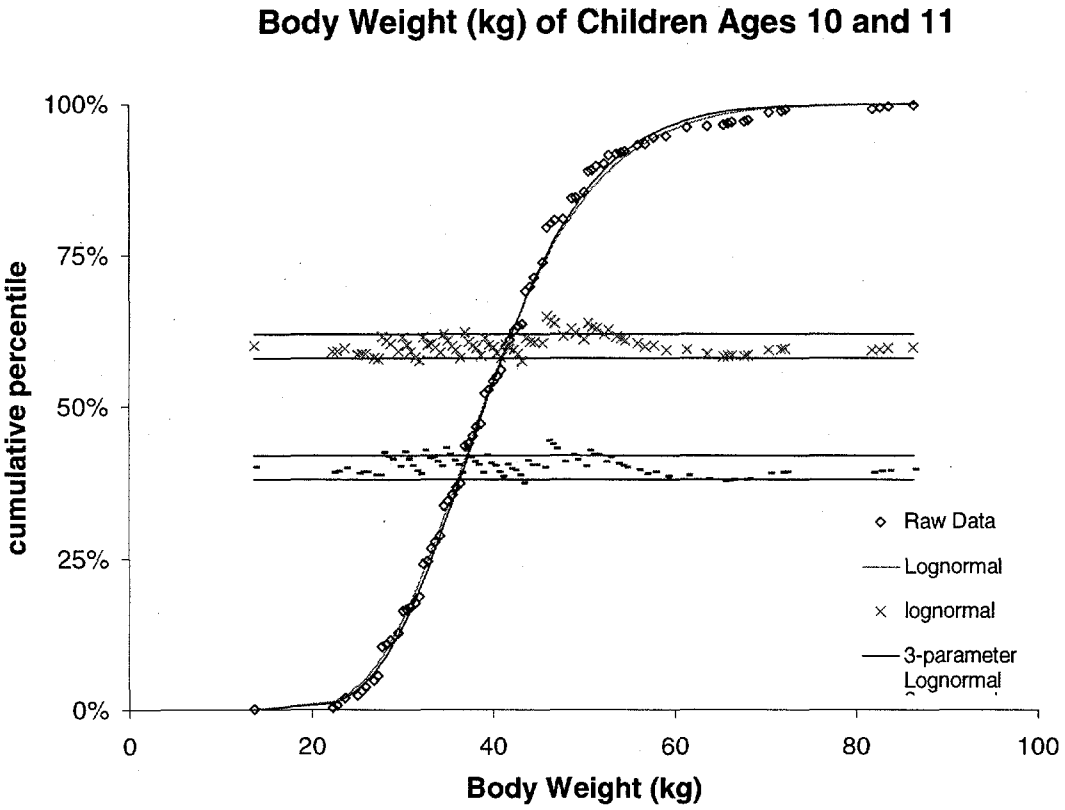


Figure 4.10: Body weight of children ages 10 and 11 years showing the performance of the lognormal and three-parameter lognormal models. The step pattern in the raw data is an artifact of the unit conversion and the self-reported values (more values were reported in units of five pounds indicating a rounding tendancy in self-reported body weights).

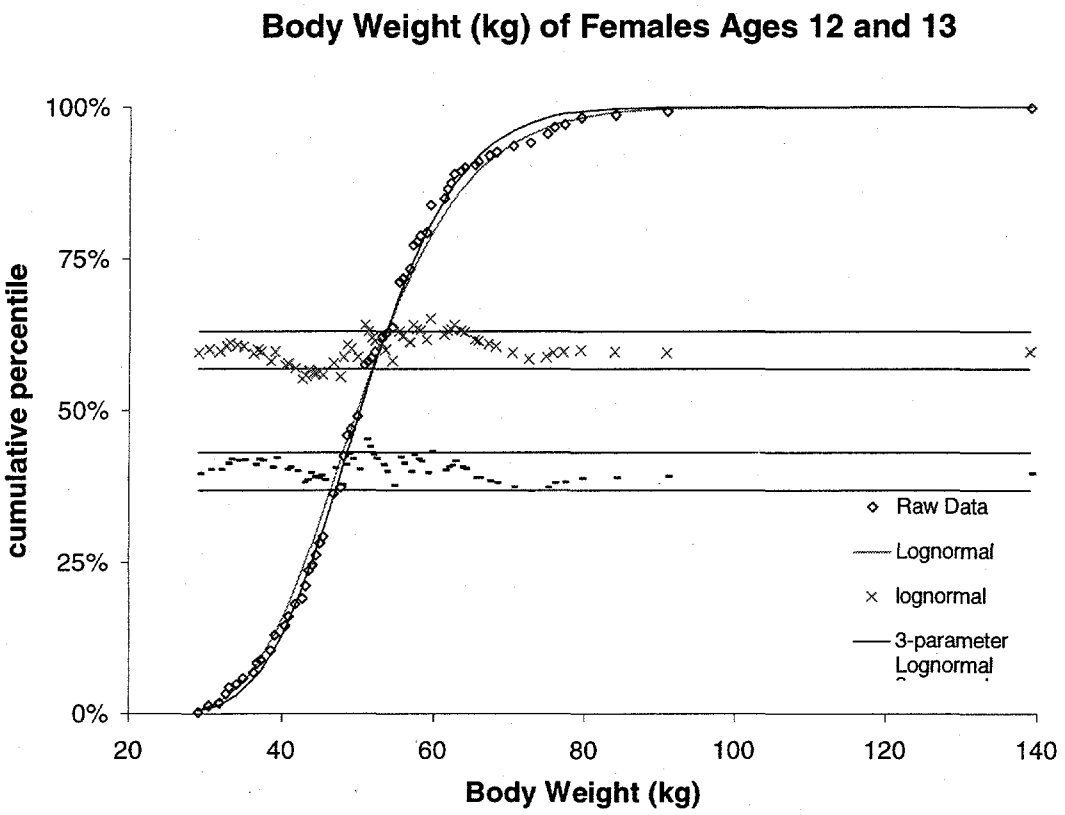


Figure 4.11: Body weight of females ages 12 and 13 years showing the performance of the lognormal and three-parameter lognormal models. Again, there is a strong outlier in the data for this data set but no justification for removal of the point could be found.

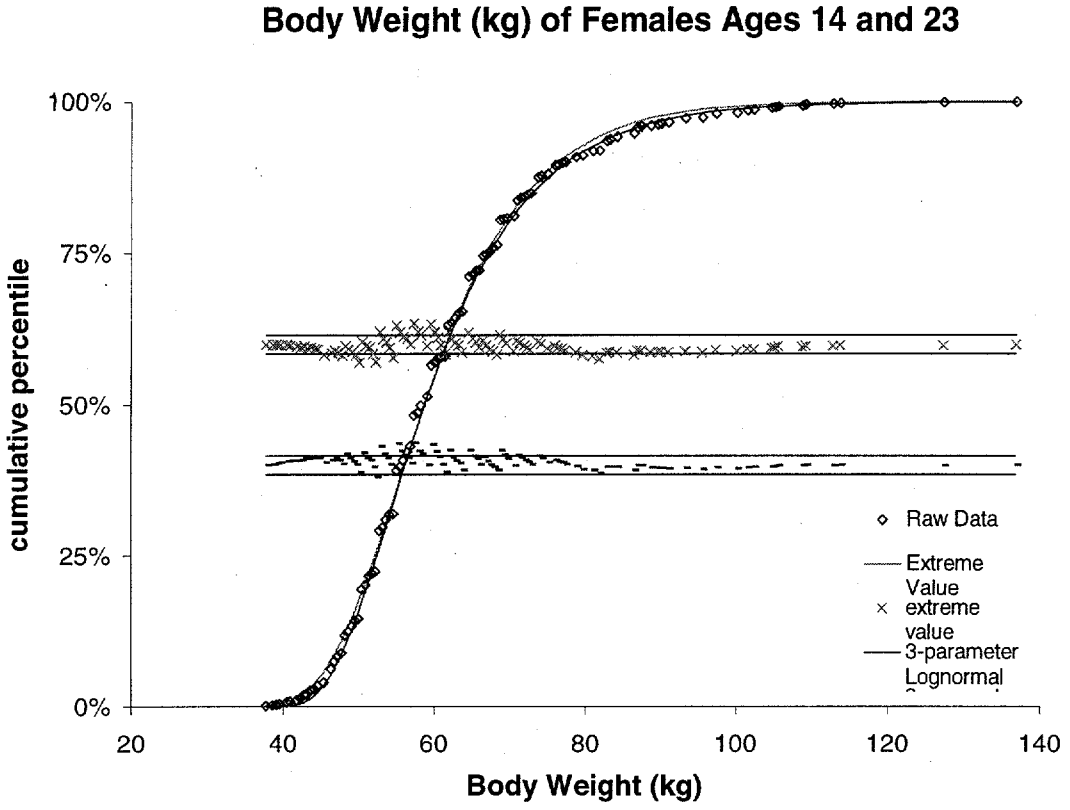


Figure 4.12: Body weight of females ages 14 and 23 years showing the performance of the extreme value and three-parameter lognormal models. The step pattern in the raw data again is thought to be an artifact of the unit conversion and the self-reported values where more values seem to be reported in units of five pounds indicating a rounding tendency in self-reported body weights.

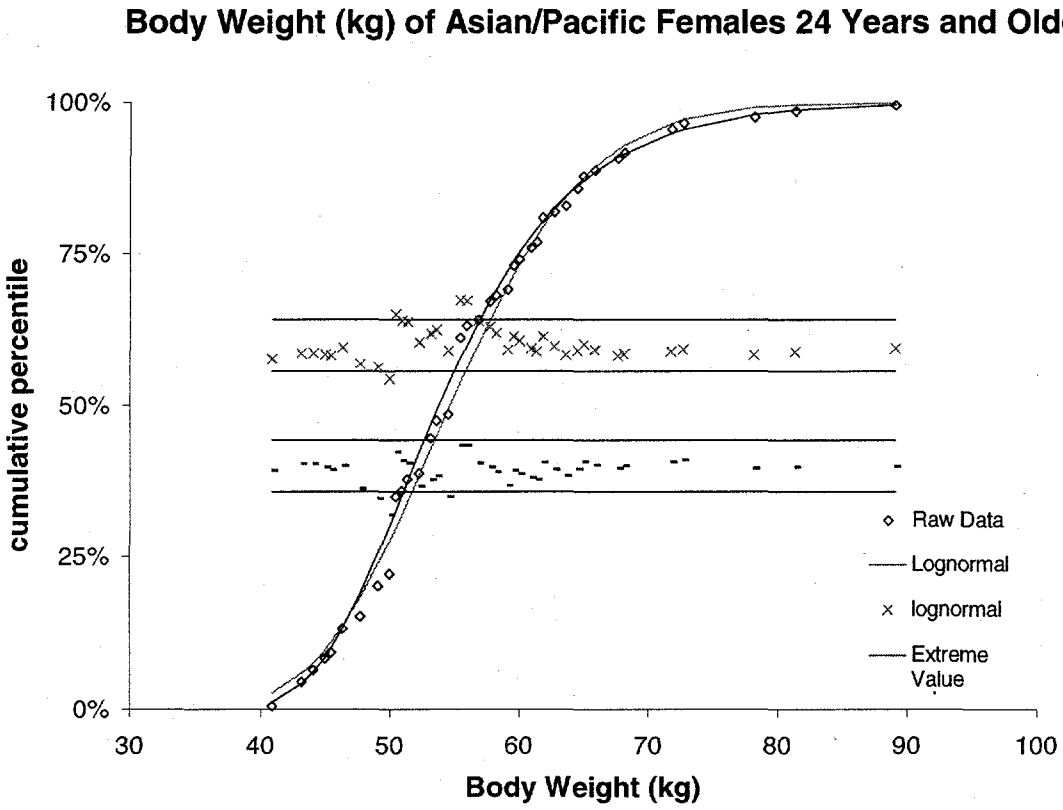


Figure 4.13: Body weight of Asian/Pacific females older than 23 years of age showing the performance of the lognormal and extreme value models.

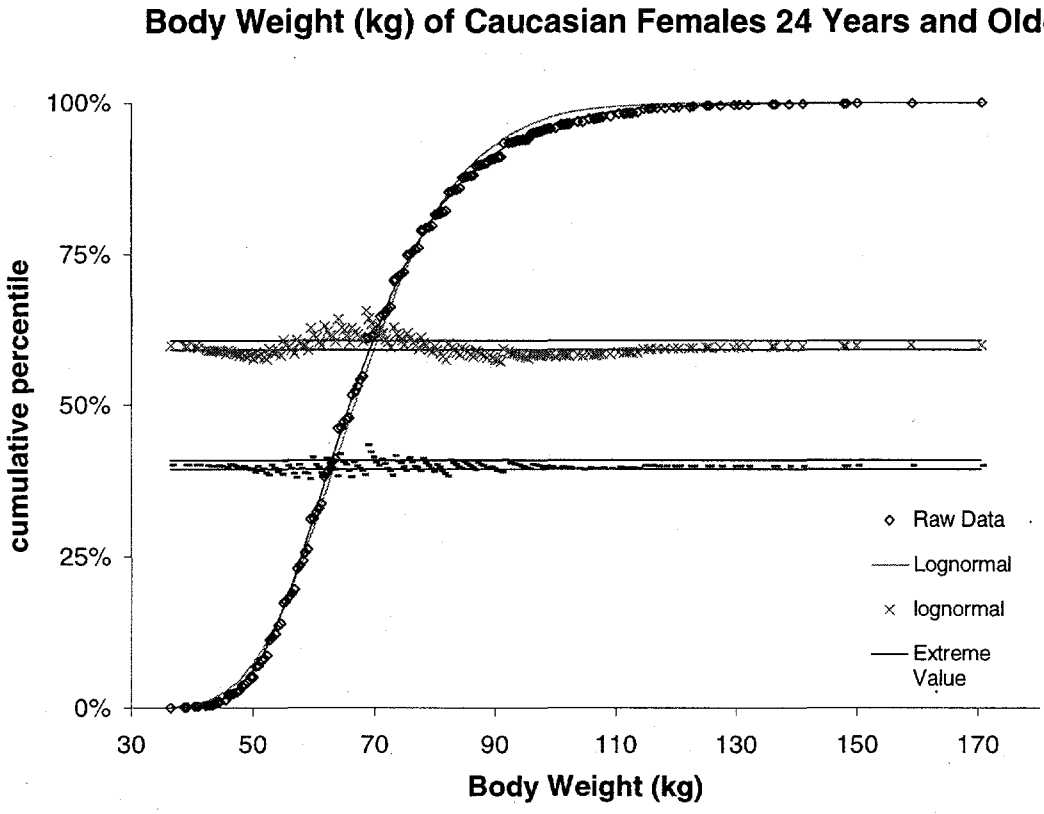


Figure 4.14: Body weight of Caucasian females older than 23 years of age showing the performance of the lognormal and extreme value models. The step pattern in the raw data is thought to be an artifact of the unit conversion and the self-reported values where more values seem to be reported in units of five pounds indicating a rounding tendency in self-reported body weights.

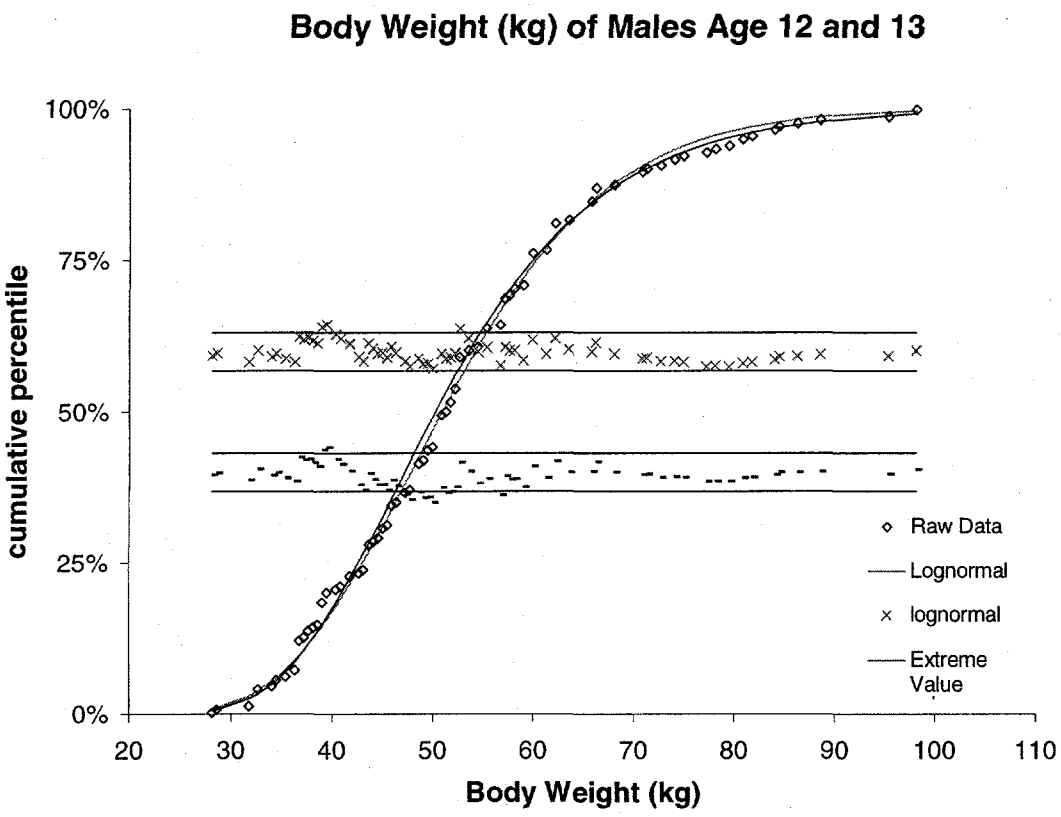


Figure 4.15: Body weight of males that are 12 and 13 years of age showing the performance of the lognormal and extreme value models.

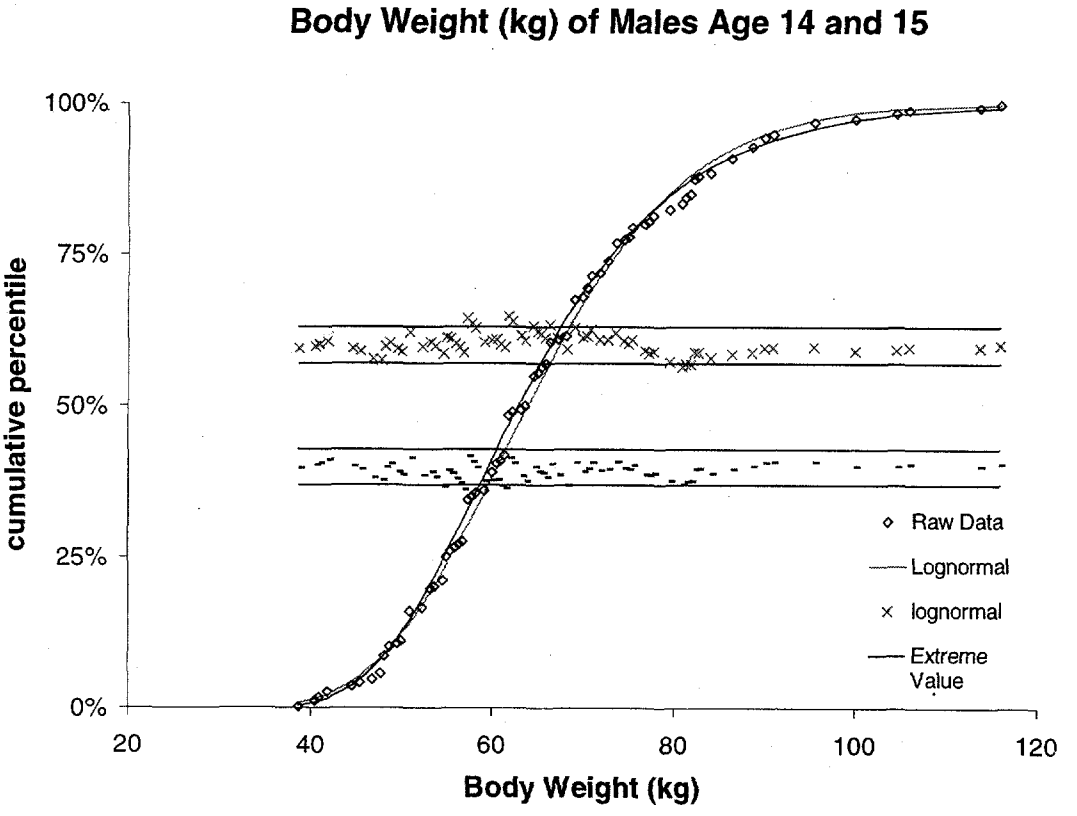


Figure 4.16: Body weight of males that are 14 and 15 years of age showing the performance of the lognormal and extreme value models.

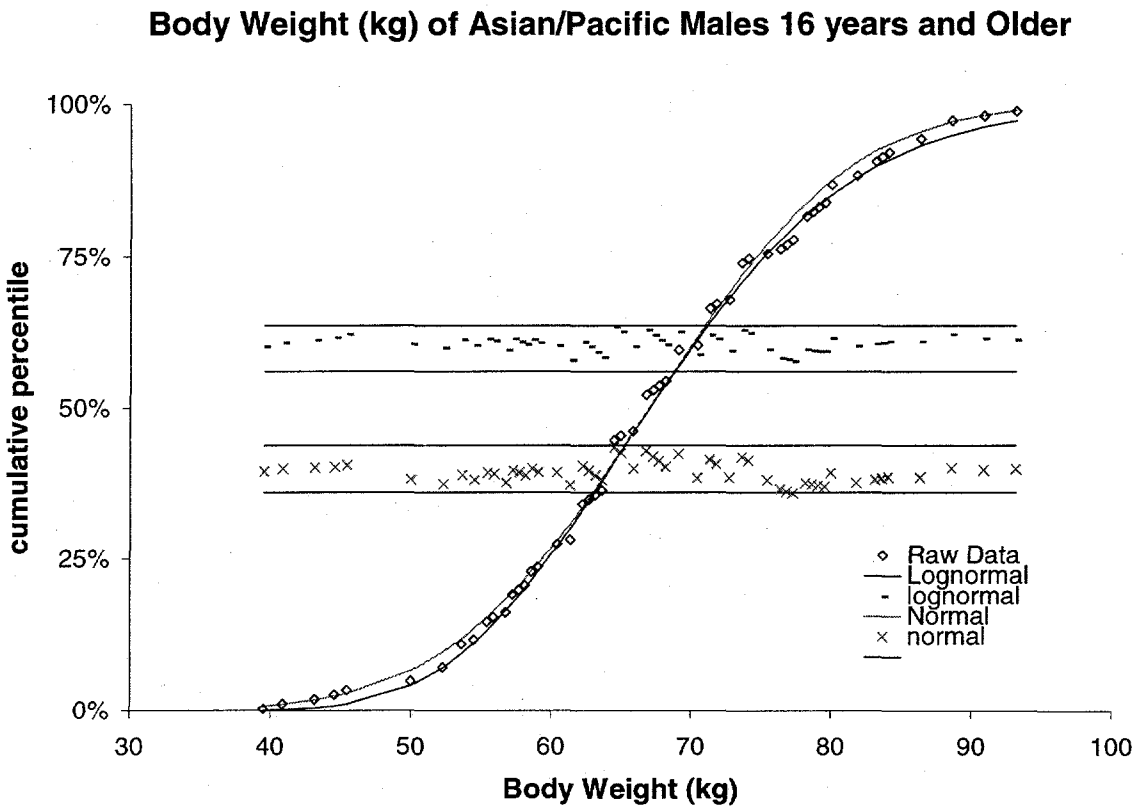


Figure 4.17: Body weight of Asian/Pacific males older than 15 years of age showing the performance of the lognormal and normal models.

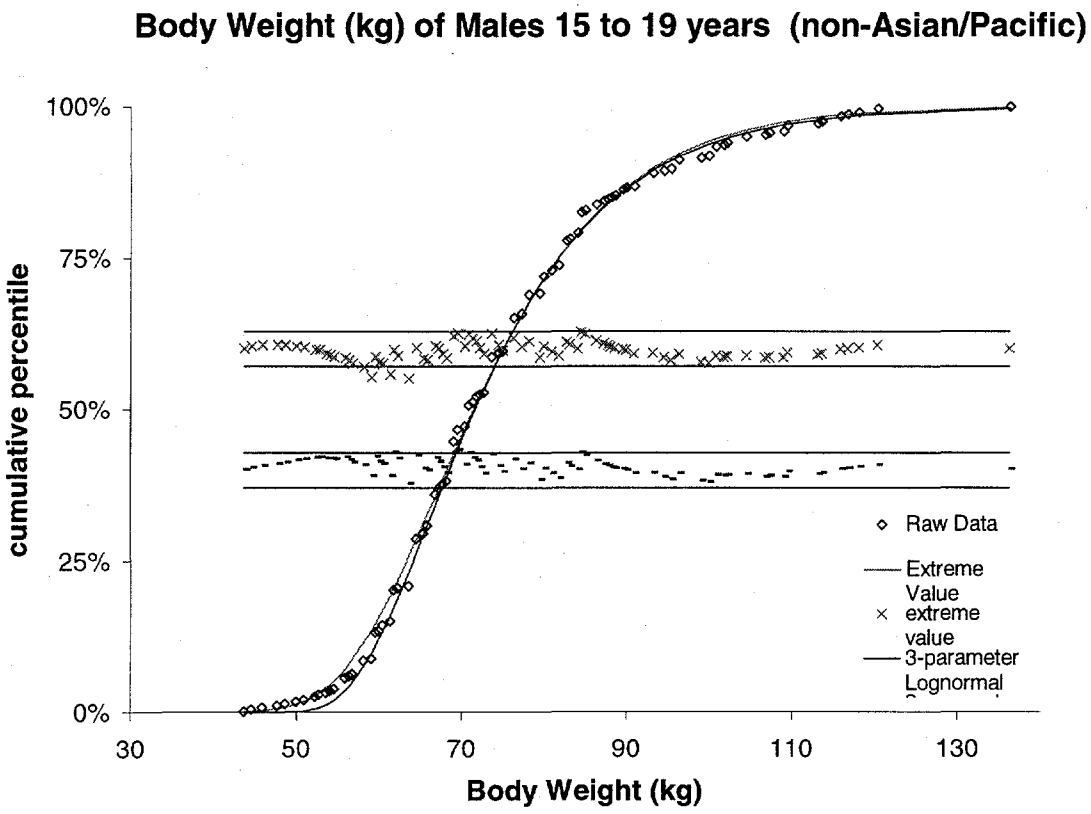


Figure 4.18: Body weight of non-Asian/Pacific males 15 to 19 years of age showing the performance of the extreme value and the three-parameter lognormal models.

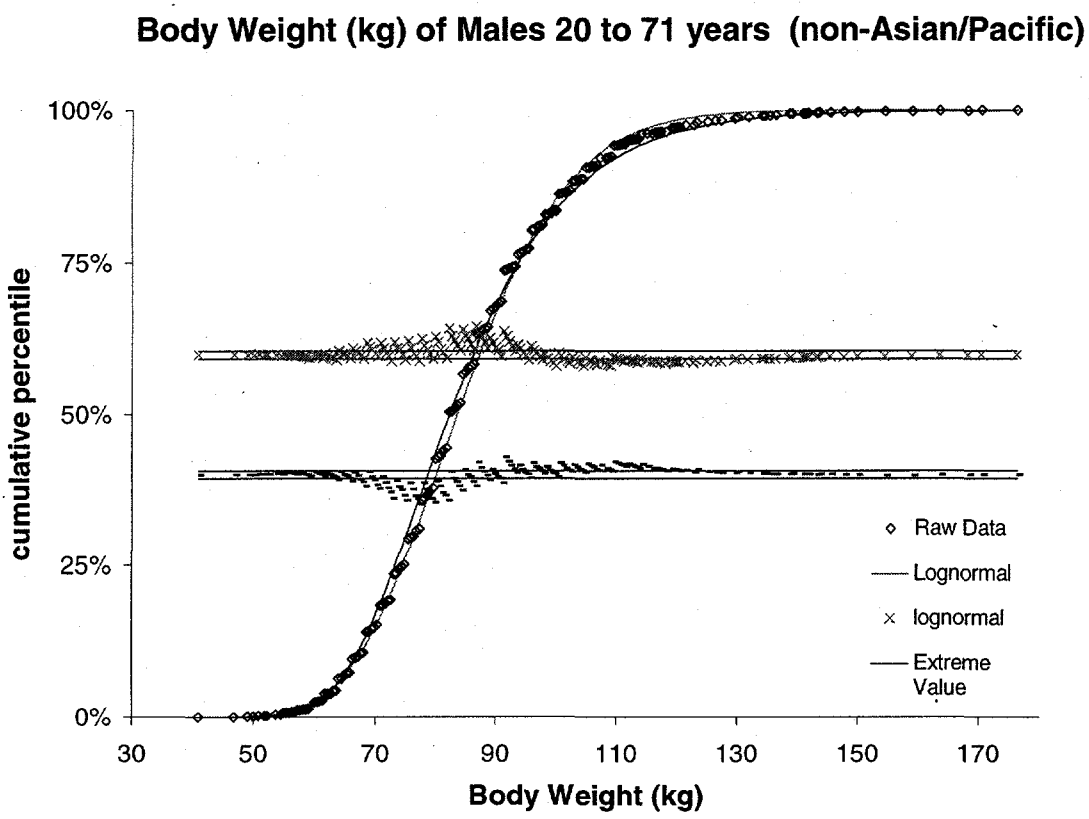


Figure 4.19: Body weight of non-Asian/Pacific males 20 to 71 years of age showing the performance of the lognormal and the extreme value models. The step pattern in the raw data is apparent due to the large sample size and possible reporting bias where weight is rounded to the nearest five pounds. This characteristic shows up more in the data sets with a large number of samples.

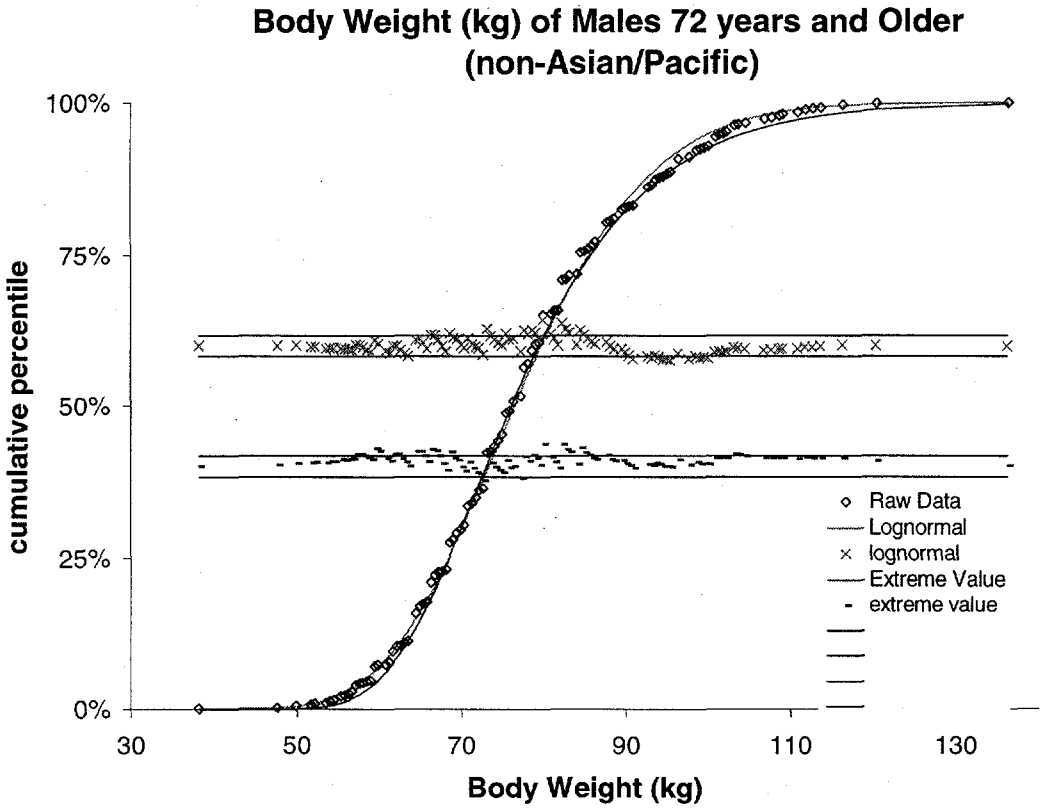


Figure 4.20: Body weight of non-Asian/Pacific males older than 71 years of age showing the performance of the lognormal and the extreme value models. The step pattern in the raw data is apparently due to the large sample size and possible reporting bias where weight is reported to the nearest five pounds. This characteristic shows up more in the data sets with a large number of samples.

Table 4.2: Initial selection and parameterization of distributions for BW

Description of the data set[a]	Distribution	n	KS[b]	A^{2c}	SS[d]	location	scale
1. Ages 1 and 2 years	Logistic[g]	1703	0.051	3.49	0.27	12.10	1.85
2. Ages 3 to 6 years	Lognormal[e]	1610	0.085	7.12	1.30	18.70	4.24
3. Ages 7 to 9 years	Lognormal	672	0.067	2.08	0.25	30.16	7.71
4. Ages 10 and 11 years	Lognormal	486	0.056	1.13	0.12	39.91	10.16
5. Females ages 12 and 13 years	Lognormal	197	0.068	0.83	0.13	51.03	11.43
6. Females ages 14 to 23 years	Extreme Value	757	0.052	1.43	0.21	55.15	9.31
7. Asian/Pacific females 24 years +	Lognormal	102	0.110	0.61	0.09	55.16	8.38
8. Caucasian females 24 years +	Extreme Value	3621	0.047	2.37	0.34	61.68	11.53
9. Black and American Indian females 24 years +	Lognormal	591	0.050	0.69	0.07	76.90	18.22
10. Males ages 12 and 13 years	Lognormal	187	0.049	0.43	0.05	52.51	13.54
11. Males ages 14 and 15 years	Lognormal	201	0.052	0.45	0.04	65.09	14.04
12. Asian/Pacific males 16 years +	Lognormal	133	0.058	0.43	0.02	66.94	11.65
13. Males ages 15 to 19 years[f]	Extreme Value	323	0.052	0.81	0.16	67.14	11.66
14. Males ages 20 to 71 years[f]	Lognormal	4268	0.055	7.58	0.58	84.61	15.12
15. Males ages 72 years +[f]	Lognormal	636	0.049	0.87	0.10	74.14	12.90

(a)

4.6 Uncertainty and variability in the body-weight distributions

The analytical uncertainty in the parametric distribution of body weight is illustrated for sample sizes of 102 and 486 in Figure 4.21. The contribution of analytical uncertainty becomes negligible for sample sizes greater than 100. Because of the relatively large sample size in each data set (n>100) the spread in the data is essentially all due to variability. As a result, 2-D Monte Carlo analyses are not beneficial or necessary for this exposure factor.

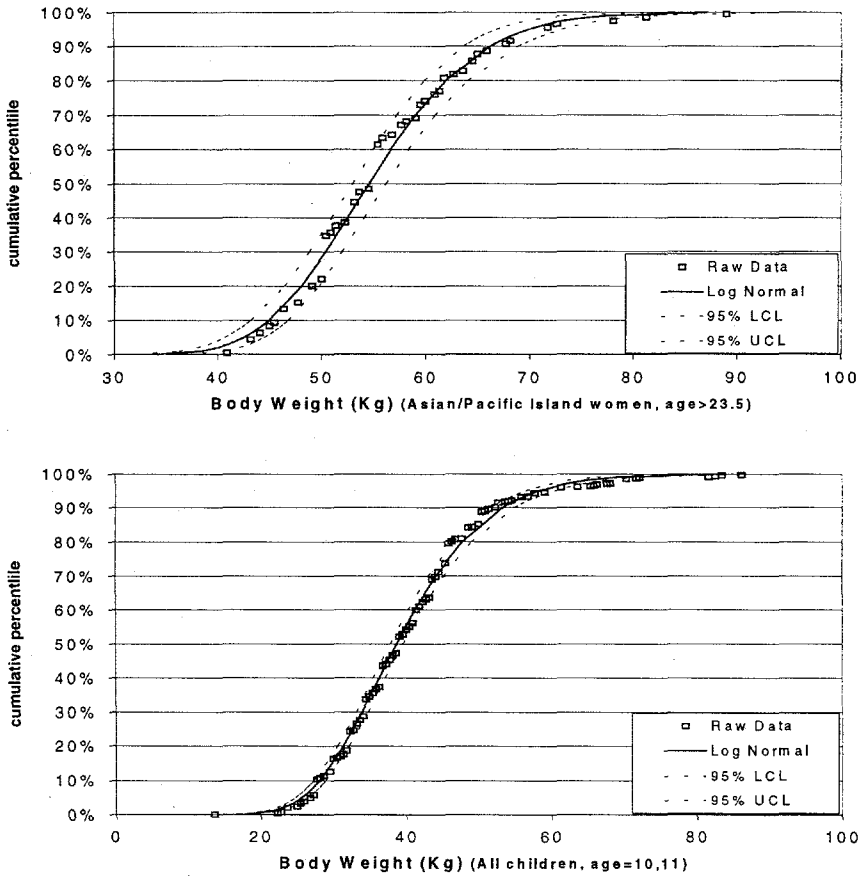


Figure 4.21: Illustration of analytical uncertainty for two distributions. The number of data points used to construct the distribution were 102 and 486 for the top and bottom figures, respectively.

A more interesting and pertinent question may be whether or not there is bias in self-reported body weights. To test for bias, cross-validation experiments could be performed on independent data sets such as the NHANESIII survey (measured values) or any of a number of smaller data sets that contain measured body weights.

4.7 Distribution scores for the body-weight

Overall, the data quantity and quality for body weight are very high. Several nationally representative surveys are available for generating distributions and performing cross-validation experiments. The national survey data is well researched with numerous layers of QA/QC and the surveys include enough demographic information to identify different demographic regions of the data set. We were able to identify standard parametric distributions that provide adequate representation of each subset of the data and visualization techniques were used to assess the quality of fit across the range of the data. The final step in the analysis is to assign scores to each distribution.

Before scoring each distribution, a clear statement of the analysis objective is required. The quality of the distributions cannot be judged without first knowing what population/scenario the candidate distribution is to be used for. In this report, the distributions are scored based on their representation of the demographic subsets from the original national survey. The scoring procedure would be identical if we were judging how well the distribution was expected to represent some other subset of the population, however, the scores would likely change. The questionnaire given in Example 2 of Section 3.3 is used to demonstrate the scoring process for the first distribution (children ages 1 and 2 years). The final score along with the scores associated with the remaining distributions are given in Table 4.3.

The first value in the score sheet is set at 3 because the sample size for 1 and 2 year old children from the CSFII was well over 250. Data relevance was scored a 2 because the body weights were self-reported. The data quality was given a 3 because of the extensive reviews received by the CSFII survey and highly representative and well documented experimental design. The theoretical basis of the distribution for body weight is given a 3 as well because of the likely influence of physical growth over the sample range (0-2 years). We assign the analytical goodness of fit a 2 although it is unclear how well either the KS or A2 tests work with very large samples. The visual performance of the model is good in the region of interest so we assign a score of 2 and no cross-validation experiments have been performed so we assign the last value a 0. Adding the right-hand column we find a total score of 15 which puts us on the border between M and HA. This illustrates the advantage of performing some form of cross-validation with a newly developed distribution. Depending on the performance of the model, a simple comparison to an independent data set could easily bring the model into the HA region or demonstrate the limitations of the distribution. Several of the distributions in Table 4.3 received scores in the M range. This is due primarily to the relatively small sample sizes (n = 50 - 250) for these demographic regions of the data. These scores could also be improved through cross validation.

Example Box 3 Scoring of body weight distribution for 1-2 year old children

For each of the following criteria, enter a number from 0 to 3 in the space to the right. For some questions the values will be arrived at quantitatively and for others the values will be subjectively assigned on a low to high scale. After filling in each of the sections, add up each score and refer to the bar at the bottom of the page to determine the appropriate score.

Sample size

For ($n \leq 10$); enter 0
For ($10 < n \leq 50$); enter 1
For ($50 < n \leq 250$); enter 2
For ($250 < n$); enter 3 ____3____

Data relevance

For irrelevant data; enter 0
For surrogate values; enter 1
For self-reported values; enter 2
For actual measurements; enter 3 ____2____

Data Quality

Score data quality from 0-3 (low to high) ____3____

Theoretical basis for distribution

Score theoretical basis from 0 to 3 (low to high) ____3____

Analytical goodness of fit

For KS or AD in 50%; enter 1
For KS or AD in 75%; enter 2
For KS or AD in 95%; enter 3 ____2____

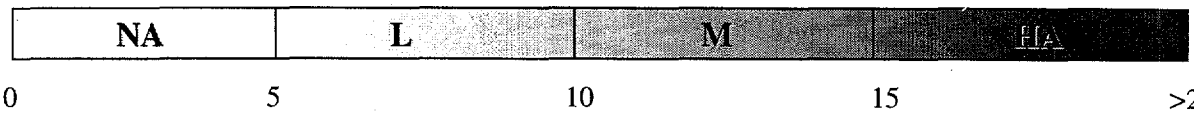
Visual performance

Poor fit across range; enter 0
High scatter but low bias; enter 1
Low scatter low bias in region of interest; enter 2
Low scatter and bias across range of data; enter 3 ____2____

Model performance in cross-validation

Enter 0 if no cross-validation has been performed
Score model performance from 0-3 (poor to good) for each independent cross-validation experiment ____0____

Add the values in the right hand column and locate the score on the following bar. **15**



0 5 10 15 >20

Table 4.2: Robustness scores for body weight distributions

Region number and data description	Robustness
1. Ages 1 and 2 years	H
2. Ages 3 to 6 years	H
3. Ages 7 to 9 years	H
4. Ages 10 and 11 years	H
5. Females ages 12 and 13 years	M
6. Females ages 14 to 23 years	H
7. Asian/Pacific females 24 years +	M
8. Caucasian females 24 years +	H
9. Black and American Indian females 24 years +	H
10. Males ages 12 and 13 years	M
11. Males ages 14 and 15 years	H
12. Asian/Pacific males 16 years +	M
13. Males ages 15 to 19 years	H
14. Males ages 20 to 71 years	H
15. Males ages 72 years +	H

5.0 Development of PDFs for Exposure Duration

In this section we provide a summary and recommendation for exposure duration PDFs. For illustrative purposes, we assume exposure occurs at or near the home so exposure duration is defined as the amount of time that an individual is expected to occupy his or her current residence. There are no direct measurements of exposure duration (ED) given as total residence time. Therefore, ED must be estimated from surrogate data such as the reported amount of time an individual has lived in their current residence or from mobility and mortality data. The use of surrogate data to estimate ED increases the qualitative level of uncertainty in the estimate and increases the level of difficulty associated with the factor analysis used to identify important subsets of the population.

However, there is an abundance of Nationally representative data that can be used with standard statistical methods to estimate distributions for ED from surrogate data. We highlight these data sources and statistical methods below.

5.1 Primary Data Source

The U.S. Bureau of the Census is currently conducting national housing surveys every other year. These surveys provide comprehensive housing statistics for the U.S. Department of Housing and Urban Development (HUD) and include information on housing (apartments, single-family homes, and mobile homes), attributes of housing units (locale, number of rooms, square footage, *etc.*), and data on household members (age, race, gender, income, education, *etc.*). The last survey year from which data have been made available to the public is 1995.

There are 45,675 occupied housing units in the 1995 American Housing Survey National (AHS-N) sample. Weighting[1] factors are provided in order to estimate housing statistics for 97,693,000 housing units in the U.S.. Housing units are identified as rental or owner occupied, in an urban or rural setting, and by geographic region (northeast, midwest, south, or west). Of particular interest to this report, the survey includes information on the age, gender, race, Hispanic origin, salary, education, and *current residence time* of each household member for a total population of 254,159,000 (after weighting housing units in the survey).

1 Appendix B of the 1995 AHS-N report (ref) discusses how housing units in the sample have been weighted to account for the probability of selection, whether the unit could be interviewed, sampling deficiencies for new constructions, and other differences in sampling estimates based on independent sources (*e.g.*, the 1980 census) for key characteristics such as region, tenure, urban or rural status, metropolitan area status, ethnicity, and race.

The size and complexity of the Housing Survey data sets necessitated the development of computer routines to facilitate the extraction and pre-processing of the data. These routines are described in the following section.

5.1.3 Computer Routine for Processing Data

Two computer programs, written in C++, were developed to manipulate data from the 1995 American Housing Survey (AHS) (BoC, 1995). When downloading the AHS data from the U.S. Bureau of the Census Web site, the user specifies what information to include for each housing unit in the downloaded data set. The information may include attributes of the housing unit (owned or rented, geographic region, square footage, etc.) as well as data on household members (age, gender, income, education, current residence time, etc.). The downloaded data consist of one record per housing unit surveyed.

In order to examine current residence time by factors such as age, gender, income, geographic region, etc., it was necessary to first select only occupied housing units, and secondly reformat the data so that there is one record per individual, rather than one record per housing unit. The first computer program works by reading the AHS data one record at a time. If the housing unit is occupied, then one record is generated for each individual in the unit. Each record contains the following:

- the number of the housing unit;
- the individual's age, current residence time, ethnicity, salary, gender, and spanish origin;
- categorical variables that describe the housing unit's census region and metropolitan region and whether the housing unit is owner occupied or a rental unit and whether the housing unit is classified as a farm.
- The output file of the computer program can then be read directly into the CART analysis program.

The second computer program extracts subsets of data from the output file of the first program. At this time, the only selection criteria used by the program are age, whether the housing unit is owner occupied or a rental unit, the census region, and the metropolitan region. These variables were identified by a preliminary CART analysis as being the most important variables in terms of explaining variance in current residence time. The computer program prompts the user for the following information.

- Type of age test to use and the upper and lower limits, a_1 and a_2, respectively. The choices are (1) $a_1 < \text{age} < a_2$, (2) $a_1 < \text{age}$, or (3) $\text{age} < a_2$. The user specifies the

number of the test (1, 2, or 3) and the value of a_1 for tests 1 and 2 and/or a_2 for tests 1 and 3.

- Whether the housing unit is (1) owner occupied, (2) a cash rental, or (3) a non-cash rental.
- In which census regions (northeast, midwest, south, and west) the housing unit should be located.
- In which of the seven metropolitan regions the housing unit should be located.

The computer program then generates an output file that contains records only for individuals who meet the age criterion and who live in housing units that meet all of the other selection criteria. The structure of the computer program is flexible so that additional selection criterion could be added in the future. This program is used to partition the data into various categories depending on the results of the CART analysis.

5.2 Definitions Associated with Exposure Duration

Exposure duration (*ED*): is the expected length of time that an individual will remain in his/her current housing unit. Exposure duration is required to calculate dose.

As noted above, the AHS-N collects data on current residence time, not on the length of time an individual lived in his/her previous residences. The U.S. Bureau of the Census (BoC, 1995) provides data on mobility, *i.e.*, whether an individual has moved within the last year. The next section explains how both of these types of data have been used to estimate *ED*.

5.3 Data analysis

Data were extracted from the AHS database and preprocessed for CART analysis using the computer routines described above. The preprocessed data included the individual sample persons tenure (owner or renter), metropolitan area (urban, suburban or rural setting), geographic region (northeast, midwest, south, or west) and various demographics such as age, gender, race, Hispanic origin, salary and education. The data also reported the *current residence time* (CRT) of each household member. As discussed in the following section, the CRT has been used to approximate the distribution of ED for the population and for various subgroups within the population using either a survival function or a semi-analytical relationship between CRT and ED. Because of the nature of the data and the methods used to estimate ED, we use the CRT variable as a proxy for ED during the factor analysis phase.

An initial CART analysis indicated a strong dependence of CRT on the age of the sample person. Prior to running the full CART analysis, the data was visualized using the Minitab data analysis software. The distribution of CRT as a function of age (Figure 5.1) clearly shows a bimodal relationship. The distribution first maximizes at age 18 (y) then drops to a minimum at age 30. After age 30, the CRT increases linearly with increasing age. A simple explanation for the bimodal shape is that the CRT of sample persons with ages 0-18 (children and adolescents living at home) is directly related to that of individuals of age 20-45 (young adults with children living at home). Including both groups in the analysis may introduce bias towards a lower estimate of ED. Children and adolescents should be analyzed separately and further investigation is warranted but beyond the scope of this study. For the remainder of this section, children and adolescents (age≤18 years) are excluded from the analysis.

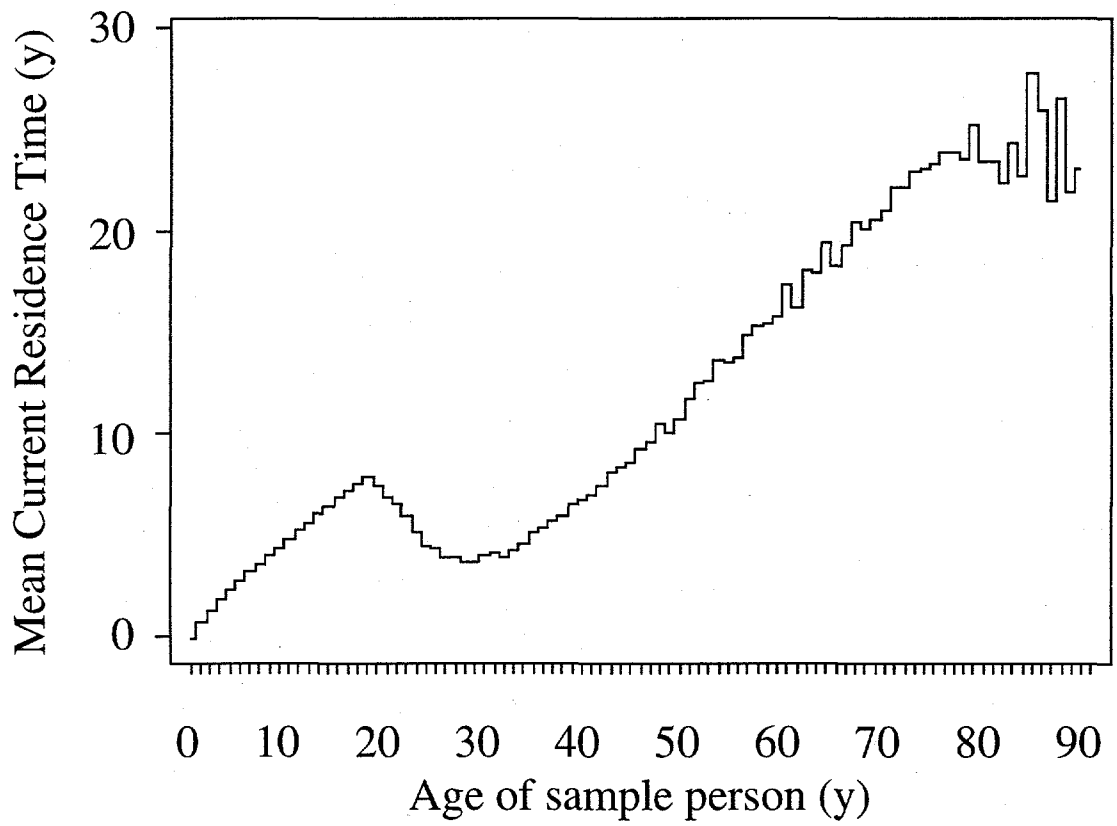


Figure 5.1: Mean current residence time (CRT) reported for each year showing the increasing trend with age and the bimodal characteristic. The height of each bar is the mean for all individuals in that age group.

Results from the CART analysis indicated that several of the factors had no measurable influence on the variance of CRT. For the factor definitions, see table 5.1. The non-influential

factors include ethnicity, gender, spanish origin and farm status. The initial CART analysis did not identify farm status as an important variable even though average CRT for farm and non-farm households was 15 and 8.5 years, respectively. The variance in CRT that is due to the farm variable is either captured by another variable (e.g., metropolitan area) or masked by the background variability within the population. This demonstrates that reliance on the CART output alone may not fully characterize the importance of some demographics variables on the exposure factor.

Table 5.1: Variable definitions used in the housing survey

Variable	Definition
HSHLD	Number of housing unit
AGE	Age of individual in years
CRT	Current residence time in years calculated as difference between survey year and the year that individual moved into the home
ETH	Ethnicity of the individual
INDSAL	Annual salary of individual in dollars
GEN	Gender of individual
SPA	Spanish origin of individual
FRM	Indicates whether housing unit is classified as a farm
MET	Metropolitan area
REG	Census region in the U.S.
TEN	Owner occupied or rental unit

Table 5.2: Relative importance of the variables in analysis of CRT

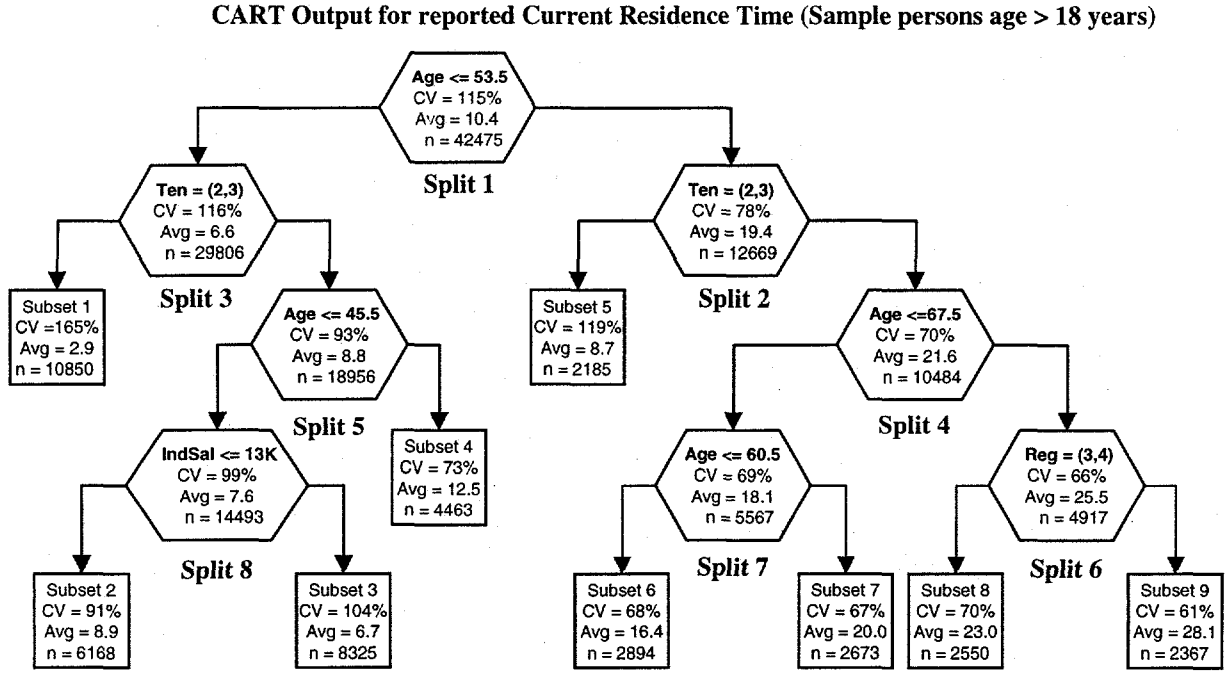
Variable	Relative Importance
AGE	0.71
TEN	0.21
INDSAL	0.04
REG	0.03
MET	0.01

For definitions of variables, see Table 5.1

The relative importance of the individual variables that were included in the analysis is given in Table 5.2 (normalized to 1). The importance of the annual salary of the sample person is questionable because the values included in the housing survey were either $0, continuous between $0 and $100,000 or greater than $100,000. The mixture of censored, continuous and classification data can potentially cause confusion in the CART analysis.

However, there is not sufficient evidence to allow the removal of socioeconomic status from the list of important variables.

Results from the CART analysis of CRT are presented in Figure 5.2 and the composition of the demographic sub-regions are summarized in Table 5.3



Legend
CV = percent coefficient of variation
Avg = average
n = sample size
Variables definitions
Age (continuous yearly values)
Reg = Region (1=northeast, 2=midwest, 3=south, 4= west)
Ten = Tenure (1=owner occupied, 2=rental unit, 3=no cash rent)
IndSal = Annual Salary of Individual in dollars (continuous 0 - 100K, categorical > 100K)

Figure 5.2: Regression tree of current residence time (CRT) for samples persons over age 18 years. The composition of the demographic subsets of the data are provided in Table 5.3.

Table 5.3: Composition of demographic sub-regions for Current Resident Time (CRT)

Sub-region	Characteristics	n	Ave.	CV
Root	Subset of the U.S. Population	42475	10	115%
1	Age≤53 years and non-owner occupied	10850	3	165%
2(3)	Age≤45 years; owner-occupied; salary≤13K	6168	9	91%
3(2)	Age≤45 years; owner-occupied; salary>13K	8325	7	104%
4	Age 45 – 53 years; owner-occupied	4463	13	73%
5	Age > 53 years; non-owner occupied	2185	9	119%
6(7)	Age 53 – 60 years; owner occupied	2894	16	68%
7(6)	Age 60 – 67 years; owner occupied	2673	20	67%
8(9)	Age > 67 years; region (south and west)	2550	23	70%
9(8)	Age > 67 years; region (northeast and midwest)	2367	28	61%

The CART analysis results in Table 5.3 excludes sample persons with ages ≤ 18 years. The analysis was repeated with inclusion of all ages and the splits were similar to those reported in Table 5.3 except that income was not included in the top 10 splits of the data. The split on income ≤ $13,000 is a very low value for annual salary for owner-occupied housing units and we were not able to explain the reason for this split. However, the results indicate that economic status may be an important factor for predicting CRT. If CRT is to be used as a surrogate or proxy for exposure duration, economic status should not be ignored. Previous estimates of ED discussed in the following section do not consider socioeconomic status or the apparent relationship between the CRT of children/adolescents and young adults.

5.4 Statistical and Computational Methods Used to Determine Exposure Duration

Israeli and Nelson, (1992) used weighted data[2] from the 1985 and 1987 AHS-N surveys to estimate expected total residence time for the following groups: all households, renters, owners, urban households, rural households, farms (subset of rural households), and households in four geographic regions (northeast, midwest, south and west). They employed a semi-analytical approach in which they derived expressions relating the fraction of households that moved into their current residence *t* years before the surveys to the fraction of households just moving in at the time of the survey that will be found in the same residence *t* years from now. Values of the fraction of households that moved into their current residence *t* years before the surveys were calculated from AHS-N data and fit with a five-parameter survival function. The expected total residence time is then a function of three of the parameters.

2 Survey data had been weighted in order to estimate housing statistics for the U.S. population. (See previous footnote.)

Johnson and Capel (1992) used a Monte Carlo approach to develop distributions of *residential occupancy time (ROP)*[3] by gender and age. In their simulations, they used population and mobility data from the U.S. Bureau of the Census and mortality data from the National Center for Health Statistics. The data that they used represent the 1987 U.S. population.

The first step in their process was to determine the number of persons in each demographic group of interest, *e.g.*, the number of males or females of a given age. Next, they developed mobility tables and mortality tables for the different demographic groups. Mobility tables give the probability that a person in the demographic group did not move during the previous year. Morality tables give the probability that a person in the demographic group will die during the upcoming year. They then applied a Monte Carlo algorithm that generates current residence time using the mobility tables and future residence time using both the mobility and mortality tables. The *ROP* is one plus the sum of the current and future residence times. Altogether *ROP*'s were generated for 500,000 persons.

Finley *et al.* (1994) discussed and included the work of both Israeli and Nelson (1992) and Johnson and Capel (1992) in their review of distributions of exposure factors. In addition, they derived a formula for and calculated the *ROP* for children born in the household based on moving rates from the U.S. Bureau of the Census. Finley et al. (1994) recommend using the estimates of Israeli and Nelson (1992) for exposure assessments that depend on housing unit characteristics such as geographic location, etc., and using the distributions of Johnson and Capel (1992) for exposure assessments for individuals of specified ages.

Price *et al.* (1998) present a somewhat different computational approach from Israeli and Nelson (1992) for calculating the total duration (*TD*) of a behavior based on the reported duration (*RD*) of and the starting age (*SA*) of the behavior. Their approach is based on relating the probability of *RD* to the probability of *TD* given the value of *SA*. Price *et al.* (1998) applied their method to surveys of anglers rather than to housing data. Their approach, however, is general and could be applied to housing data. In the context of housing data, *TD* would be the total time an individual lives in a residence, *RD* would be the reported residence time when the individual is surveyed, and *SA* would be the age of the individual when he/she started living in the residence (not the individual's age at the time of the survey).

The Oregon Department of Environmental Quality (DEQ, 1998) recommended using a custom distribution that samples the percentiles reported by Johnson & Capel (1992) because the data represents the age dependant exposure duration.

3 Johnson and Capel (1992) use the term "residential occupancy period" or *ROP* instead of *ED*.

Rather than present new distributions for exposure duration, we summarize the strengths and limitations of currently used distributions in Section 5.5 and asses whether the considerable effort needed to generate new distributions is warranted.

5.5 Presentation of the Exposure Duration Distributions

Israeli and Nelson (1992) calculated *average total residence times*[4] in years for the following groups: all households (4.55 ± 0.60), renters (2.35 ± 0.14), owners (11.36 ± 3.87), farms (17.31 ± 13.81), urban households (4.19 ± 0.53), rural households (7.80 ± 1.17), and households in the northeast (7.37 ± 0.88), midwest (5.11 ± 0.68), south (3.96 ± 0.47), and west (3.49 ± 0.57). In addition, they provided a table of total residence time for these same groups corresponding to selected values of the fraction of households just moving in at the time of the survey that will be found in the same residence t years from now.

Johnson and Capel (1992) give their results in terms of the mean and selected percentiles of the *ROP* for all ages for both genders, males only, and females only and also the mean and selected percentiles of the *ROP* for ages 3, 6, 9, ..., 90 for both genders, males only, and females only. The estimated mean *ROP* for all males is 11.1 years, for all females 12.3 years, and for the entire population (both genders, all ages), 11.7 years.

Table X in Finley *et al.* (1994) gives selected percentiles of residential occupancy by residence types (all households, renters, owners, urban households, rural households, and farms). Table XI in their report gives selected percentiles of residential occupancy period by age for "from birth", and ages of 3, 12, 21, 30, and 60 years. Table X is based on the work by Israeli and Nelson (1992), and most of Table XI is based on the work by Johnson and Capel (1992). The percentiles for "from birth" were calculated by Finley *et al.* (1994).

5.6 Uncertainty in the Exposure Duration Distributions

Israeli and Nelson (1992) calculated standard errors for the average total residence times (given in the previous section) from the standard errors calculated for the five parameters determined as part of the curve fitting process. Israeli and Nelson (1992) also give the standard errors for the five parameters in the paper.

Underlying the Monte Carlo process used by Johnson and Capel (1992) are the assumptions that the probabilities in the mobility and mortality tables used are independent of both the calendar year to which they are applied and the history of moving of the person being represented in the simulation. The mobility and mortality rates were determined from 1987 data.

4 The values of the average total residence times are given together with their standard errors derived from the standard

Johnson and Capel (1992) note that the rates "are unlikely to be representatives of rates in effect during earlier decades" and that it would be "difficult to predict the applicability of these rates to future decades." The authors acknowledge that any bias that results from these uncertainties would be difficult to quantify.

Given the sample sizes available through the housing survey and the mortality/mobility data tables, quantitative uncertainty in the estimates and distributions previously reported are likely to be negligible. However, qualitative uncertainty arising from the use of surrogate and/or proxy data to calculate ED, regardless of the calculation method applied, may be extensive. There is also uncertainty in the way that the previous studies partitioned the data. From the CART analysis performed in this section, we a very different set of important factor influencing CRT and the estimate of ED.

5.7 Scores for the Exposure Duration Distributions

The only indication of quantitative or method robustness is from Johnson & Capel (1992) who performed five Monte Carlo simulations for 500,000 sample persons per simulation. The results of the five simulations in terms of mean ROP and selected percentiles are nearly identical. However, concern arises in that the different methods reported in section 5.5 produce somewhat different predictions of ED.

Overall, the data used to construct the distributions for exposure duration are adequate and represent the national population and demographic sub-regions within the population although it appears that information about Native Americans living on reservation land are not included in the housing survey. The parametric distributions presented in the previous studies do a good job of representing the data that was used to generate them. However, the use of a surrogate variables to predict ED and the limited effort to identify significantly different demographic subgroups within the population lead to a low score for data relevance. In addition, to our knowledge, the currently used distributions have not been tested against independent samples (independent samples of measured exposure duration do not exist at this time). Because of the high degree of qualitative uncertainty, we recommend a robustness score of medium (M) to low (L). This exposure factor clearly needs further consideration.

5.8 Recommended Improvements to the Exposure Duration Distributions

The references cited above provide *ED* distributions for the following groups.

- Johnson and Capel (1992): gender and selected ages (ages 3, 6, 9, ..., 90) based on 1987 census and health statistics data.

- Israeli and Nelson (1992): geographic regions (northeast, midwest, south, or west); renter or owner; urban, rural or farm based on 1985 and 1987 AHS-N data.

Neither provides *ED* distributions by socioeconomic status. Such distributions could be determined by applying the analytical/statistical procedures of either Israeli & Nelson (1992) or Price *et al.* (1998) to the 1995 AHS-N data or the Monte Carlo procedure of Johnson & Capel (1992). In addition, the data used to develop current distribution was split on variables that were not found to be important in this analysis (gender, multiple age groups) and the strong relationship between young adults and children living at home was not accounted for.

This section does not attempt to recommend a family of distributions for exposure duration. Rather, it demonstrates and suggests that further work is warranted and that demographically appropriate distributions should be developed for exposure duration. We recommend a reanalysis of both the methods and the data used to estimate ED. The new distributions should be tested and validated using the approach introduced in section 3. In addition, effort should be directed towards developing survey questions to collect data that is not only representative of the national population but is also relevant to the estimate of exposure duration.

6.0 Development of PDFs for Exposure Frequency

In this section we summarize efforts for development and analysis of PDFs for exposure frequency (EF). EF is the frequency in hours per day or days per year that an individual is in contact with the hazard that is being assessed. In practice, information on EF should be collected for each important exposure location and activity. However, for this study we look only at the fraction of the day that an individual spends indoors at his/her primary residence. The method developed here can be applied to other exposure locations/activities as needed.

Exposure frequency is developed from an analysis or survey of personal activity patterns. Several assessments of personal activity have been performed. Typically these assessments are based on diary data prepared by individuals over a set time period – usually 1 to 7 days. This type of data leads to some important questions regarding representiveness of the data used to develop the exposure factor distributions. The over-riding assumption is that within a survey sample, individual diary reports will converge on an unbiased representation of the population.

However, certain activities reported over short time periods are clearly not representative. For example, a significant fraction of the population report (in 3-day diary records) time spent indoors at home as either 100% of the day or 0% of the day. Intuitively, we know that this is not a realistic behavior over the duration of a typical exposure event. To get a reasonable picture of personal activity patterns and to ascertain the distribution of activity for an individual over an extended period of time, data should be collected for several nonconsecutive days. Unfortunately, available data is collected on consecutive days and as a result, we need to make an *a-priori* decision about representativeness. It must be determined whether it is best to construct truncated distributions that capture extreme values in the bounded data set or to treat extreme values as outliers and remove them from the analysis. Given that the reported values for EF are likely an artifact of the method used to collect the data, the removal of these data points cannot be justified. Rather, we state in advance that the distributions resulting from the currently available data (short-term diary data) should be considered to represent short-term behavior of an individual in the population. The distribution of long-term behavior would likely converge on a value other than 0% and 100%.

6.1 Sources of Data

The data for calculating exposure frequency was taken from the National Human Activity Patterns Survey (NHAPS) conducted between 1992-1994. NHAPS consisted of daily diary data results including 9386 different respondents (*excluding Alaska and Hawaii*) over 8 seasonal quarters. Respondents were between 0-93 years and the following demographic data was provided:

- age in yrs (0-90 yrs) and in months (if <1 yr)
- Gender
- race (White, Black, Asian, Hispanic, or other race)
- Hispanic origin
- pregnancy status
- education (grade or years of school completed)
- employment status(full-time, part-time, unemployed, student, retired)
- weekday or weekend and season
- region

  Northeast = Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont;

  Midwest = Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin,

  South = Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia; and

  West = Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming), gender, urbanization (MSA, central city, MSA, outside central city, or Non-MSA),

- and year of survey.

The data were taken from both the NHAPS A and NHAPS B questionnaires. The questions in the NHAPS A and NHAPS B questionnaires are related to the air-exposure pathway and water exposure pathway, respectively. Three different types of questionnaires were administered in the survey. These include an adult survey, a child survey and a proxy survey (for children to young to answer). The fraction of time spent indoors at home was calculated based on the location variables defined as other, own home indoors, own home kitchen, own home living room/family room/den, own home dining room, own home bathroom, own home bedroom, own home study/office, own home garage, own home basement, own home moving from room to room, own home utility room/laundry room, and own home, other verified.

Also included in the survey was information (numeric groupings, 0, 1-2, 3-5, etc.) on the amount of 8 oz glasses of orange juice, lemonade, Kool Aid, or other drinks made with tap water (drink yesterday)? And, qualitative information on the sources of drinking water (i.e., in terms of public water system, private well, or other, and the use of bottled water). The amount [gallons] of water used each week was also recorded. This information is related to other exposure factors included in this study but the qualitative nature precludes its use for either distributional analysis or for use as cross-validation data. The NHAPS survey also includes information on the year moved into current home. This may be useful for estimating or cross-validating estimates of exposure duration at a later date.

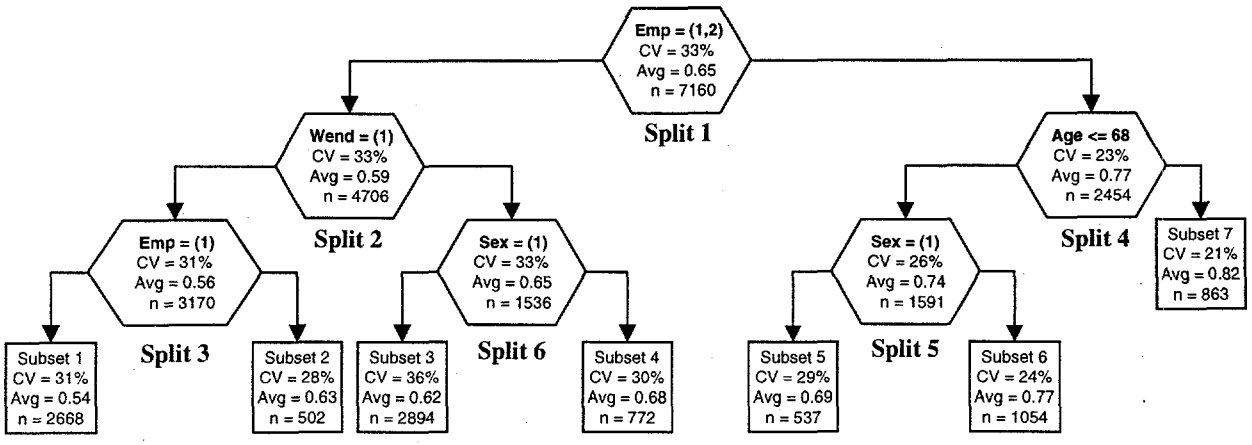
As with other nationally representative surveys, the NHAPS survey includes weighting functions for each sample person to relate the data to the 1990 US Census. Weighting is typically necessary when estimating population characteristics when the characteristics are expected to be different for demographic subgroups within the population (Snedecor and Cochran, 1989, pp 431-455). The weighting functions were developed for (1) over-sampling on weekends, (2) probability of sampling adults, (3) probability of household selection, (4) disproportionate weekday ratios, (5) unrepresentative male/female ratios, (6) disproportionate sampling on weekends, and (7) unrepresentative ratios among ten age groups. There is also an overall weighting variable for each individual in the survey. However, the weighting functions were not required in this analysis because the data set was decomposed into statistically different demographic subsets of the population prior to distribution development (Section 3.2). The reader is referred to Blaire (1995) for additional information and details about the NHAPS-data collection methodologies and sample questionnaires. A condensed questionnaire is also available in the Appendix of USEPA (1996b).

6.2 Data Classification and Distribution Analysis

The original data set for fraction of day spent at home indoors was modified prior to analysis. Modifications include the removal of individuals with missing values for age, sex, race and pregnancy status. The reported quarter of survey was converted to season (spring, summer, fall and winter). The initial CART analysis indicated that the reported time spent indoors at home was strongly dependent on employment status. The majority of sample persons younger than 19 years of age had missing values for employment status. Therefore, all individuals with reported age $\leq$ 18 years (1886 sample persons) were placed in a separate category. The final data set included 7160 sample persons greater than 18 years of age.

The CART analysis was set up using the default options. The analysis was set for regression tree with v-fold cross validation (n=10) and the minimum cost tree was generated using the least squares method. The results are illustrated Figure 6.1 below.

CART Output for reported Fraction of Day Spent Indoors at Home (Sample persons age > 18 years)



Legend
CV = percent coefficient of variation
Avg = average
n = sample size
Variables definitions
Age (continuous yearly values)
Emp = employment status (1= full time, 2= part time, 3= not employed)
Wend = day of week (1= weekday, 2= weekend)
Sex = gender (1 = male, 2 = female)

Figure 6.1: Classification and regression tree showing the decomposition of the original data set for exposure frequency into demographic sub-regions. The tree excludes sample persons younger than 19 years of age. See text for further explanation.

There is a clear difference in EF for individuals employed full-time, part-time and unemployed. In addition, for employed individuals (both full- and part-time), EF is dependent on whether or not it is a weekday or weekend. Interestingly, the reported EF depends on gender during weekend days or when the individual is unemployed. Employed females on the weekend are not significantly different than unemployed males but the data sets cannot be logically recombined. Finally, individuals 68 years of age and older spend the most time indoors at home (average = 0.82). The compositions of the terminal nodes in the regression tree are given in Table 6.1.

Table 6.1: Composition of final EF nodes

Sub-region	Characteristics	n	Ave.	CV
Root	data set for U.S. population (age>18 years)	7160	0.65	33%
1(2)	Employed full-time (weekday)	2668	0.54	31%
2(1)	Employed part-time (weekday)	502	0.63	28%
3(4)	Employed male (weekend)	2894	0.62	36%
4(3)	Employed female (weekend)	772	0.68	30%
5(6)	Unemployed male younger than 68 years of age	537	0.69	29%
6(5)	Unemployed female younger than 68 years of age	1054	0.77	24%
7	Unemployed with age ≤ 68 years of age	863	0.82	21%

The sub-region number refers to the terminal region in the CART tree illustrated in figure 6.1. The number in parenthesis indicates the "sister" node. Sister nodes are generated by splitting a single node and as such can be recombined if it is determined that the difference between the two nodes is not significant. The sub-region number does not reflect the order of importance. For data splitting order, refer to figure 6.1.

6.3 Presentation of Distributions

The output from the CART analysis was used to construct individual data sets for each terminal node in Table 6.1. ECDFs for each resulting demographic region of the sample are illustrated in Figure 6.2. Figure 6.2 includes all sample persons (over the age of 18 years) subdivided into the compositions defined in Table 6.1. Two interesting characteristics of the distributions warrant mention prior to fitting the parametric models to the data.

First, the distribution of full-time employed individuals (weekday) clearly deviates from the smooth form typical of most parametric distributions. This indicates that all of the factors required to explain the variance in the population (demographic subset) were not included in the survey. For example, we were not able to determine if the employed individual was on vacation or home sick on the day(s) of the survey. Fitting the distribution of employed individuals (weekday) requires a mixture model (distribution constructed from two or more parametric distributions) at least until additional information about the population can be gathered.

The second characteristic illustrated in Figure 6.2 is related to the significant number of sample persons reporting the fraction of the day spent indoors at home as 1. This results in an ECDF that does not seem to reach 100%. This is most apparent in Figure 6.2 for the category of unemployed individuals with age greater than 67 years and is least apparent for weekday EF of individuals who are employed either full- or part-time. Fitting parametric distribution to this form of EDF requires that the parametric distribution be truncated at 1. However, rather than simply truncate the distribution and discard the values greater than the specified upper bound (1 in the case of EF), a procedure is used to reduce the values that exceed the upper bound (1) to a value that equals the upper bound.

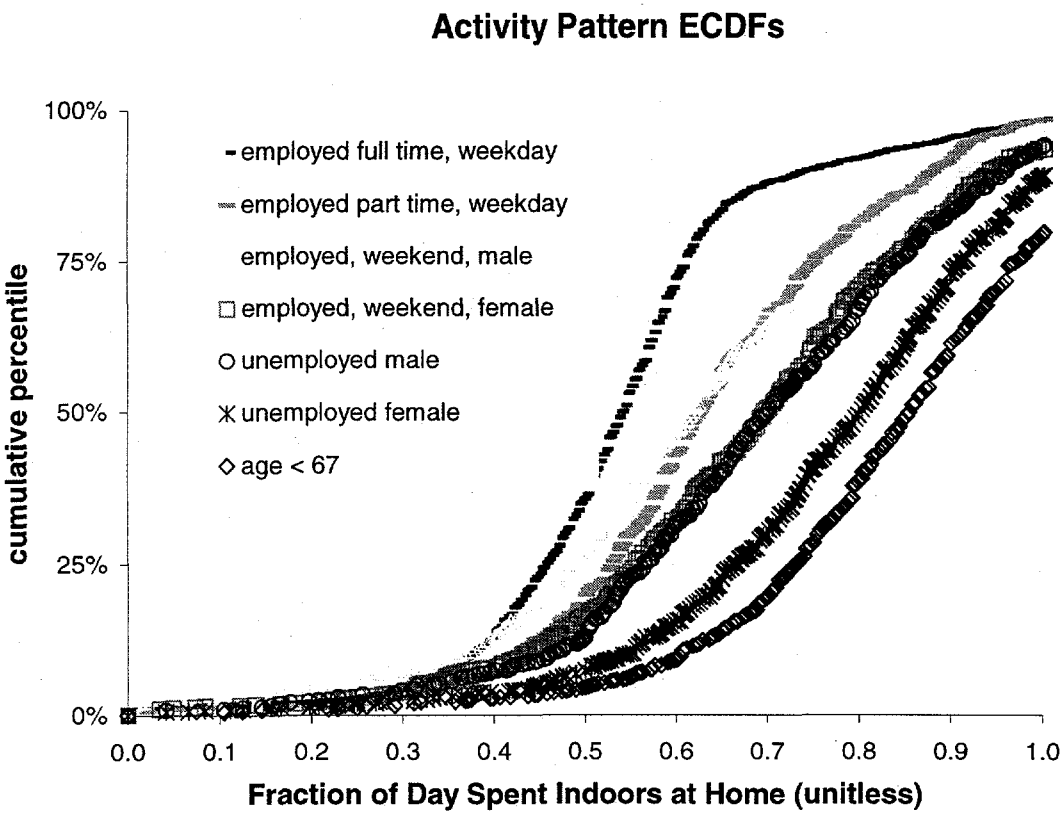


Figure 6.2: Empirical distributions of the sub-regions of EF as determined using a CART analysis of the data collected and reported in the NHAPS survey. For distributions that do not reach 100%, (age > 67) the remainder of the sample reported fraction of day spend indoors at home equal to 1. See text for further explanation.

Results from fitting the parametric distributions to the terminal nodes in Table 6.1 are presented in Figures 6.3 to 6.9. The figures include both the ECDF of the raw data (open triangles) and the best parametric distribution (solid line) along with the 95% bounded residual points (open crosses and solid lines). In all of the distributions, the logistic model performed the best based on the least squares between the predicted and reported values for EF. The Weibull model also performed well but the cost of adding a parameter to the model was not deemed necessary. In Figure 6.3, a uniform distribution was used along with the logistic model to produce an adequate fit of the data. The final results are presented in Table 6.2.

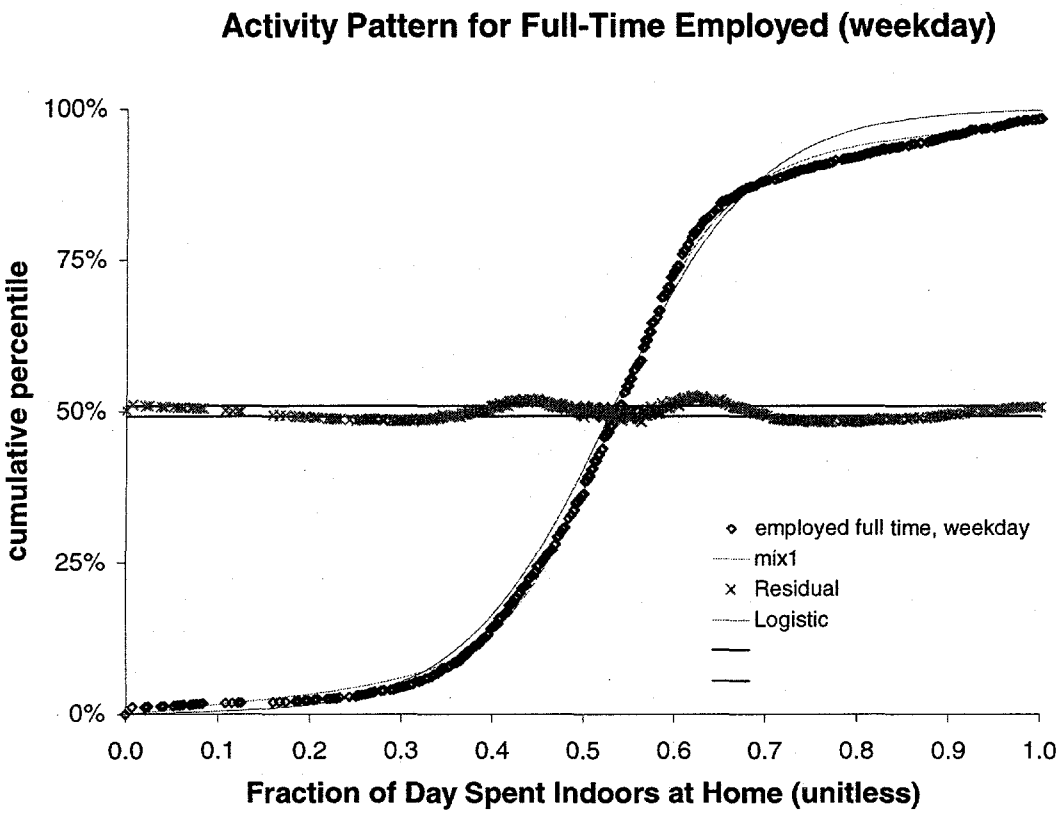


Figure 6.3: Parametric distributions and ECDF for full-time employed individuals reporting time spent indoors at home on a weekday. The logistic distribution failed to explain the apparent uniform nature of the upper tail of the distribution (residuals not shown). However, a mixture model using a combination of logistic and uniform distributions provide a reasonable fit of the data. The mixture model was optimized using least squares between the estimated and reported EF along with the solver routine in the spreadsheet program . The model was generated using the following form:

$$X \approx \text{Logistic}[x_1,\mu_1,\sigma_1] + (1-\pi)\text{Uniform}[x_1,a_1,b_1]$$

where μ_1 and σ_1 are the mode and scale of the logistic distribution, a_1 and b_1 are the lower and upper bound of the uniform distribution and π is the mixing variable (Burmaster and Wilson, 1999).

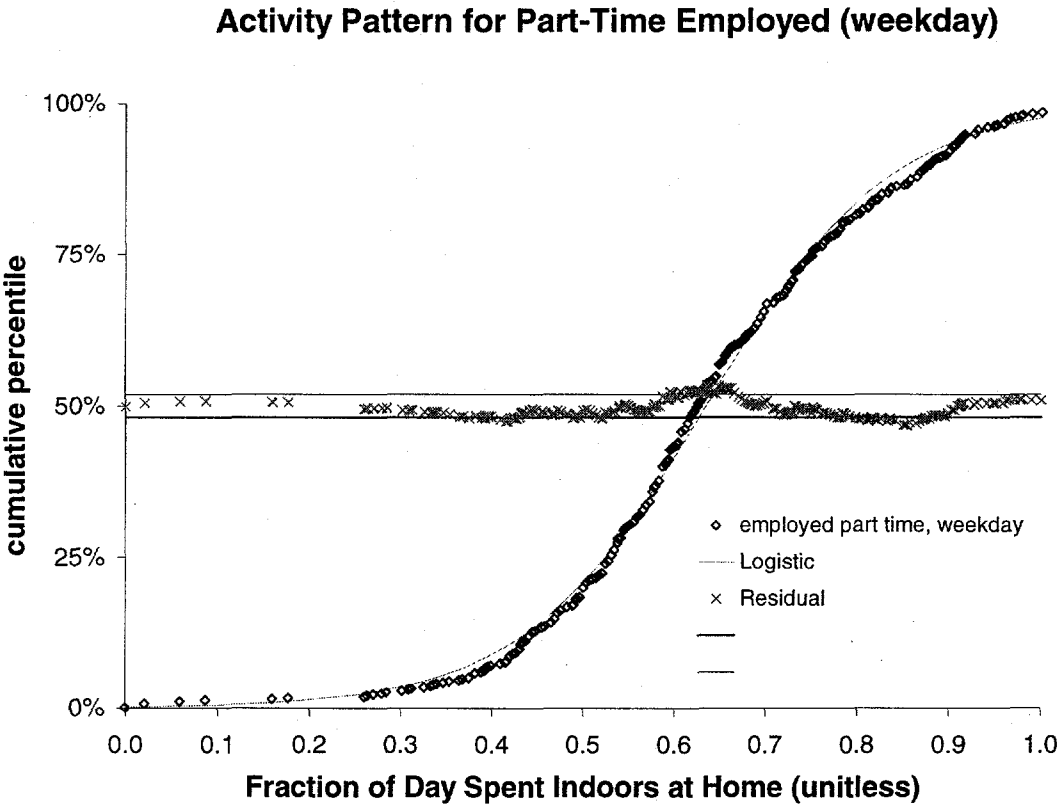


Figure 6.4: Parametric distributions and ECDF for part-time employed individuals reporting time spent indoors at home on a weekday. The logistic distribution was able to provide an adequate fit to the data.

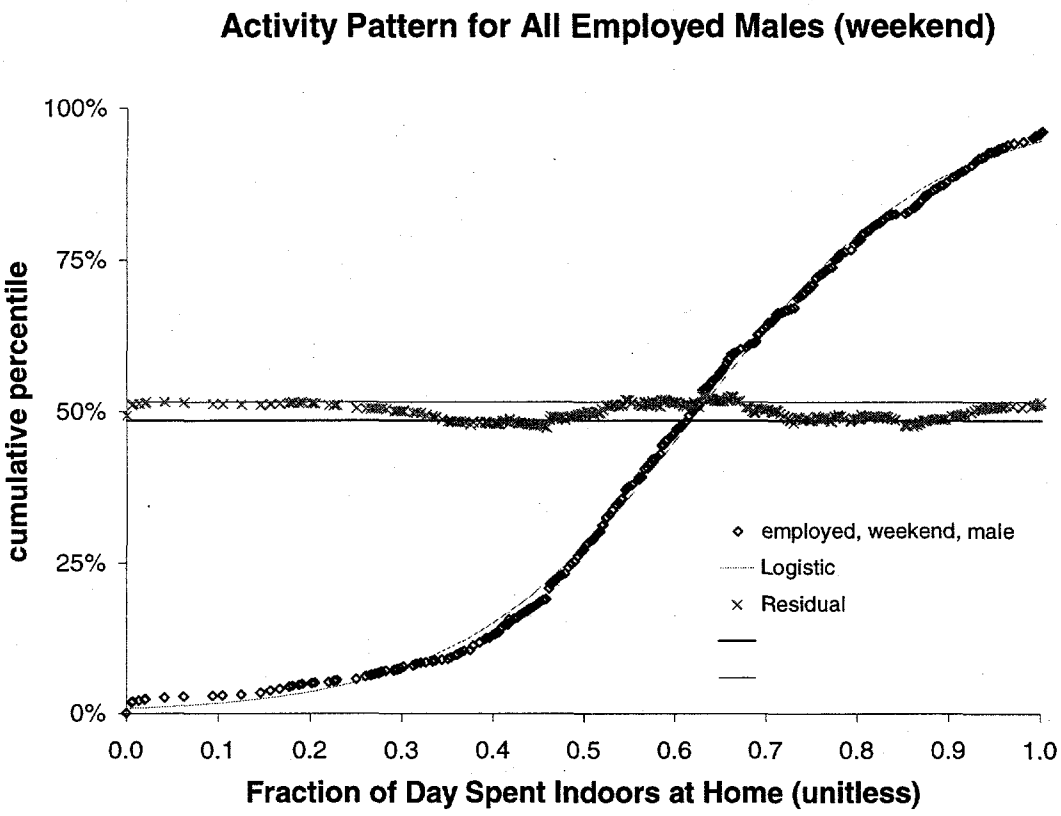


Figure 6.5: Parametric distributions and ECDF for employed males reporting time spent indoors at home on a weekend. The logistic distribution was able to provide an adequate fit to the data.

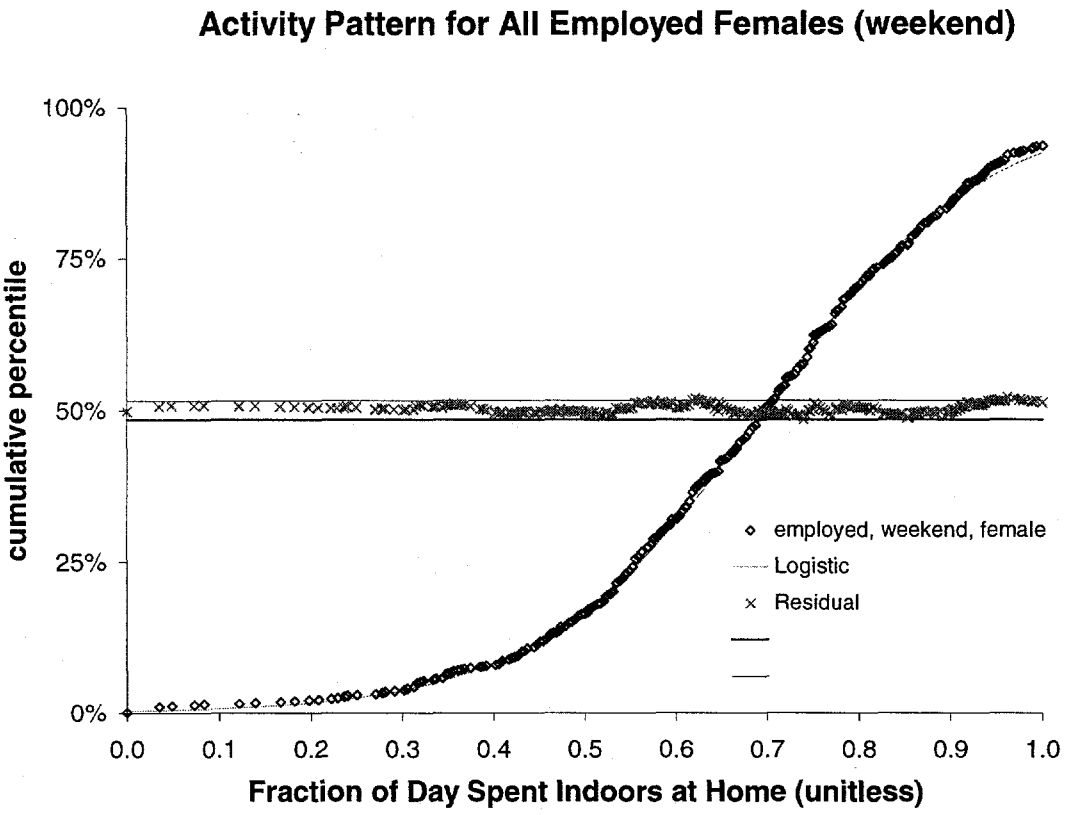


Figure 6.6: Parametric distributions and ECDF for employed females reporting time spent indoors at home on a weekend. The logistic distribution was able to provide an adequate fit to the data.

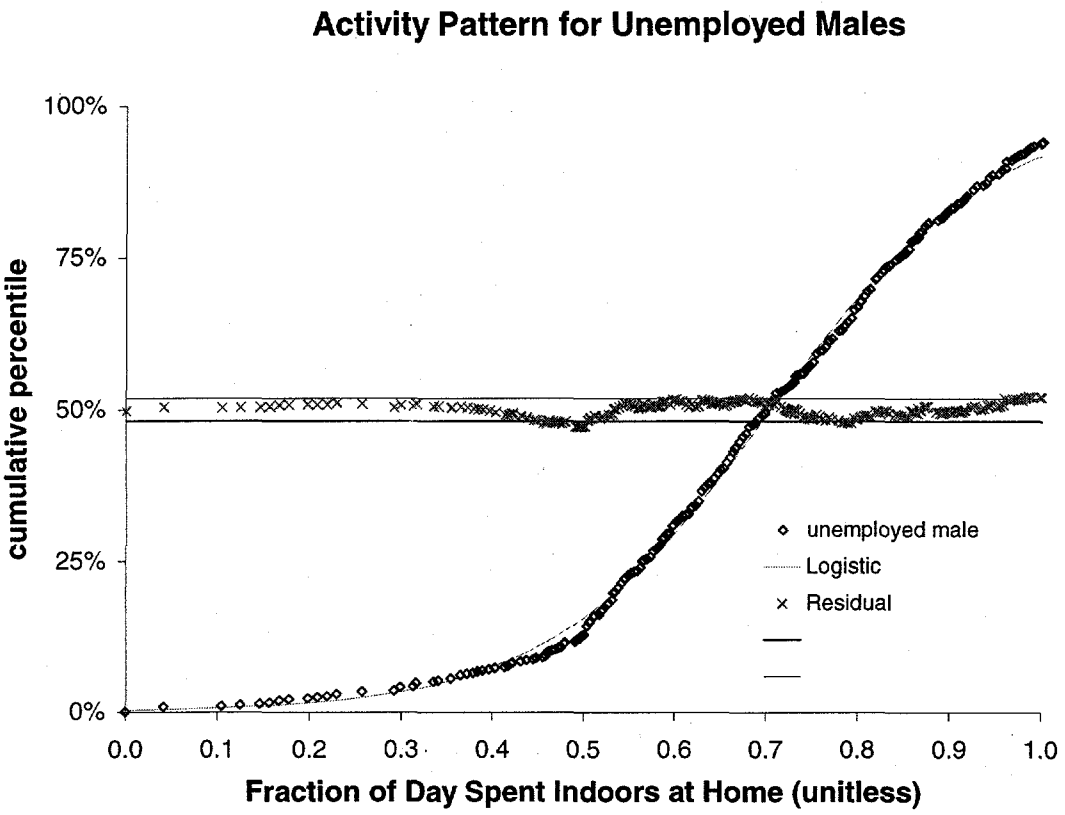


Figure 6.7: Parametric distributions and ECDF for unemployed males reporting time spent indoors at home. The logistic distribution was able to provide an adequate fit to the data.

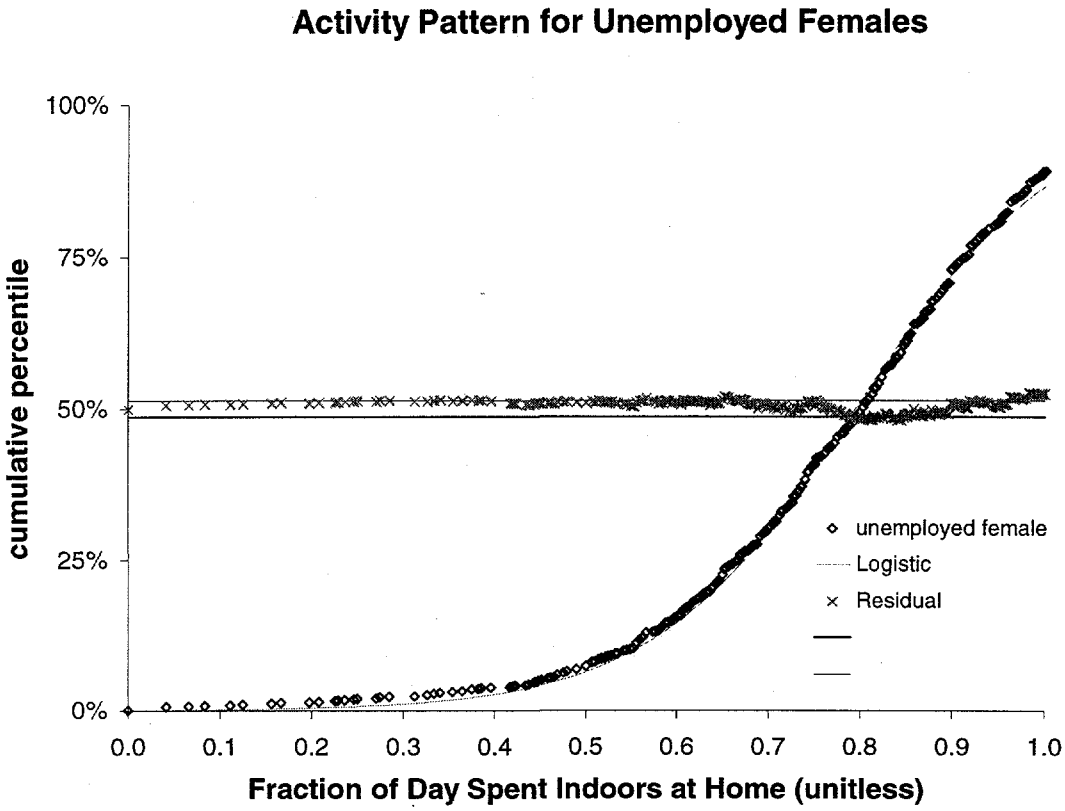


Figure 6.8: Parametric distributions and ECDF for unemployed females reporting time spent indoors at home. The logistic distribution was able to provide an adequate fit to the data.

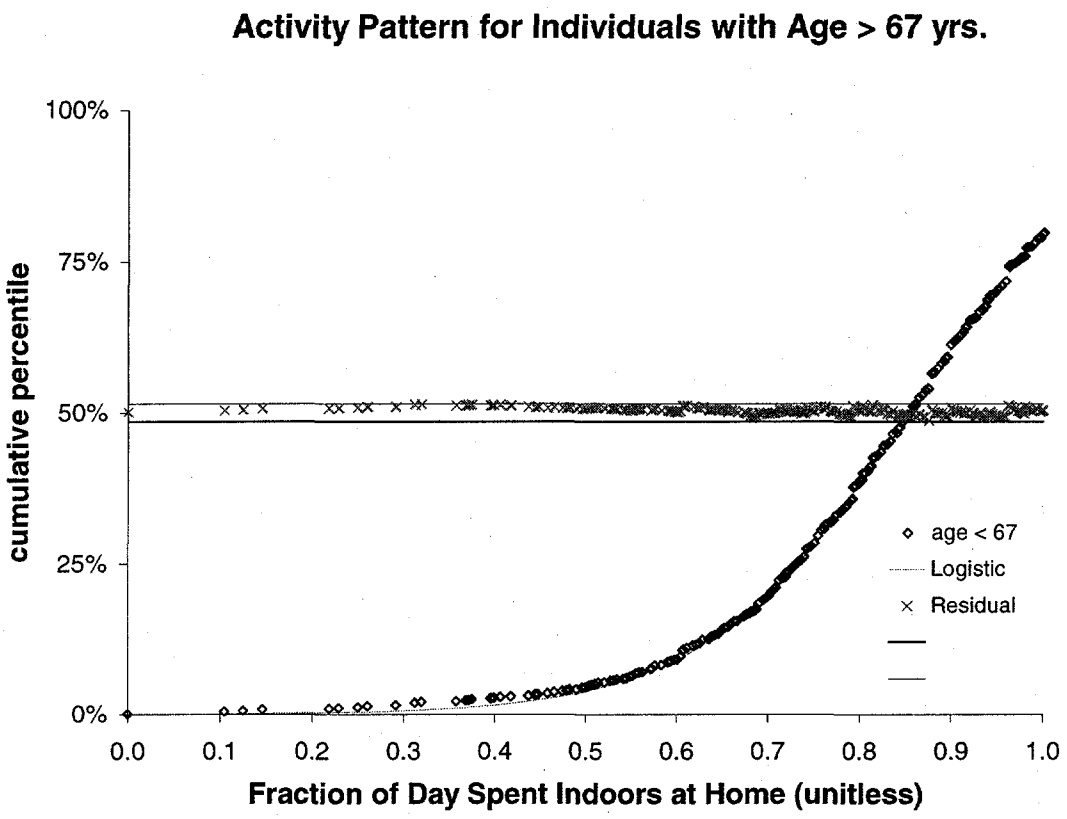


Figure 6.9: Parametric distributions and ECDF for unemployed individuals over 67 years of age reporting time spent indoors at home. The logistic distribution was able to provide an adequate fit to the data.

Performance of the pseudo-truncated parametric models can be tested by generating a distribution of random variable (sample size equal to that of the reported sample) using the selected parametric model. This is demonstrated for terminal node 7 (unemployed individuals over 67 years of age). The results should fall on a diagonal line between zero and one. The performance of the model is shown in figure 6.10. The figure includes a secondary x-axis across the top of the chart showing the approximate percentiles. The secondary axis shows that the model performs well for values above the 5 percentile. Given that EF is used in the numerator of the risk calculation, performance at the low end of the distribution is not considered as critical as the upper end performance and as such, the model is deemed to be satisfactory.

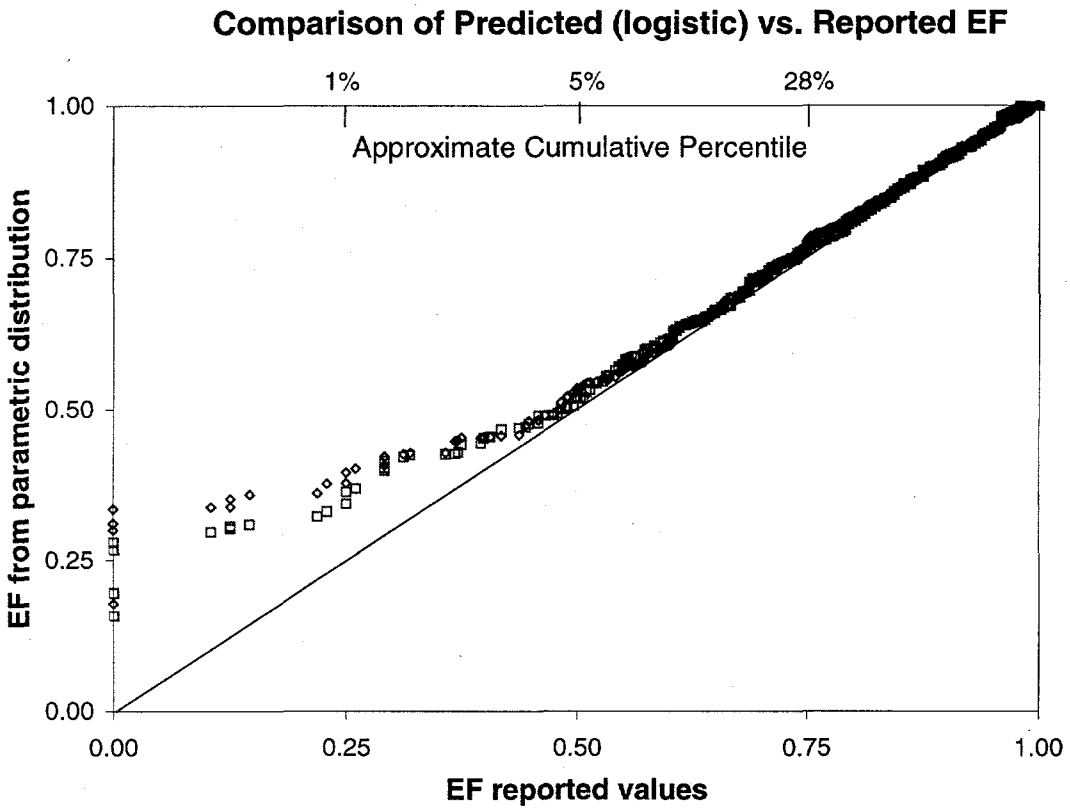


Figure 6.10: Comparison of the random numbers generated using the pseudo-truncated logistic distribution to reported values for time spent indoors at home. The diagonal line indicates perfect agreement between the values. The values across the top axis show the approximate percentiles of the data showing that the upper 95% of the distribution performs well.

Table 6.2: Initial selection and parameterization of distributions for EF

Description of the data set[a]	Distribution	n	location	scale	Multiplier[b]
1. Employed full-time (weekday)	Logistic	2668	0.53	0.057	0.82
	Uniform		0	1.14	
2. Employed part-time (weekday)	Logistic	502	0.63	0.10	
3. Employed male (weekend)	Logistic	2894	0.62	0.13	
4. Employed female (weekend)	Logistic	772	0.69	0.12	
5. Unemployed male	Logistic	537	0.70	0.12	
6. Unemployed female	Logistic	1054	0.79	0.11	
7. Individuals over 67 years of age	Logistic	863	0.85	0.11	
Individual 18 years of age and younger	Logistic	1886	0.67	0.11	

(a)

The theoretical basis for the logistic and mixture distributions used to fit the reported EF values is unclear. Although the pseudo-truncated approach used to account for the fraction of respondents reporting an EF of 1 seems to perform well (Figure 6.10), the theoretical basis for the model is unclear. The visual performance of the parametric models (residuals) was very good, however, the analytical goodness-of-fit scores (KS and AD) were not calculated for these distributions. Finally, the ability of the recommended distribution to forecast samples from independent but related surveys was not measured. As a result, the final scores for the EF distributions are low (L) to medium (M) as reported in Table 6.3.

Table 6.3: Scores for Exposure Frequency distributions

Description of the data sets	Robustness score
1. Employed full-time (weekday)	L
2. Employed part-time (weekday)	L
3. Employed (full- and part-time) male (weekend)	M
4. Employed (full- and part-time) female (weekend)	M
5. Unemployed male 67 years of age and younger	M
6. Unemployed female 67 years of age and younger	L
7. Unemployed individuals over 67 years of age	M
All respondents 18 years of age and younger	M

The best way to improve upon the scores for EF is to demonstrate that the self-reported short-term diary data is both relevant and representative. In addition, a better understanding of the theoretical basis for logistic and mixture distributions is warranted.

7.0 Development of PDFs for Inhalation Rates

In this section we provide a summary on the development and scoring of PDFs for inhalation rate. Inhalation rate is dependent on activity and is often estimated from other metabolic factors (METS) such as heart rate or caloric intake. There is a wealth of information in the sports medicine literature about inhalation rate and its relation to physical exertion but these studies are often targeted towards healthy individuals and as a result, provide little information on potentially important demographic subsets of the population. In addition, the relationship between physical activities that a person performs throughout the day and the level of exertion in a laboratory setting using a treadmill has not been clearly established.

Adams (CARB, 1993) designed a study to test how well METS predict inhalation rate for a small but representative sample. The study found that heart rate actually did a poor job predicting inhalation rate. A separate study reached a similar conclusion (Mermier et al., 1993). The reason for discrepancy between heart rate/inhalation rate correlation from laboratory treadmill studies and from field studies may be related to the use of upper body and lower body muscle. Upper body muscle used in field activities consumes more energy than walking which in turn increases oxygen demand (Adams, 1993). However, given the ease of measuring heart rate and the intuitive relationship between physical exertion and both heart rate and inhalation, METS data may ultimately be the best available surrogate for predicting inhalation rates. However, identifying the relationship between METS and IR is beyond the scope of this project. Rather, this section uses actual measurements of IR at various activity levels to develop and score representative distributions.

7.1 Sources of data

The California Air Resources Board, 1993 and the Human Performance Laboratory at University of California, Davis, performed laboratory and field studies of individuals inhalation rate (IR) while resting (three levels), active (two levels), and while performing common activities such as cleaning house, mowing the yard and driving. Resting protocol involved lying down, sitting, and standing while the active data were collected at running and walking modes on a treadmill (w/ speed [mph] included).

Measurements taken during active and resting protocols included ventilation rate, [L/min] (measured at body temperature, standard pressure), heart rate [beats/min], breathing frequency [breaths/min] and volume of oxygen consumption [L/min]. Anthropometric and demographic data (age, height in [cm], weight in [kg], and body surface area, gender and race) as well as field data during specific activities were also included in the data set.

Active and resting data were given for each age and a cross validation group of children (6-12 yrs) was also included in the study. Within each age group, males and females were included as well as members from 4 ethnic groups: Black, Hispanic, Caucasian, and Asian. Data was collected for both active and resting experiments on 20 male and 20 female subjects (19 Males in the active >59 age category) except the young children. For the young children, data on six male and six female subjects was collected for the active and resting protocol.

Layton (1993) presented a second study that may be of interest for predicting the distribution of long term average breathing rate. Three approaches were described for calculating breathing rates primarily for the purpose of assessing the quantitative dose of airborne nuclides. In his first approach, food-energy intake from the USDA 1977-78 Nationwide Food Consumption Survey was used to estimate the oxygen demand (volume of oxygen consumed in production of 1 kJ expended energy) and related to inhalation rate. The IR was adjusted for underreporting of foods. The second approach involves calculating IR from the ratio of total energy expenditure to basil metabolism rate (BMR), as well as ventilatory equivalent, average oxygen uptake (volume of oxygen consumed in production of 1 kJ expended). The third approach incorporates time activity data from Sallis, et al. (1985). IR's were calculated based on age, gender and specific activity.

For the initial development of IR distributions, we use the CARB study (CARB, 1993) because of its relationship to activity level and the availability of anthropometric measurements.

7.2 Definitions Associated with Inhalation Rate

- Alveolar Ventilation Rate:= the absorbed dose of oxygen. Equivalent to: 'tidal volume – space in lungs'. Approximately 70% of the total ventilation
- Tidal Volume:= Volume of air respired per respiratory cycle/breath.
- Lung Volume:= total/max volume of air that can occupy lung
- Ventilatory Equivalent:= Ratio of minute volume (i.e., ventilation rate) [L/min] to oxygen uptake [L/min]
- Basal metabolic rate:= minimum amount of energy required to support basic cellular respiration while at rest and not digesting food.
- Minute volume:= volume of air exhaled per minute

7.3 Data Classification and Distribution Analysis

The original data was obtained from the California Air Resources Board in the form of multiple spreadsheets – each containing information related to a specific activity or field experiment for a group of individuals. The data were combined into a single set including all available information for each individual in the study. For the initial classification described in this section, only the laboratory experiments were used. Field studies can be analyzed at a later date to determine how each field activity relates to the laboratory activity level used to construct the distributions (as done by Beals et al., 1993). The original data set reported multiple measurements (typically 5) at each activity level for each individual. For the factor analysis, we look only at the average value reported for the activity but note that there was significant variability between measurements for each individual (inter-individual variability).

Activity was categorized and included in the analysis to determine the degree to which we could separate the activities given the variance in the data. The final data set included activity, age, gender, race and BW although the sample size for race was much too small to make any conclusions about differences in inhalation rate. A preliminary analysis shown in Figures 7.1**a** found that normalizing the intake rate to body weight significantly reduced the scatter in IR which is in agreement with an earlier analysis of the data (Beals et al., 1996). Adams (CARB, 1993) recommended using body surface area to normalize IR but the body surface area is a function of height and body weight. Thus, all IR data was normalized to body weight ($l\ kg^{-1}\ min^{-1}$) prior to analysis.

For younger children, there is a strong dependence on age as shown in figure 7.1b. This is thought to be due to changes in metabolic rate as children grow (CARB, 1993; Beals et al., 1996). In addition, physiological differences between males and females are expected to results in gender specific differences in IR as (CARB, 1993) however, the sample size may be too small to detect these differences. The appropriate level of partitioning in the data was left to the objective analysis in CART.

The regression tree analysis with v-fold cross validation (n=10) and the minimum cost tree was generated using the least squares method. The CART results support the suspected dependence of IR on age and activity. However, the analysis did not clearly support separation of the three resting categories into individual levels. This is confirmed visually in Figure 7.2 and 7.3 for adults and children, respectively. Gender differences were not clearly established in this analysis but that may be due to the small sample size. The data is separated into only 6 subsets, shown in Table 7.1 and defined as resting, walking running for children under 12 years of age and for persons 12 years and older.

Figure 7.1: Figure **a** shows an age dependent plot of inhalation rate and figure **b** shows the reduction of scatter in the data when the inhalation rate is normalized to body weight. Figure **b** also shows the age dependence of the normalized inhalation rate for individuals younger than 12 yrs.

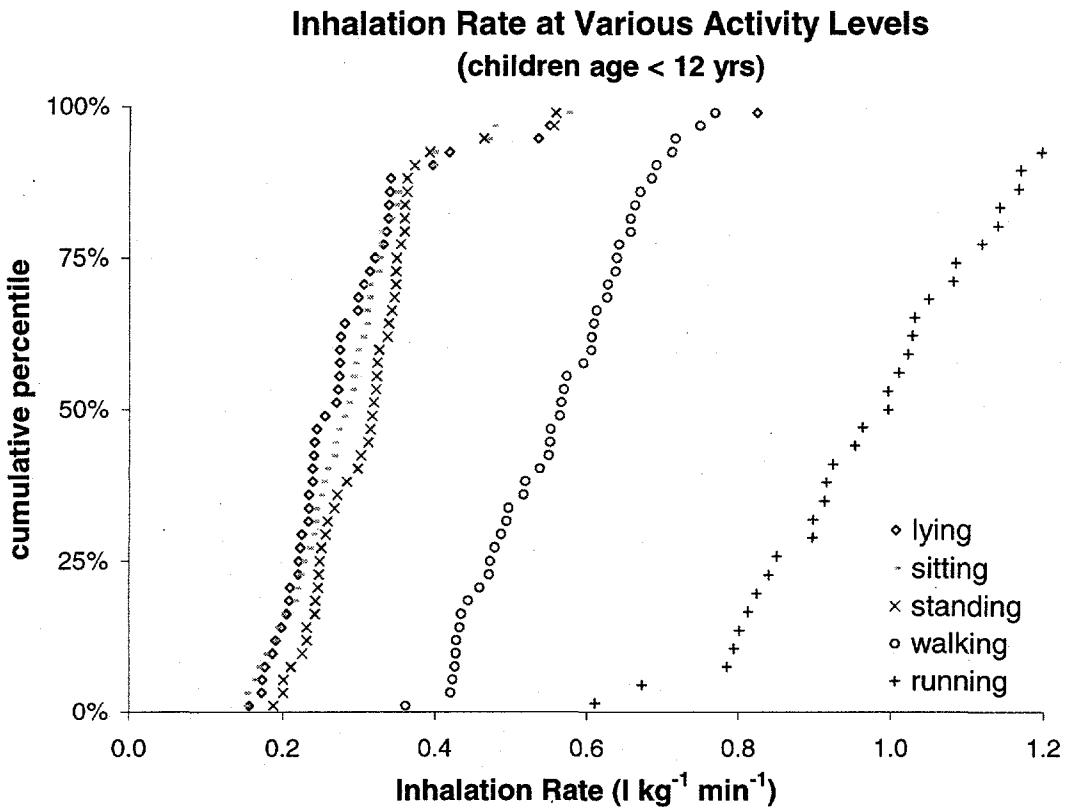


Figure 7.2: Plot of the normalized inhalation rate for different activity levels for children under 12 years of age. Note that there is little difference between the three "resting" activities. The resting categories are combined resulting in three activity categories for the distributional analysis.

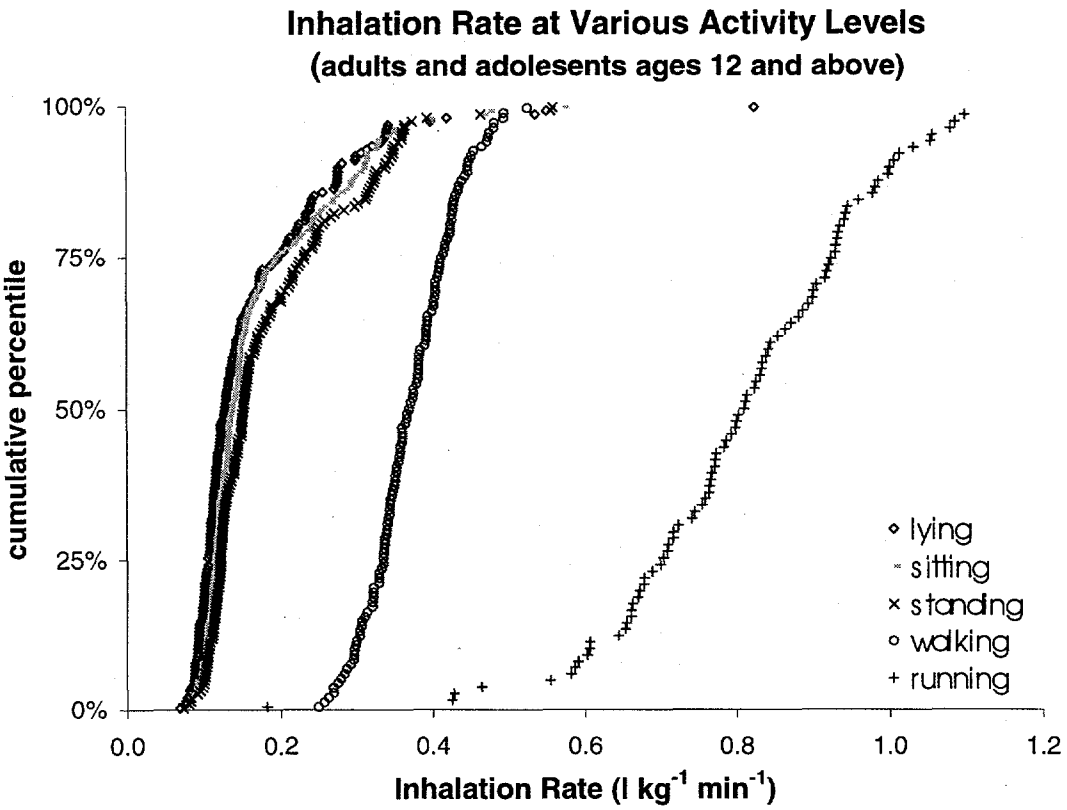


Figure 7.3: Plot of the normalized inhalation rate for different activity levels for individual 12 years of age and older. Again, there is little difference between the three "resting" activities. The resting categories are combined resulting in three activity categories for the distributional analysis.

Table 7.1: Composition and Summary Statistics for Final IR Data Sets

Characteristics	n	Ave.	CV
Children under 12 years of age resting	138	0.29	32%
Children under 12 years of age walking	46	0.56	18%
Children under 12 years of age running	33	0.98	17%
Individuals 12 years and older resting	378	0.13	27%
Individuals 12 years and older walking	125	0.37	15%
Individuals 12 years and older running	93	0.80	21%

7.4 Presentation of Distributions

The data subsets listed in Table 7.1 were analyzed and parametric distributions selected for each group. Figures 7.4 to 7.6 illustrate the ECDFs for children under 12 years of age and Figures 7.7 to 7.9 include all individuals 12 years and older.

The best distributions identified by the Anderson Darling goodness of fit test for the children under 12 years of age were the Extreme Value, Triangular and Normal for resting, walking and running, respectively. However, when one considers the uncertainty due to the small sample size (illustrated by the 95% confidence interval about the residuals) it was not possible to distinguish between the performance of those distributions and the performance of the Lognormal distribution. Therefore, given its general acceptance and ease of use, the Lognormal distribution was selected over the other parametric models for the inhalation rate of children under 12 years of age performing the three different activities.

The best distributions identified for the individuals 12 years and older were the Extreme Value, Weibul and Logistic for resting, walking and running, respectivily. Again, for the resting and walking data sets, the Lognormal distribution performed as well given the quantitative uncertainty in the data. For the running data set the Logistic distribution was used. Although a theoretical basis for this model is not apparent, the fit was clearly better than that of the other parametric models and, as such, we could not justify using another model.

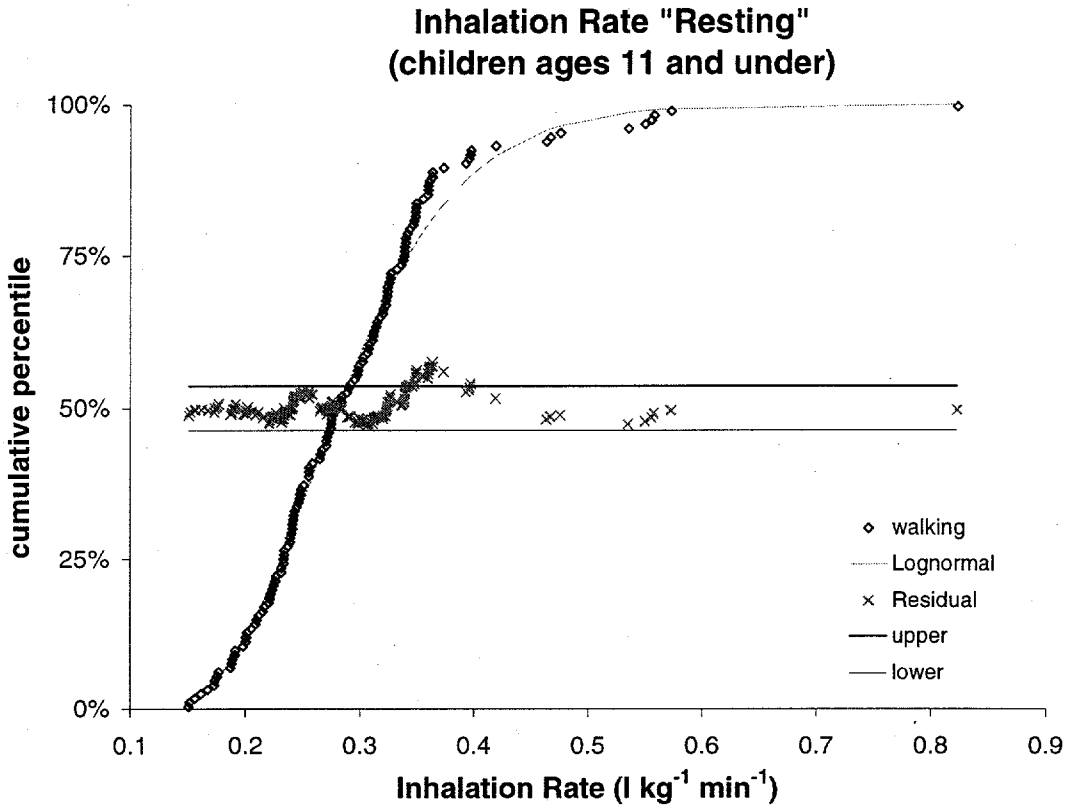


Figure 7.4: Plot of the ECDF for Children under 12 years of age during the "resting" activities along with the fitted Lognormal distribution and residuals. The shape of the upper tails of the ECDF indicates a possible mixture model. This may be due to the different levels of "resting" used in the study (lying, sitting and standing) however, the affect was not great enough to warrant further decomposition of the data set.

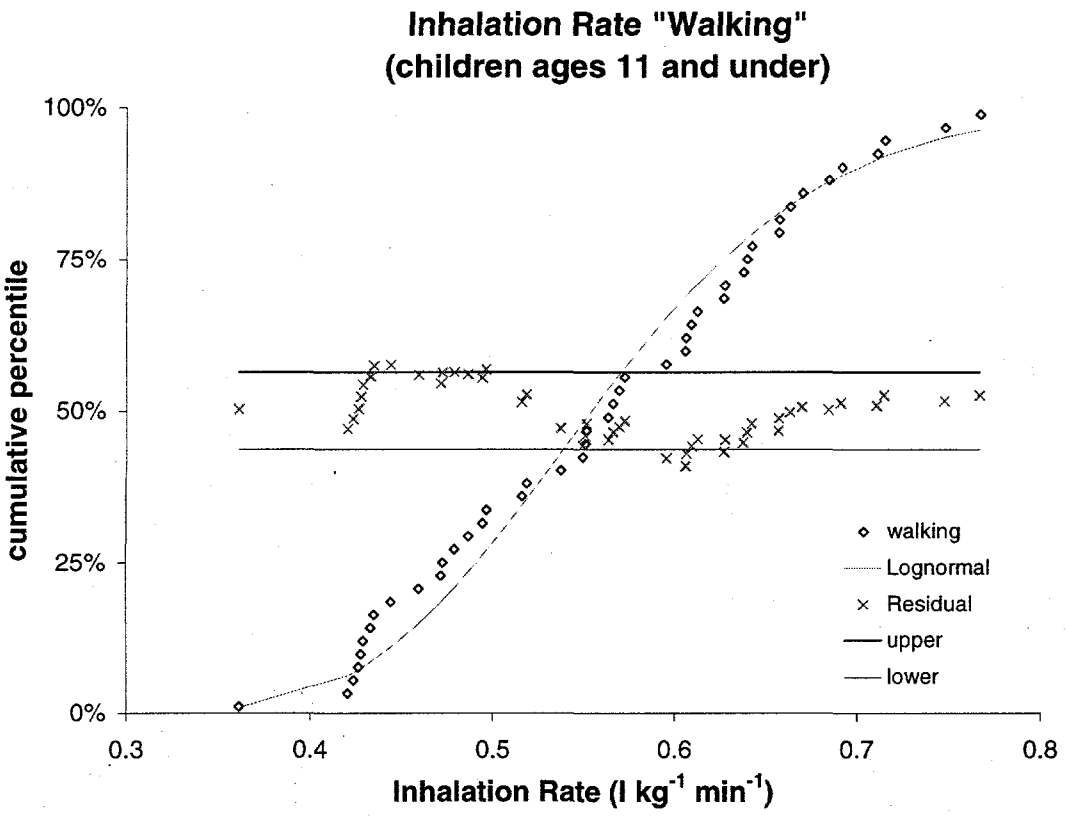


Figure 7.5: Plot of the ECDF for Children under 12 years of age in a "walking" activity along with the fitted Lognormal distribution and residuals.

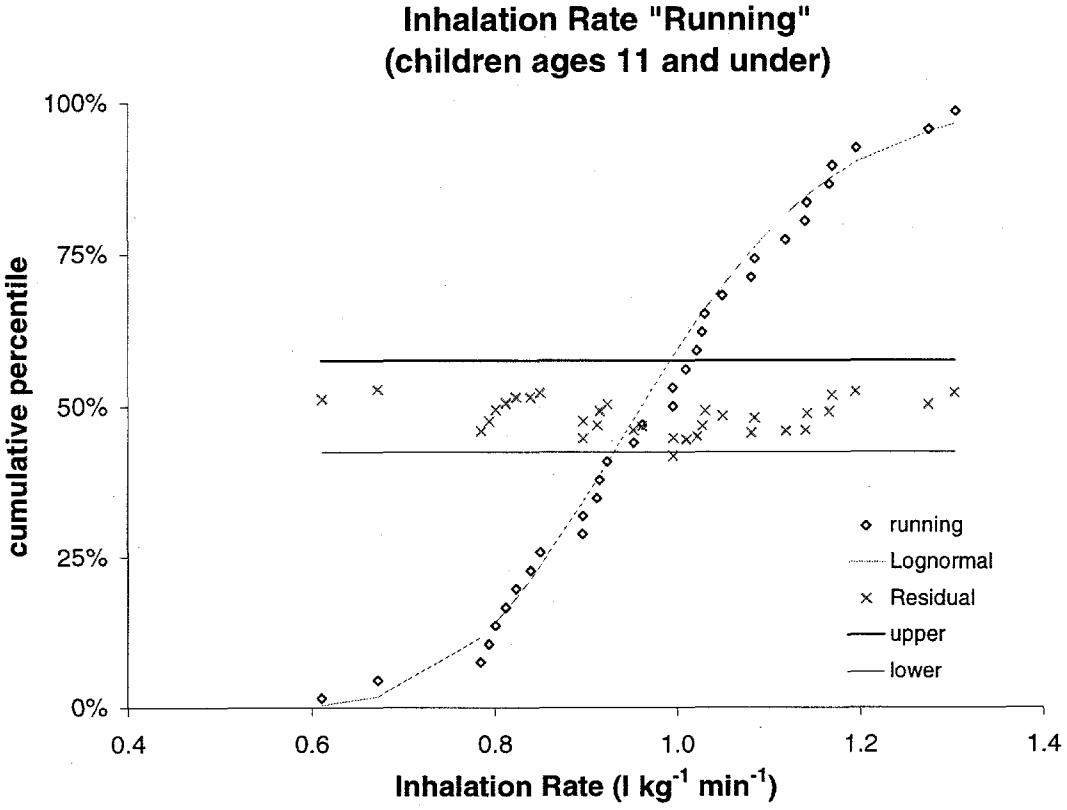


Figure 7.6: Plot of the ECDF for Children under 12 years of age in a "running" activity along with the fitted Lognormal distribution and residuals.

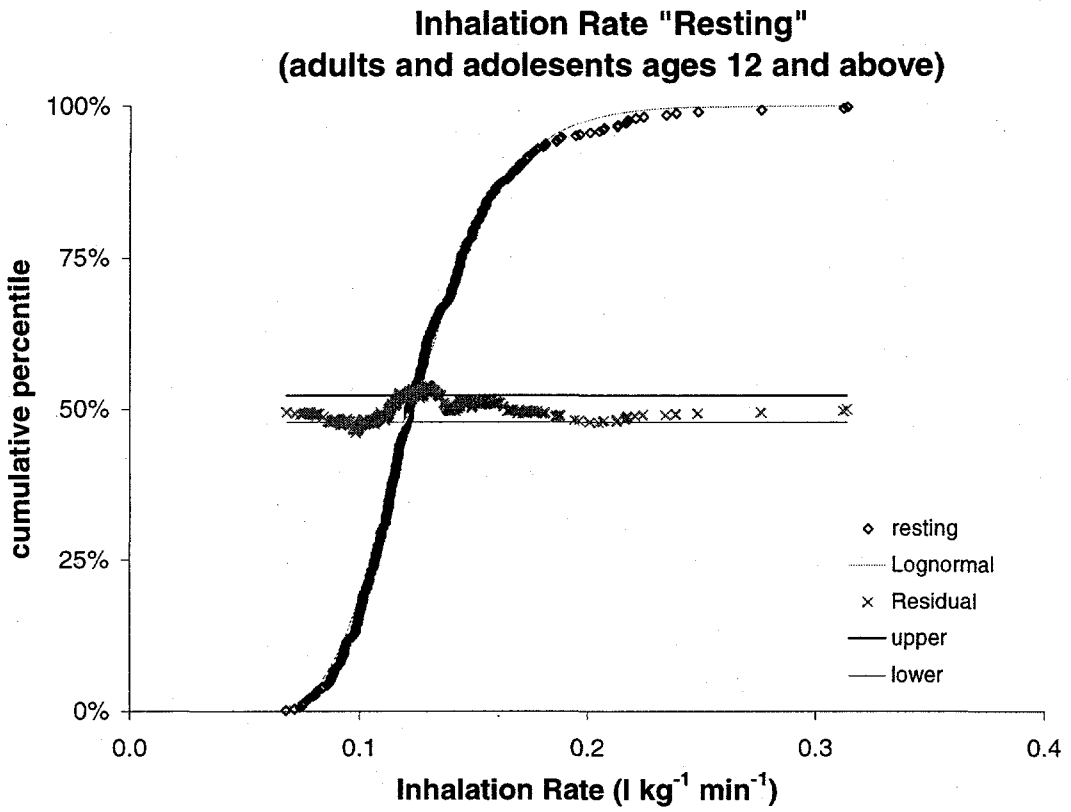


Figure 7.7: Plot of the ECDF for Individual 12 years of age and older in a "resting" activity along with the fitted Lognormal distribution and residuals. The upper tail has a similar shape as that of the children in a "resting" activity but again, the small deviation from the does not warrant further decomposition of the data.

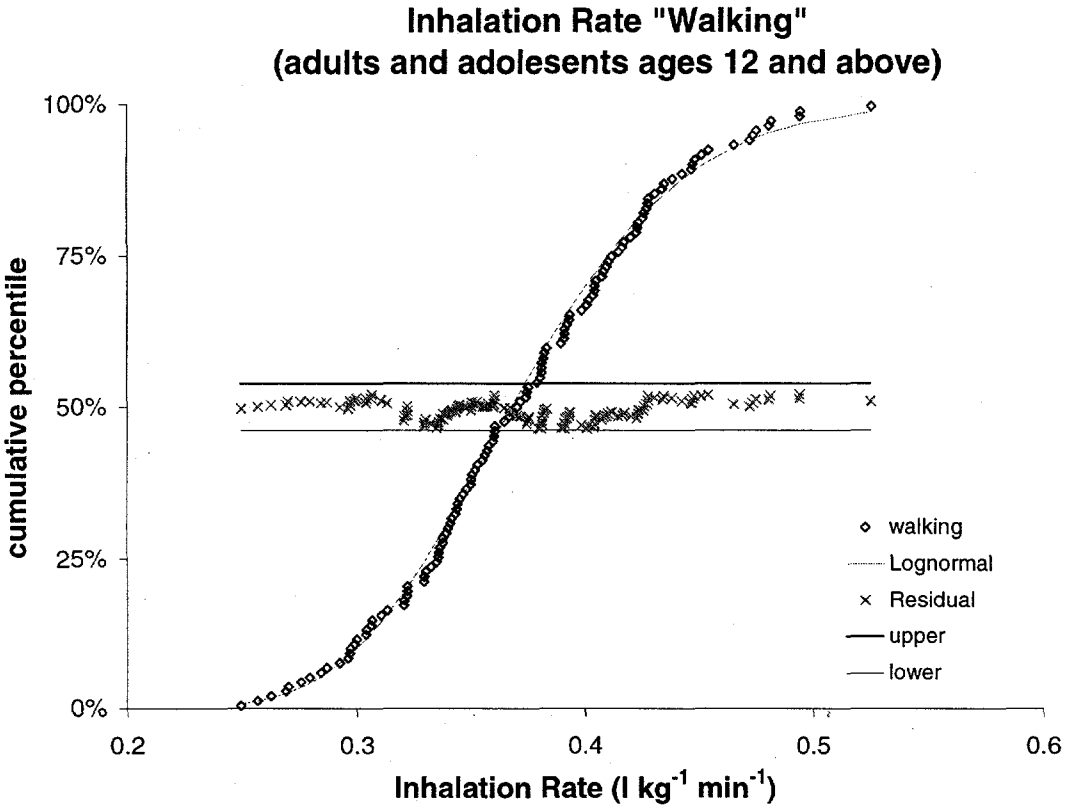


Figure 7.8: Plot of the ECDF for Individuals 12 years of age and older in a "walking" activity along with the fitted Lognormal distribution and residuals.

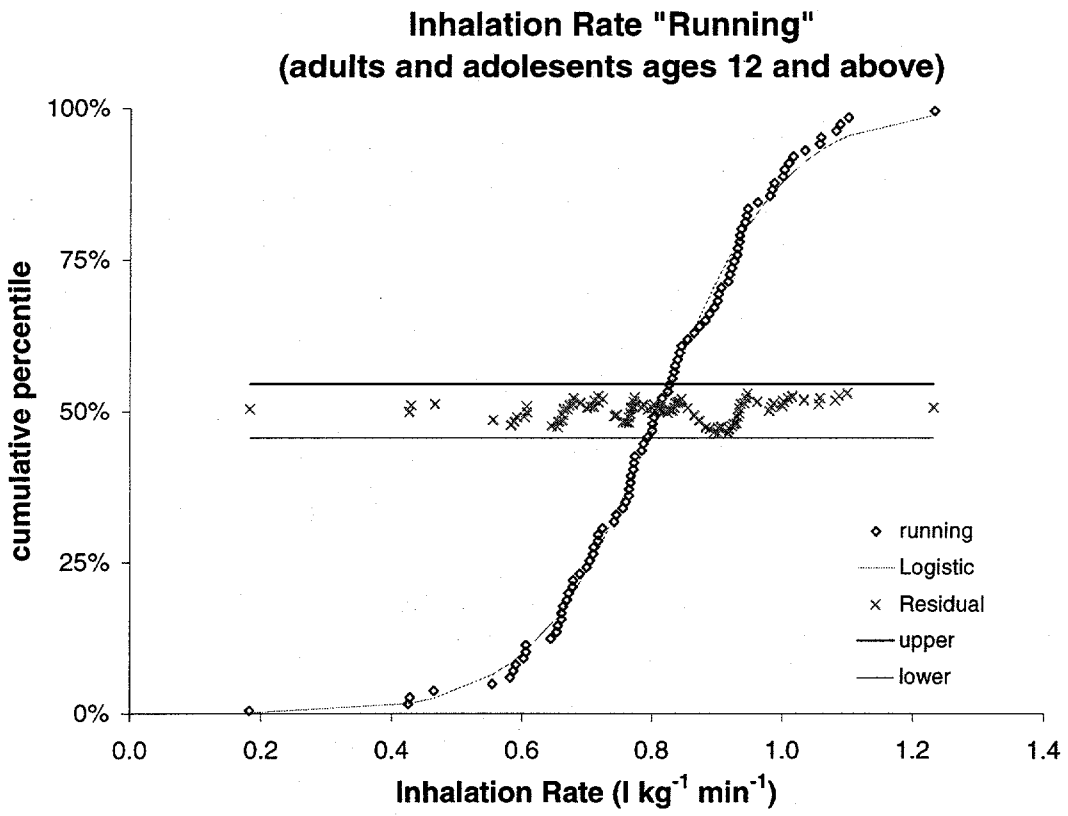


Figure 7.9: Plot of the ECDF for Individuals 12 years of age and older in a "running" activity along with the fitted Logistic distribution and residuals. Although a theoretical basis for the Logistic model was not apparent, the model performed much better than any of the other parametric models tested.

Table 7.2: Initial Selection and Parameterization of Models for IR

data description	Distribution	n	location[a]	scale[b]
Children under 12 years of age resting	Lognormal	138	0.29	0.09
Children under 12 years of age walking	Lognormal	46	0.56	0.10
Children under 12 years of age running	Lognormal	33	0.98	0.17
Individuals 12 years and older resting	Lognormal	378	0.13	0.03
Individuals 12 years and older walking	Lognormal	125	0.37	0.06
Individuals 12 years and older running	Logistic	93	0.81	0.09

(a) the arithmetic mean for the lognormal model and the logistic model

(b) the arithmetic standard deviation for the lognormal and the scale parameter for the logistic model

7.5 Uncertainty and Variability in the Inhalation Rate Exposure Factor

The greatest source of analytical uncertainty in the distributions of inhalation rate is due to the relatively small sample sizes use to construct the distributions. The uncertainty can be visualized by the width of the 95% confidence interval around the residuals in Figures 7.4 through 7.9. A two dimensional analysis is likely appropriate for this exposure factor. One way to reduce uncertainty is with a series of cross validation experiments. However, data is not readily available for such an experiment at this time.

Qualitative uncertainty in the distribution arises from the fact that the inhalation rates are highly dependent on activity and that the measurements used to develop the distributions are from laboratory experiments rather than actual field studies. Another weakness is that there is a physiological reason for women to have a different inhalation rate than men yet this was not apparent in the small sample used to develop the distributions in this report. Although it is not clear how well the distributions can apply to individual demographic subgroups within the population (race, gender), the study used here was well designed and efforts were taken to construct a sample that was representative of the population as a whole (CARB, 1993).

There is a significant amount of qualitative uncertainty associated with the distributions for children under 12 years of age across all activity levels. Overall, these distributions do a good job representing the combined data for all children. However, it is clear that the inhalation rate for young children is inversely related to the age of the individual and as a result, the distributions will likely under predict IR for the younger members of the group and over predict IR for the older members. A larger sample size is required to identify the where to split the data to best capture this age dependence (as was done for body weight).

7.5 Scores for the inhalation rate distributions

The data used to construct the distributions for inhalation rate were of high quality and were relevant to the three basic activity levels (resting, walking and running) used in the study. A limited amount of information is available for determining the relationship between the basic activity levels actual activities performed throughout the day.

The experimental design produced reliable direct measurements of IR and the selected parametric models do a good of representing the data. However, the small sample size results in a significant amount of analytical uncertainty for some of the distributions.

The theoretical basis for using Lognormal distributions to describe inhalation rate data is probably sound (no negative values, long upper tails). The basis for the logistic used to fit the distribution of IR for "running" individuals 12 years of age and older is unclear. The visual performance of the parametric models was good. Analytical goodness-of-fit scores and the ability

of the recommended distribution to forecast samples in cross validation experiments was not measured. As a result, the final robustness scores for the EF distributions are low (L) to medium (M) as reported in Table 7.3.

Table 7.3: Robustness scores for Inhalation Rate distributions

data description	Robustness score
Children under 12 years of age resting	L
Children under 12 years of age walking	L
Children under 12 years of age running	L
Individuals 12 years and older resting	M[a]
Individuals 12 years and older walking	M
Individuals 12 years and older running	M

a. the model for individuals 12 years and older "resting" was on the high end of the medium score. The possibility of a mixture model in the upper tail precluded the highly applicable score.

As with most of the exposure factors considered in this report, the best way to improve upon the robustness scores is to demonstrate that the parametric models perform well with independent data.

8.0 Development of PDFs for Water Consumption Rates

In this section we report on the development of PDFs for water intake. Estimating exposure to hazards in drinking water requires knowledge about the amount and the source of water that the exposed individual consumes. Water is either consumed directly as drinking water or indirectly through ingestion of food and beverages that have been prepared with drinking water. Several well-designed surveys have been completed and extensive data are available with information about the intake of water. However, the questions included in the national surveys do not specifically address the consumption of tap water in all forms (direct from tap and indirect through food or beverage). Rather, these surveys include information about the amount of tap water consumed as drinking water and the type and amount of food/beverages ingested. Another limitation of available data is that information pertaining to the source of drinking water (tap, bottled, at or away from home) is qualitative.

To estimate drinking water consumption, previous authors have created and used databases for estimating the amount of tap water used in the preparation of each food/beverage based on standard recipes or on directions provided with packaged or canned foods (Ershow and Cantor, 1989; Levallois et al., 1998). Using these databases in combination with the reported values for food ingestion, the authors of both studies estimated the amount of drinking water consumed directly or through the ingestion of food and beverage.

Food coding databases and recipe databases are available from the US Department of Agriculture along with results from the national surveys. However, a critical analysis of all food intake data is beyond the scope of this report. To demonstrate the method described in this report, we consider only the "total water" intake from all sources including tap water, extrinsic water (added to food/beverage) and intrinsic moisture in food and beverage.

8.1 Sources of data

The Continuing Survey of Food Intake by Individuals 1994-96 (CSFII 1994-96) was used to develop distributions for total water intake levels. The total water intake includes drinking water consumed directly, drinking water added to food and intrinsic moisture in food products (i.e., moisture in fresh fruit). The CSFII was the 10th national food consumption survey conducted by the USDA. It contains information on the sources of water used for cooking purposes, in preparing beverages and as plain drinking water. The survey participants were asked at an interview how much plain drinking water they drank in the previous 24-hour period. The amount of water included in foods and beverages as well as the amount of plain water drunk yesterday was used to estimate the total water consumed in a 24-hr period. In record time 40 ("daily intakes: nutrients, fatty acids") the water intake [g] ("water" variable) was obtained,

which includes any water in food and beverages [g]; excluding plain drinking water. Added to the "water" variable on either day 1 or 2 was the corresponding amount of water (drunk yesterday, i.e., on day 1 and/or 2). Record type 25 also included qualitative information on the fraction of water that was drunk at home (all, most, some, none, or don't know), the day of intake (day of week) and month of intake and the type of "water from home" as either, tap water/drinking fountain, bottled water, other, or don't know, or not ascertained.

Each year of the CSFII 1994-96, a nationally represented sample (over-sampling of low income individuals) was selected and asked to recall their food intake over the past 24 hour period via interview. Two nonconsecutive days of food and nutrient intake data were collected for 16,108 individuals and the day 1 response rate was 80% in all 3 years of the survey. Other data collected, includes health related variables such as self assessed body weight (Section 4.1) and body mass index as well as the activity level of the survey participant during the time of recall (i.e., frequency of vigorous exercise and number of hours of television or videotapes watched yesterday).

Individual sample person variables included: age from 0 to 90 yrs where ages over 90 years was reported as 90, gender, race (white, black, Asian/Pacific, Native American, or Other), ethnic origin (i.e., Hispanic, including Mexican, Peurto Rican, Cuban, or Other Hispanic). The work status for all household members (>=15 yrs), and pregnancy or lactation status and breastfeeding status of children 3 years old or less. The region of US (Northeast, Midwest, South, and West) and whether urban/rural area (Metropolitan Statistical Area (MSA) (**)- central city, MSA-outside central city, or non MSA), and household income from previous year before taxes and income as a percentage of the poverty threshold are also given as well as data on the income from the previous month by source.

Another source of data is the NHANES III survey. For the NHANES III , the total water intake in previous 24 hour period was defined by summing the DRPQ2A and DRPNWATE variables for each sample person in NHANES III. DRPQ2A is defined as the "quantity of plain drinking water reported either in total fluid ounces per day or by specifying the number of glasses of water and the volume per glass using standardized measurement aids". All responses for DRPQ2A were converted to fluid ounces. If the respondent answered "None," meaning that no plain drinking water is usually consumed, the amount of water was reported to be 000 fluid ounces; other quantities of plain drinking water were recorded as xxx fluid ounces. The volume of plain drinking water is in addition to water found in foods and beverages. Water from foods and beverages is included in the variable DRPNWATE [g]. Plain drinking water and spring water usually were excluded from the dietary recall unless beverages were diluted with plain water or water was a component of a combination food that was reported by components such as a homemade fruit and water drink.

During the dietary interview, information collected from the 24 hour recall for all sample persons and the food frequency for ages 12-16, was automatically recorded by an automated computer system (NCC, 1992). If the SP was less than 12 years old, a proxy interview was conducted with a parent or guardian. The Nutrition Coordinating Center (NCC) at the University of Minnesota is responsible for the design of the automated Dietary Data Collection (DDC) system and construction of the foods databases. The development of the DDC was supported by the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI) the National Center for Health Statistics (NCHS), and the Food and Drug Administration. Sample questions pertaining to water intake include: "How much plain water do you usually drink in a 24 hour period? (include only tap water and spring water)?".

8.2 Data Classification and Distribution Analysis

Prior to analyzing the water consumption data, the two-day average values for plain drinking water intake and the two day average values for total moisture intake (intrinsic and extrinsic) from food and beverage excluding plain drinking water were combined. Sample persons with missing values were removed from the analysis. These include missing values for "day two" consumption, zero values for intrinsic moisture intakes on day two, zero values for water intake and missing body weight values. The CART analysis was set for regression tree with v-fold cross validation (n=10) and the minimum cost tree was generated using the least squares method.

Initial analysis indicated that body weight contributed significantly to variability in the population. Therefore, prior to the CART analysis, the water intake was normalized to body weight ($l\ kg^{-1}\ day^{-1}$). The affect of normalizing the data to body weight is illustrated in Figure 8.1. As with the inhalation rate data, the water intake is inversely related to age for children under 12. Do to this strong dependence of intake on age for children, individuals less than 11 years of age were excluded from the CART analysis. For the younger ages, the data can be split into appropriate subgroups. These splits can either be selected to match the BW subgroups or taken from a CART analysis of the data for children. Our preliminary results (not shown) indicate optimal splits for children under 12 years of age at 8 months, 18 months, 3.5 years, 6.5 years and 12 years. It must be noted however, that these subgroups would necessary be averaged over whatever age categories are used.

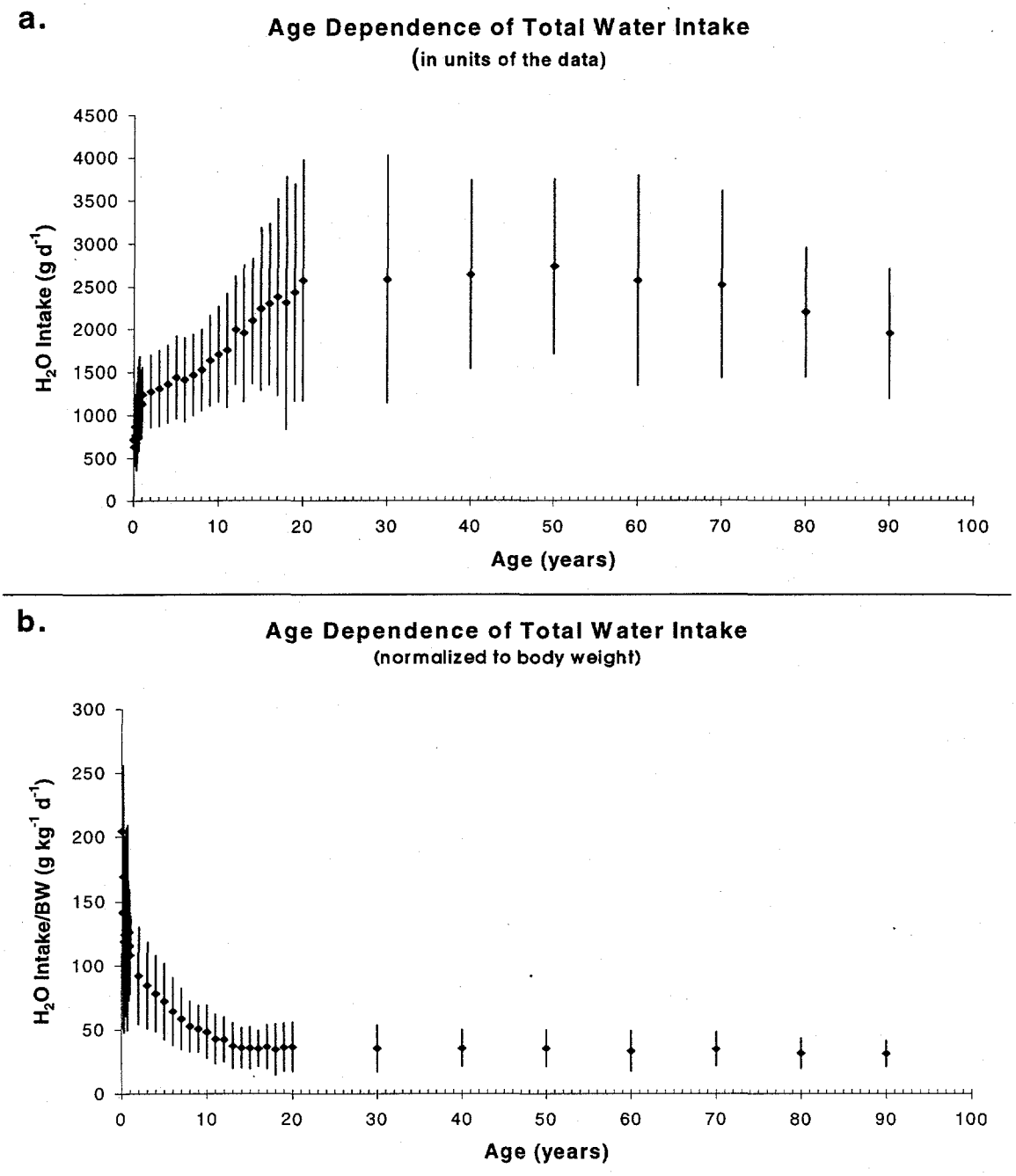


Figure 8.1: Plot (**a**) illustrates the age dependence for water intake plotted in the units of the data. Intake increases until about 20 years of age, remains relatively constant to 70 years then decreases with age. When the data is normalized to body weight as shown in plot (**b**), consumption is relatively constant above 12 years of age. Only decade values are shown for sample persons older than 20 years. The error bars indicate one standard deviation for values included in each yearly bin.

Results from the CART analysis for individuals older than 11 years of age are shown in the tree diagram in Figure 8.2. Interestingly, the main split in the data was on "Region". The data indicate that individuals in the Northeast and the South consume less water than those living in the Midwest and West. This split may be due to factors not included in the analysis (i.e., environmental factors such as temperature) but this could not be confirmed. The data also indicate a dependence on race where Black and Native American (in the Northeast and South) ingest less water on a body weight basis than the remainder of the population.

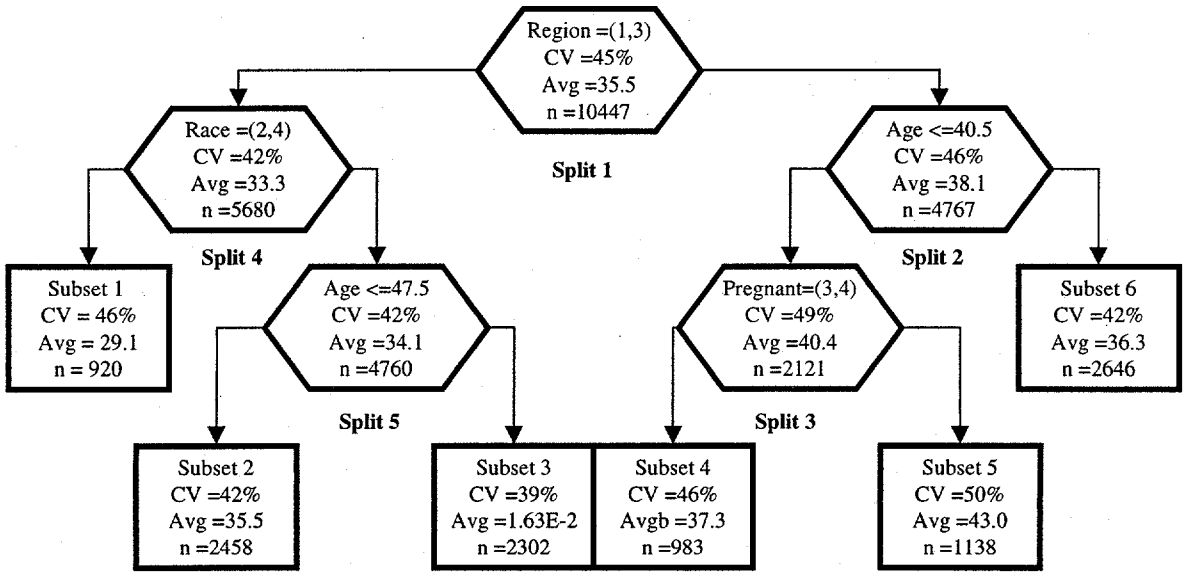
The split on pregnant and lactating status (split 3) separates all "non-pregnant" women from males, indicating females under 41 years of age consume less water than males in the same region. This also implies that pregnant women consume more water than non-pregnant women do (Ershow et al., 1991) although the CART analysis cannot distinguished them from males in the same region and age group. Depending on the analysis objective, one would probably want to develop a separate distribution for pregnant and lactating women (Burmaster, 1998).

Each of the data sets from the CART analysis are plotted as empirical distributions in Figure 8.3. The results indicate that three of the distributions can be recombined without significant loss of information. One could combine distribution (2) "White, Asian/Pacific Islander, Other living in the Northeast and South" with "Non-pregnant women living in the West and Midwest" and "Individuals ≥ 41 years of age in the West and Midwest".

Table 8.1: Composition and Summary Statistics for Final Water Intake Data Sets

Characteristics of the Subgroups	n	Ave.	CV
All data	10447	35.5	45%
1. Black and Native American in Northeast and South	920	29.1	46%
2. White, Asian/Pacific and Other in Northeast and South (age<48 y)	2458	35.5	42%
3. White, Asian/Pacific and Other in Northeast and South (age≥48 y)	2302	32.6	39%
4. Non-pregnant women in West and Midwest (age<41y)	983	37.3	46%
5. Men and pregnant women in West and Midwest (age<41y)	1138	43.0	50%
6. Individuals in West and Midwest (age≥41 y)	2646	36.3	42%

CART Output for Water Intake Normalized to Body Weight (g kg^{-1} d^{-1})
(all sample persons 12 yrs. and older)



Legend
CV = percent coefficient of variation
Avg = average
n = sample size
Variables definitions
Region (1=northeast, 2=midwest, 3=south, 4= west)
Age (continuous yearly values greater than 11)
Race (1=white, 2=black, 3=asian/pacific islander, 4=native american and 5=other
Preg (1=pregnant, 2=lactating, 3 =pregnant and lactating 4 = not pregnant or lactating, 5=not female)

Figure 8.2: Classification and regression tree showing the decomposition of the original data set for water intake rate normalized to body weight. Sample persons younger than 12 years of age are excluded from the analysis (see text for explanation). The next data split would occur on Subset 6 splitting individuals living in urban from those living in rural areas.

8.3 Presentation of distributions

The output from the CART analysis was used to construct individual data sets for each demographic group described in Table 8.1. ECDFs for each resulting demographic region of the sample are illustrated in Figure 8.3. Figure 8.3 includes all sample persons (over the age of 11 years) subdivided into the compositions defined in Table 8.1. Figure 8.3 indicates that three of

the datasets can be recombined without significant loss of information. These include "Individuals who are not Black or Native American living in the Northeast and South", "Non-pregnant women living in the West and Midwest" and "Individuals ≥ 41 years of age in the West and Midwest". However, The systematic determination of which distributions can be recombined is the subject for future work (See section 4.5). Parametric distributions where assigned to each of the data sets listed in Table 8.1.

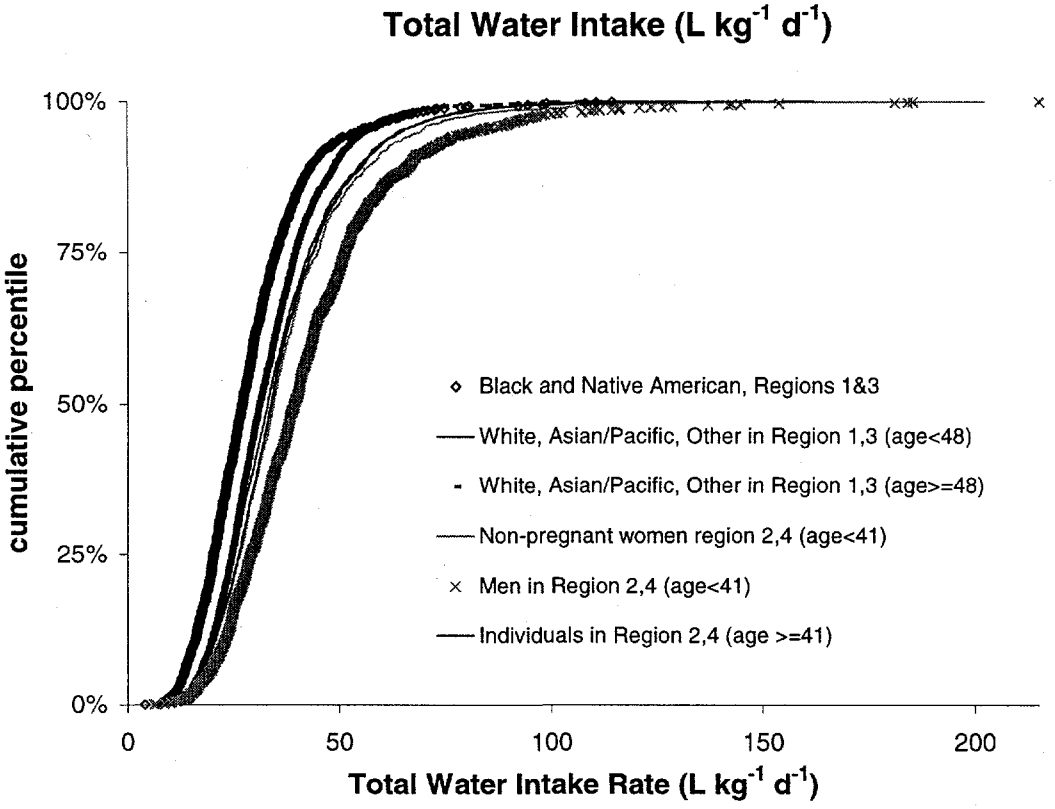


Figure 8.3: Plot of all empirical distributions resulting from the datasets identified in the CART analysis. The three distributions that are plotted as lines can likely be recombined without loss of information.

Distributions for each demographically independent data set in Table 8.1 are presented in Figures 8.4 through 8.9. Distributions for children less than 12 years of age are not included. Although the Extreme Value and the Lognormal distributions performed equally well in most cases, the Lognormal was selected because of its theoretical basis (values for water intake must

be greater than zero) and because of its ease of use. Each of the distributions are summarized in Table 8.2.

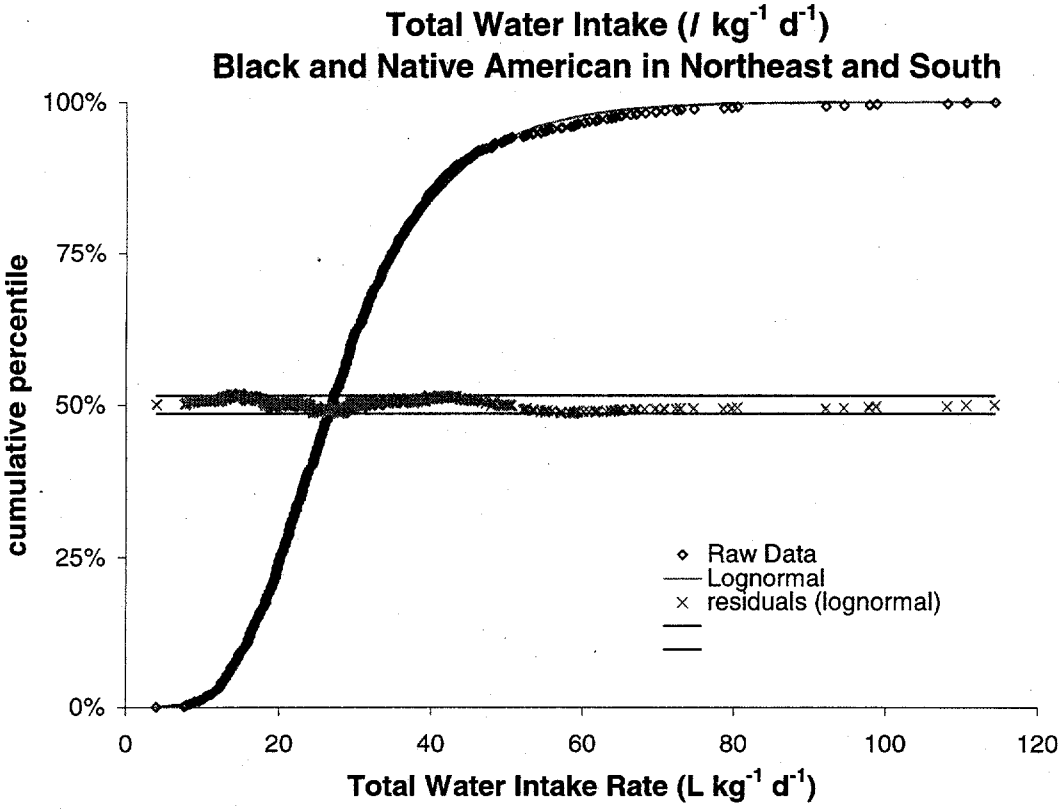


Figure 8.4: Distribution for total water intake for Black and Native American individuals living in the Northeast and South plotted along with the fitted Lognormal, the residuals between the parametric model and the data, and the 95% confidence interval for the residuals. The parametric model slightly over predicts water intake in the upper tail but the results are still within the expected range of precision (confidence interval around residuals).

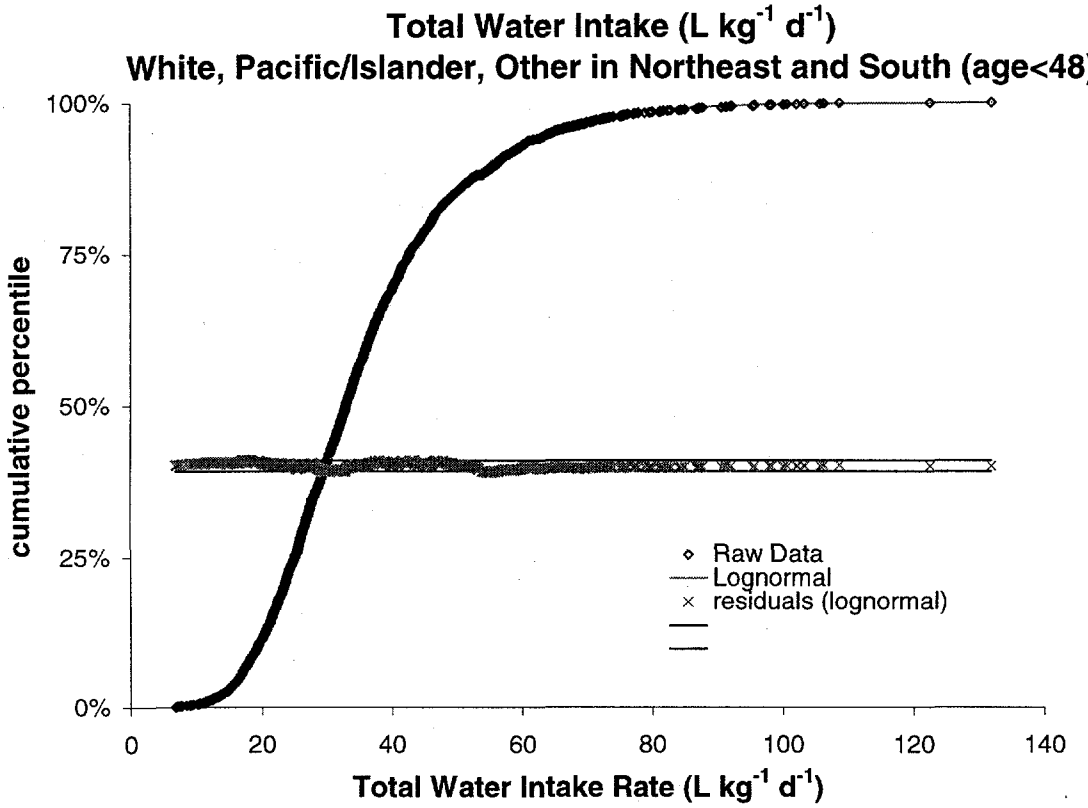


Figure 8.5: Distribution for total water intake for individuals who are not Black or Native American living in the Northeast and South and are less than 48 years of age. The empirical cumulative distribution is plotted along with the fitted Lognormal, the residuals between the parametric model and the data, and the 95% confidence interval for the residuals.

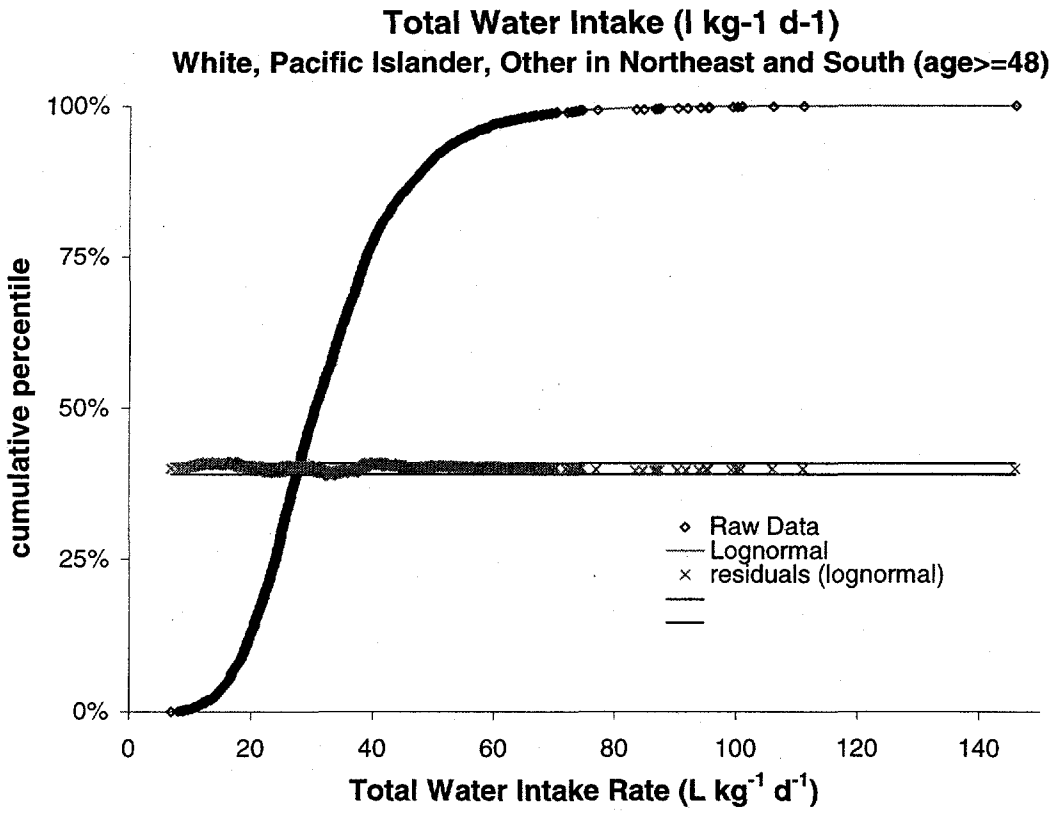


Figure 8.6: Distribution for total water intake for individuals who are not Black or Native American living in the Northeast and South who are greater than or equal to 48 years of age. The empirical cumulative distribution is plotted along with the fitted Lognormal, the residuals between the parametric model and the data, and the 95% confidence interval for the residuals. The slight lack of a smooth transition in the ECDF between the 25th percentile and the 90th percentile indicate a possible mixture model but the confidence interval of the residuals does not warrant increasing the complexity of the parametric model.

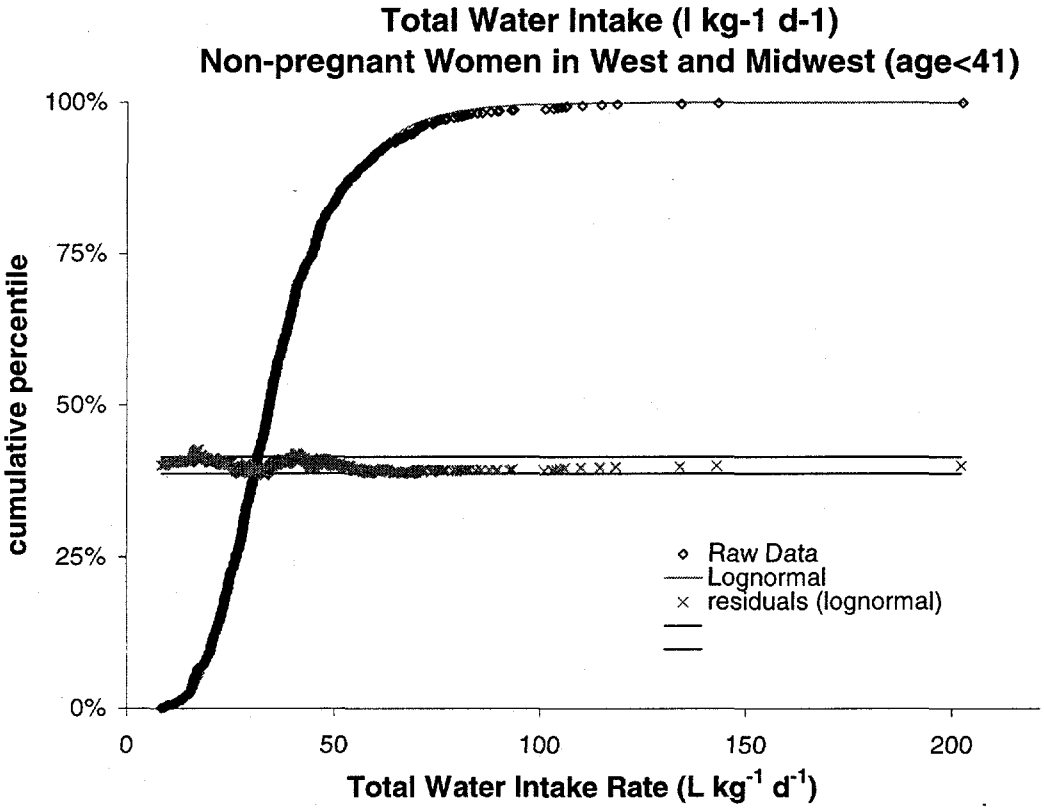


Figure 8.7: Distribution for total water intake for non-pregnant women in the West and Midwest who are less than 41 years of age. The empirical cumulative distribution is plotted along with the fitted Lognormal, the residuals between the parametric model and the data, and the 95% confidence interval for the residuals.

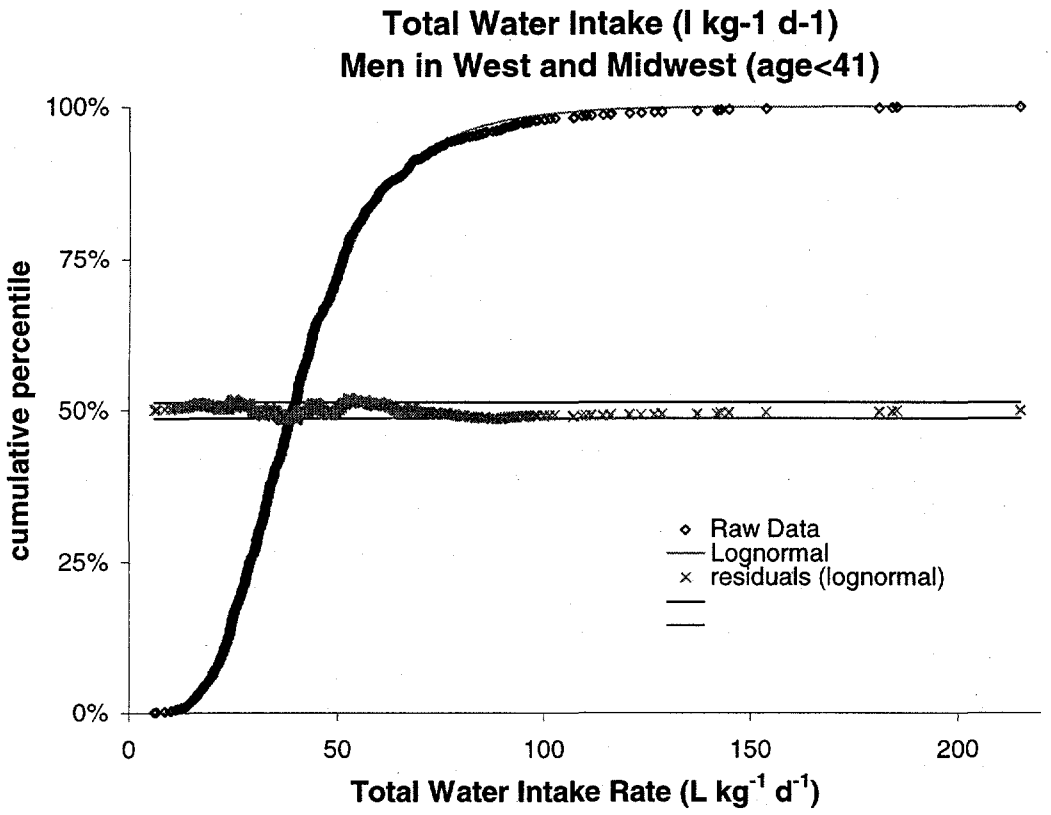


Figure 8.8: Distribution for total water intake for men living in the West and Midwest who are less than 41 years of age. The empirical cumulative distribution is plotted along with the fitted Lognormal, the residuals between the parametric model and the data, and the 95% confidence interval for the residuals. The parametric model slightly over predicts water intake in the upper tail but the results are still within the expected range of precision (confidence interval around residuals).

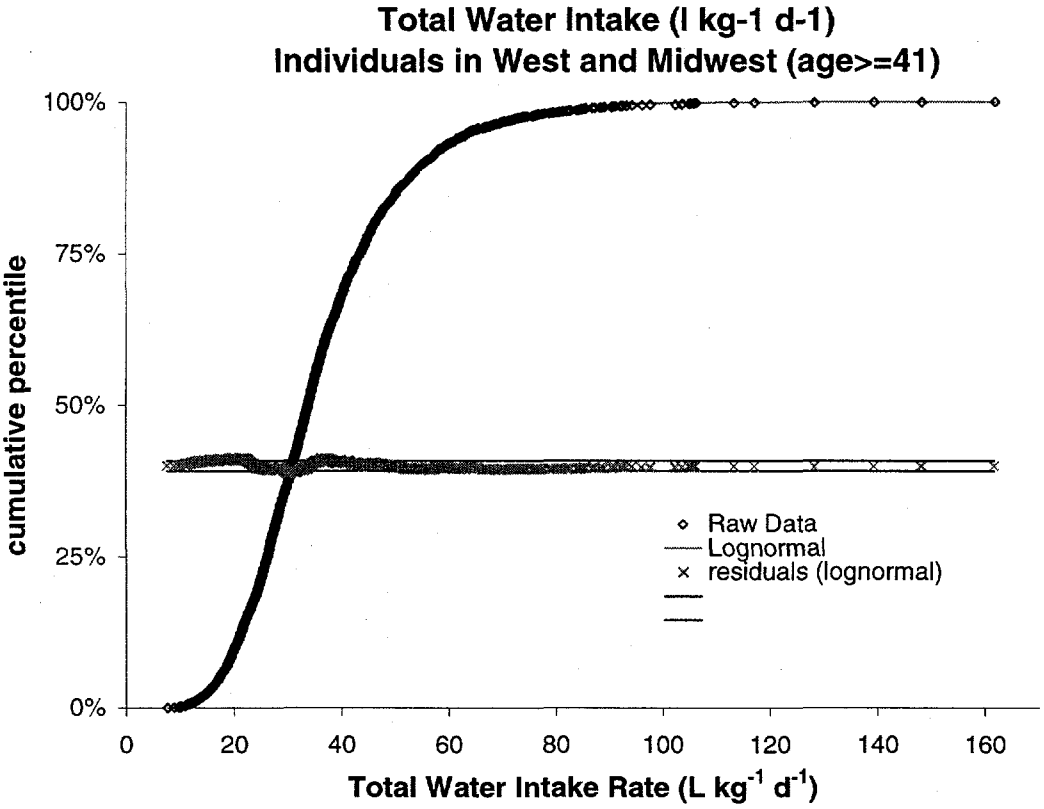


Figure 8.9: Distribution for total water intake for individuals living in the West and Midwest who are 41 years of age and older. The empirical cumulative distribution is plotted along with the fitted Lognormal, the residuals between the parametric model and the data, and the 95% confidence interval for the residuals.

Table 8.2: Parameterization of Models Selected for Total Water Intake

data description	Distribution	n	location[a]	scale[b]
1. Black and Native American in Northeast and South	Lognormal	920	29.11	13.08
2. White, Asian/Pacific and Other in Northeast and South (age<48 y)	Lognormal	2458	35.55	15.46
3. White, Asian/Pacific and Other in Northeast and South (age≥48 y)	Lognormal	2302	32.57	12.80
4. Non-pregnant women in West and Midwest (age<41y)	Lognormal	983	37.23	16.48
5. Men in West and Midwest (age<41y) [c]	Lognormal	1138	42.93	20.33
6. Individuals in West and Midwest (age≥41 y)	Lognormal	2646	36.32	15.09

a. The arithmetic mean for the lognormal model

b. The arithmetic standard deviation for the lognormal

c. Pregnant women are included in the distribution for men (age < 41 years) in west and midwest. Inclusion or exclusion of pregnant women from this set does not change the distribution.

8.4 Uncertainty and variability in the ingestion-rate distributions

Because of the large sample sizes for each data set, analytical uncertainty in the distributions of water intake is not expected to contribute significantly to variance in this exposure factor. The small analytical uncertainty is indicated by the narrow 95% confidence interval around the plot of residuals. A two dimensional analysis of uncertainty is not necessary on the basis of analytical uncertainty alone.

However, qualitative uncertainty in the data used to develop these distributions may be significant. The data used to develop the distributions in this section are based on 24-hour recall data. Although the approach was validated in previous studies (Ershow and Cantor, 1989), a recent study found a possible bias in recall data. More water was reported as consumed over a 24-hour period on recall than from diary data (Levallois et al., 1998). However, the sample size used in this study was relatively small (n=125). To our knowledge, the inconsistency between 24-hour recall data and diary data has not yet been resolved.

There is also a significant amount of qualitative uncertainty about the reliability and relevance of food intake data converted to drinking water intake. Critical evaluation of the data bases used to estimate the amount of drinking water used in the preparation of food and beverage products was beyond the scope of this study. In addition, the qualitative nature of questions used in the survey lead to uncertainty about the source of drinking water (tap or bottled).

All of the distributions presented in Figures 8.4 through 8.9 had strong outliers in the upper tails of the data with Z-scores greater than 6 and occasionally greater than 8. It is often

recommended that Lognormal distributions of drinking water intake be truncated to prevent such outliers. However, *the data presented in this section does not support truncating the data.* Rather, it demonstrates that uncharacteristically large values are indeed possible and should be included in the analysis unless additional information is available that can show that these excessively large values are not real.

8.5 Scores for the ingestion-rate distributions

The data used to construct the distributions for water intake were of high quality. The Continuing Survey of Food Intake by Individuals has been adapted and improved over several decades. As a result, the data collected in the survey is highly representative of the population and the sample size is more than adequate even when separated in to demographic subset. However, the relevance of the data is somewhat questionable. There is limited evidence that recall data may overestimate the amount of water ingested and there is uncertainty in the authors mind about the approach and assumptions used to estimate indirect drinking water consumption from self reported food intake data. Thus, the values used to estimate water intake must be considered both self-reported and surrogate.

The theoretical basis of the Lognormal distribution is acceptable and both the goodness of fit and visual performance across the range of data for all distributions was excellent. However, there is still a question as to whether or not the distributions should be truncated. Removal of the outliers (values greater than 6 standard deviations from the mean) did not change the parametric distribution. However, the question is not whether the outliers will change the distribution. Rather, the question is whether one should use the occasional extreme value generated by the Lognormal distribution when performing a probabilistic risk assessment. Extreme values occur regularly in the samples of water intake and efforts to determine if these outliers could be removed from the data set were inconclusive.

Although the quality and quantity of data are extremely high and the parametric models do an excellent job fitting the data, concern about the relevance of the food intake data to water consumption and lack of information about the source of drinking water result in a score of medium (M). These are summarized below in Table 8.3.

Table 8.3: Robustness scores for Total Water Intake distributions

data description	Robustness score
Black and Native American in Northeast and South	M
White, Asian/Pacific and Other in Northeast and South (age<48 y)	M
White, Asian/Pacific and Other in Northeast and South (age≥48 y)	M
Non-pregnant women in West and Midwest (age<41y)	M
Men in West and Midwest (age<41y)	M
Individuals in West and Midwest (age≥41 y)	M

The robustness scores that are reported in Table 8.3 are specific to Total Water Intake. Total water includes plain drinking water, drinking water in food and beverage and intrinsic water in food/beverages. If we were to apply these distributions to drinking water they would clearly need to be modified. Drinking water accounts for approximately 50% to 60% of total water (Ershow and Cantor, 1989). Attempts to modify the distributions to represent only drinking water would likely lead to a significant penalty in the score of robustness reducing most of the distributions from medium (M) to low (L).

9.0 Conclusions, Findings and Recommendations

Exposure assessments use a number of factors that are both variable and uncertain. As a result, the magnitude of these factors can not accurately be represented by a single value in a risk assessment. A range of values reflecting both the population variability and the uncertainty that results from limited and imprecise data must be used to characterize these exposure factors. Methods are readily available for developing distributions for exposure factors when relevant data exists. Although standard goodness of fit techniques can be used to test the performance of a distribution in mapping or fitting existing data sets, little effort has gone into developing a method for scoring the expected performance of these models in new applications.

In an effort to develop a practical and reliable method for evaluating the performance and robustness of PDFs, LBNL has collected and critically evaluated data for the following exposure factors:

- body weight
- exposure duration (amount of time living at a residence)
- exposure frequency (fraction of the day spent at the exposure location)
- inhalation rates and
- total water intake

For each of these exposure factors available data was critically reviewed and analyzed to identify important demographic regions of the population. The original data sets were decomposed into these regions and PDFs were constructed for each demographic subset of the population. The most appropriate distribution for each subset was selected based on a combination of standard procedures and on a simple graphical method. Lessons learned during the data collection, evaluation and distribution development process were used to design a scoring system based on the quantity, quality and relevance of the data and on our ability to identify a parametric model (or other distributional form where appropriate) that adequately describes the data.

9.1 Findings

The validity or quality of the PDFs that are used in a probabilistic exposure analysis directly influence the reliability of the decisions that are based on the outcome of the analysis. Many times default distributions are prescribed by regulatory agencies, consulting organizations or in the peer review literature. In such situations, there is a risk that policy guidelines can be looked on as fact. Default values need to be clearly represented as to their quality or appropriateness for various exposure scenarios.

PDFs are developed from data sets and there are a number of methods for making the best fit of a distribution to the data. When these methods are applied, one obtains a distribution that provides an optimum fit to the data used in the analysis. However, once this process is completed, the resulting distribution does not provide the user of that distribution with a quantitative measure of how well the distribution replicates either the underlying data or the true variability of the exposure factor being represented. What is needed to address these issues is some measure of the quality, reliability and relevance of the distribution as it relates to the current application.

The scoring procedure introduced in this report is a questionnaire designed to combine quantitative and qualitative information about the data and distribution into a single scenario-specific measure for the quality of a given parametric model (or other form of distribution). Although the final scores fall on a continuum from **not applicable** to **highly recommended,** the continuum is partitioned into four basic regions defined as Highly recommended for use (H), Medium (M), Low (L) and Not Applicable for use (NA). The questions are designed to elicit information about:

- The quantity of data used to construct distributions,
- Relevance of the data (actual measurement, self-reported or surrogate value),
- Analytical goodness of fit for standard distributions,
- Theoretical basis for standard distributions,
- Visual performance of the model across the range of data including the percentiles of greatest interest to the particular analysis objective,
- Extent to which variability and uncertainty can be represented, is the amount of measurement or reporting error known, and
- Ability of the recommended distribution to forecast samples from independent but related surveys and/or data sets.

Although the final form of the questionnaire and scoring system should come from extensive open debate among experts from a wide range of disciplines, an initial format has been developed from the above list of criteria and demonstrated in this report.

Some of the criteria in the questionnaire are quantitative where the value given is dependent on an actual measurement of sample size or fit. For other criteria such as data quality the score falls on a continuum from very poor to very good. To assign a score to these criteria, it is essential that the user become familiar with all facets of the data. The more intimate a person is with a given data set, the more qualified that person is for judging the quality of the data for a given task.

9.1.1 Score for Body Weight

The body-weight distributions score mostly high as a result of an abundance of representative and directly relevant data that is well described by standard distributions. Some subsets of the body-weight data sets—age, gender, etc. groupings—score medium on the robustness scale because the sample sizes are small among certain demographic regions of the population. In addition, no cross-validation experiments have yet been run with the distributions. No effort was made to reduce the number of demographic regions or subgroups of body weight. These distributions are expected to apply equally well across the population for any well-defined exposure scenario and probabilistic risk assessment.

9.1.2 Score for Exposure Duration

Exposure duration can be defined in a number of ways. For this report, we assume that the hazard originates at or near the home and ED is defined as the length of time that and individual is expected to remain in their current residence. Exposure duration is estimated from reported values for "current residence time" or the amount of time that an individual has occupied his/her current home. Data used to construct the distributions for exposure duration were representative of the national population and subgroups within the population although it was no clear whether Native Americans living on tribal land were included in the analysis.

The parametric distributions presented in the previous studies do an excellent job of representing the data that was used to generate them. However, the use of a surrogate values to predict ED and the limited effort to identify significantly different subgroups within the population lead to a recommended robustness score of low to medium (L-M). This distribution clearly needs further consideration.

9.1.3 Score for Exposure Frequency

The definition of Exposure Frequency is also scenario specific. Knowing the fraction of the day that an individual spends performing an activity at a given location is critical when assessing exposure through multiple pathways. For this report, EF is generally defined as the fraction of the day spent indoors at home. The data used to construct the distributions for exposure frequency were highly representative of the population and the parametric distributions developed for EF do a good job representing the data. However, the relevance of information from short-term diary data is of concern. As a result, the distributions of EF receive a score of low to medium (L-M).

9.1.4 Score for Inhalation Rate

Inhalation rate is strongly dependent on activity and as a result, a single estimate of inhalation rate is not feasible. The data selected for use in this report were from a small but

representative study that included measurements collected during field activities and during controlled laboratory activities. Only results from the laboratory portion of the study were used here. The parametric distributions selected for use do a good job of representing the data. However, the sample sizes for directly measured inhalation rates is too small to allow analysis of demographic/physiological difference in inhalation rate and time constraints precluded a critical analysis of the relationship between inhalation rate and activity. As a result, we assign the distributions of inhalation rate a score of medium (M) with the caveat that they are only appropriate for the three standard activity levels performed in the controlled laboratory study. The applicability of these distributions to an actual analysis would likely result in a moderate to heavy penalty against the score.

9.1.5 Score for Water Intake

Exposure to water borne contaminants is dependent on both the amount and the source of water consumed by the exposed individual. A large amount of quality data is available for constructing distributions for water intake. Although the data is highly representative of the population it is not clear how relevant the data is for estimating water consumed from a particular source (home drinking water). For this report, water intake was defined as "total water", which includes direct tap water, indirect tap water and intrinsic water ingested with food and beverage. Although the quality and quantity of data are extremely high and the parametric models do an excellent job fitting the data, concern about the relevance of the food intake data to water consumption and lack of information about the source of drinking water result in a score of medium (M) for water intake.

9.2 Recommendations

(i) To score the exposure factor distributions, it is critically important that the user have a clear and complete understanding of (1) the data used to develop the distribution in question, (2) the procedure used to construct the distribution and (3) the population that the distribution will be used to represent. Whether this understanding comes from developers of the distributions, the user of the distributions or a combination of the two is not readily apparent but clear documentation of all phases of the process are critical.

(ii) When a large amount of data is available, CART is an efficient and effective tool for identifying the optimum way to split complex data along demographic lines. Further splits in the data may be necessary for political or policy reasons but that is beyond the scope of this report.

(iii) Systematic methods for incorporating sensitivity/uncertainty analysis should be developed and use to determine when and to what degree the demographic subsets of data

identified by CART can be recombined into the minimum number of subsets in the family of distributions for each exposure factor.

(iv) Future work should be directed towards better understanding how to fit truncated distributions and how truncated distribution influences the calculation of dose/risk.

(v) Although not included in the body of this report, we found that model-free methods show promise as a tool for learning more about the underlying shape of distributions but more work is needed to determine just how useful they might be.

(vi) Neither set of currently available exposure duration (ED) distributions include information on ethnicity or socio-economic status. Such distributions could be determined by applying the analytical/statistical procedures of either Israeli and Nelson (1992) or Price *et al.* (1998) to the 1995 AHS-N data or the Monte Carlo procedure of Johnson and Capel (1992). In addition, the information was split on variables that were not found to be important in this analysis (gender, multiple age groups) and the strong relationship between young adults and children living at home was not accounted for. A reanalysis of the methods and data used to estimate ED and the construction of new distributions that can be tested using the methods introduced in section 3 is warranted.

(vii) All of the exposure factors included in this report can benefit from cross-validation experiments designed to test the performance of the parametric models (or other distributional forms) against independent data sets.

(viii) A better understanding of the relevance of short-term diary data for estimation of activity patterns and exposure frequency is warranted.

(ix) Direct measurements for inhalation rates are limited. Resources should be directed towards collection of quality data that can provide a better understanding of the physiological differences and inter-individual variability.

(x) The relevance and reliability of nutrition studies for estimating water intake should be verified using a series of small scale studies designed specifically for the estimate of source and amount of water consumed by various demographic subsets of the population.

References

Beals, JAJ; Funk LM; Fountain R; Sedman R. (1996) "Quantifying the Distribution of Inhalation Exposure in Human Populations – Distribution of Minute Volumes in Adults and Children" Environmental Health Perspectives, V104: 974-979.

Blaire, J (1995). "Estimating Exposure to Pollutants through Human Activity Pattern Data: The National Microenvironmental Activity Pattern Survey." Annual Report, Survey Research Center, U. of Maryland by J. Robinson.

BoC (1995) American Housing Survey for 1995, Current Housing Reports H150/95RV. US Department of Commerce, Economic and Statistics Administration, Bureau of Census.

Breiman, L., J. Friedman, R. Olshen and C. Stone, "Classification and Regression Trees," Pacific Grove: Wadsworth, 1984

Breiman, L. "Some Properties of Splitting Criteria," Statistics Department, University of California, Berkeley. 1992.

Burmaster, DE; Crouch, EAC (1997) "Lognormal distributions for body weight as a function of age for males and females in the United States, 1976-1980" Risk Analysis, V17 N4:499-505.

Burmaster, DE (1998). Lognormal Distributions for Skin Area as a Function of Body Weight. Risk Analysis. 18(1):27-32

Burmaster and Murray (1998). A Trivariate Distribution for the Height, Weight, and Fat of Adult Men. Risk Analysis. 18(4).

Burmaster, D.E., (1998) "Lognormal distributions for total water intake and tap water intake by pregnant and lactating women in the United States" Risk Analysis, 18(2):215-219.

Burmaster, D.E. and A.M. Wilson (in press) "Fitting Second-Order Mixture Models to Data with Many NonDetects Using Maximum Likelihood Estimation" Submitted to Human and Ecological Risk Assessment.

Canadian Ministry of National Health and Welfare (1981). Tap Water Consumption in Canada. Document no. 82-EHD-80. Public Affairs Directorate, Dept. Of National Health and Welfare, Ottawa, Canada

Cantor, et al (1987). National Cancer Institute (NCI) Study. Bladder Cancer, drinking water source and tapwater consumption: A Case Control Study. J Natl Cancer Inst. 79(6): 1269-1279.

CARB (1993); Measurements of Breathing Rate and Volume in Routinely Performed Daily Activities. Contract no. A033-25. 185 pps

Carey, M (1990). Occupational Tenure, Employer Tenure and Occupational Mobility. Occupational Outlook Quarterly. Summer 1990: pp 55-60.

Carey, M. (1988) Occupational Tenure in 1987: Many Workers Have Remained in their Fields. Monthly Labor Review. October 1988, pps 3-12

CHAD v1.0 (1997). Consolidated Human Activity Database, NERL (National Exposure and Research Laboratory).

Crespo, CJ, et al (1996). Leisure Time Physical Activity Among US Adults: Results from the Third National Health and Nutrition Examination Survey. Archives of Internal Medicine 156, January 8, 1996: 93-98.

D'Agostino, R.B. and M.A. Stephens (editors), Goodness of Fit Techniques. Marcel Dekker, Inc. New York, 1986.

Eisenberg J, McKone T. 1998. Decision Tree Method for the Classification of Chemical Pollutants: Incorporation of Across-Chemical Variability and Within-Chemical Uncertainty. Environ Sci Technol 32:3396-3404.

Eisenberg JN, Seto EYW, Olivieri AW, Spear RC. 1996. Quantifying Water Pathogen Risk in an Epidemiological Framework. Risk Analysis 16:549-563.

Eisenberg JNS, Bennett DH, McKone TE. 1998. Chemical Dynamics of Persistent Organic Pollutants: A Sensitivity Analysis Relating Soil Concentration Levels to Atmospheric Emissions. Environ Sci Technol 32:115-123.

Ershow & Cantor (1989): Life Sciences Research Office, Federation of American Societies for Experimental Biology.

Ershow et al (1991). Intake of Tap Water and Total Water by Pregnant and Lactating Women. American Journal of Public Health. 81:328-334.

Field, RW, et al (1998). Retrospetive Temporal and Spatial Mobility of Adult Iowa Women. Risk Analysis 18(5): 575-584.

Funk, et al (1998). Quantifying the Distribution of Inhalation Exposure in Human Populations: 2. Distributions of Time Spent by Adults, Adolescents, and Children at Home, at Work, and at School. Risk Analysis 18(1); 47-56.

Greenleaf JE, et al (1966). Water Consumption by man in a warm environment: a statistical analysis. J. Appl. Physiol. 21(1): 93-98.

Hamill, P.V.V; T.A. Drizd; C.L. Johnson; R.B. Reed and A.F. Roche: NCHS Growth Curves for Children Birth-18 Years. Washington, DC, DHEW Pub. No (PHS) 78-1650, Series 11, No. 165, 1977.

Hamill, P.V.V; T.A. Drizd; C.L. Johnson; R.B. Reed ; A.F. Roche and W.M. Moore (1979) "Physical Growth: National Center for Health Statistics Percentiles" Am. J. Clin. Nutr. 32:607-629.

Hill, M.S. Patterns of time use. In: Juster, F.T. Et al, Time, goods, and well being. Ann Arbor, MI: Survey Research Center, Institute for Social Research, U of MI, 1985; 135-166

Klepeis, NE, AM Tsang, and JV Behar (1996). Analysis of the NHAPS Respondents from a Standpoint of Exposure Assessment (FINAL report). NERL, ORD, USEPA Contract no. 68-01-7325. EPA/600/R-96/074. July 1996

Layton, DW (1993). Metabolically consistent breathing rates for use in dose assessments. Health Physics 64(1): 23-36.

Levallois, P, N.Guevin, S. Gingras, B. Levesque, J.-P. Webber and R. Letarte (1998). New pattens of drinking water consumption: results of a pilot study. Sci of Tot. Envir. 209: 233-241.

Linn, W.S et al (1992) Documentation of activity patterns in "high risk" groups exposed to ozone in the LA area. In: Proceedings of the Second EPA/AWMA Conference on Tropospheric Ozone, Atlanta Nov 1991, pp 701-712. AWMA, Pittsburgh, PA.

Linn, W.S. Et al (1993) Activity Patterns in Ozone exposured construction workers, J. Occ Med. Tox. 2(1) 1-14.

Najjar, M.F. And Rowland, M (1987). Anthropometric reference data and prevalence of overweight: US 1976-1980. NCHS, US DHHS, pub no. (PHS)87-1688; 1987. Data from NHANES Series 11, No. 238; Hyattsville, MD.

NAS (1977). Drinking Water and Health. Vol I. Washington, DC; National Academy of Sciences-National Research Council.

National Association of Realtors (1993), The Homebuying and Selling Process: 1993. The Real Estate Business Series. Washington, DC: NAR.

NHAPS CD-ROM

ODEQ Oregon Department of Environmental Quality (1998) "Guidance for use of probabilistic analysis in human health risk assessments – Interim Final"

Pennington (1983) Revision of theTotal Diet Study food list and diets. J Am. Diet. Assoc. 82: 166-173. cited in USEPA EFH (1996)

Pilote L, et al. 1996. Determinants of the Use of Coronary Angiography and Revascularization After Thrombolysis For Acute Myocardial Infarction. N Eng J Med 335:1198-1205.

Price, PS et al (1998). An Empirical Approach for Deriving Information on Total Duration of Exposure from Information on Historical Exposure. Risk Analysis 18(5):611-619.

Myers, L.; J. Lashley and R. Whitmore (March, 1998) "Development of Statistical Distributions for Exposure Factors –Final Report" RTI Research Triangle Institute. U.S.EPA Contract 68D40091.

Myers, L.; J. Lashley and R. Whitmore (April 1999) "Options for Development of Pametric Probability Distributions for Exposure Factors – Final Report" RTI Research Triangle Institute. U.S.EPA Contract 68D40091.

Robinson, J.P, and J. Thomas (1991). Time Spent in Activities, Locations, and Microenvironments: A CA-National Comparison Project Report. Las Vegas, NV: USEPA, Environmental Monitoring Systems Laboratory.

Roy, M and Courtay, C. Daily activitities and breathing parameters for use in respiratory tract dosimetry. Radiation Protection Dosimetry, 35: 179-186; 1991.

Sallis, J. et al (1985). Physical activity Assessment Methodology in the Five-City project. Am J. Epidemiology 121:91-106; 1985.

Schwab, M et al (1992). Using Longitudinal Data to Understand Children's Activity Patterns in an Exposure Context: Data from the Kanawha County Health Study. Env. International 18:173-189

Sell (1989). The Use of Children's Activity Patterns in Development of a Strategy for Soil Sampling in West Central Phoenix. Report for the Arizona DEQ.

Shamoo, D.A et al (1991) Activity Patterns in a panel of outdoor workers exposed to oxidant pollution. J. Expos. Anal. Environ. Epidem. 1(4): 423-438.

Silvers, et al (1994). How Children Spend their time: A sample survey for Use in Exposure and Risk Assessments. Risk Analysis, 14(6); 931-944

Snyder, W.S. et al (1975) and ICRP (1981). Report of the Task Group on Reference Man. Pub no. 23. Oxford Pergammon Press.

Spear RC, et al. 1991. Modeling Benzene Pharmacokinetics Across 3 Sets of Animal Data - Parametric Sensitivity and Risk Implications. Risk Analysis 11:641-654.

Spear RC, Grieb TM, Shang N. 1994. Parameter Uncertainty and Interaction in Complex Environmental Models. Water Resour Res 30:3159-3169.

Spier, C.E et al. (1992) Activity Patterns in elementary and high school students exposed to oxidant pollution. J Exp Anal Environ. Epid. 2(3): 277-293

Tarter, E.T. and M.D. Lock, (1991)"Model-Free Curve Estimation: Mutuality and Disparity of Approaches" Journal of Official Statistics, 7:219-23

Timmer, S.G. Et al. How children use time. In Juster, F.T., et al . Time, goods and well-being. U of MI, 1985: 353-382.

Tronstad R. 1995. Importance of Melon Type, Size, Grade, Container, and Season in Determining Melon Prices. Journal of Agricultural and Resource Economics 20:32-48.

Tsang and Klepeis (1996). Descriptive Tables from a detailed analysis of the NHAPS data. Report no. EPA/600/R-96/148. July 1996. Prepared for the US EPA by Lockheed Martin, Contract No. 68-W6-001, Delivery Order No. 13

US Army (1983). Water Consumption Planning Factors Study. Directorate of Combat Developments, United States Army Quartermaster School, Fort Lee, Virginia.

US EPA (1990). Exposure Factors Handbook. Office of Health and Environmental Assessment. EPA/600/8-89/043, March 1990.

US EPA (1997). Exposure Factors Handbook, Office of Health and Environmental Assessment. EPA/600/P-95/002A, Washington, DC, August 1997.

US EPA (1996). Exposure Factors Handbook. Science Advisory Board Review Draft. Exposure Assessment Group. EPA/600/P-96/002B a, b and c. Washington, DC, August 1996.

USEPA (1984) Office of Radiation Programs. An Estimation of the daily average food intake by age and sex for use in assessing the radionuclide intake of individuals in the general population. EPA-52/1-84-021

USEPA (1992) Dermal Exposure Assessment: Principles and Applications. Washington DC, Office of Health and Environmental Assessment EPA No 600/8-91-011B.

USEPA (1996a). Analysis of the National Human Activity Pattern Survey (NHAPS) Respondents from a Standpoint of Exposure Assessment: Percentage of Time Spent, Duration, and Frequency of Occurrence for Selected Microenvironments by Gender, Age, Time-of-day, Day-of-week, Season, and US Census Region. Office of Research and Development. July 1996. Final Report. EPA 68-01-7325.

USEPA (1996b). Descriptive Statistics Tables from a Detailed Analysis of the National Human Activity Pattern Survey (NHAPS) Data. July 1996. EPA /600/R-96/148.

Wiley JA et al (1991b) Study of Childrens Activity Patterns, Final Report. Survey Research Center, UCB. Prepared for CARB, Contract No. A733-149, September 1991.

Wiley JA et al. (1991a) Activity Patterns of CA Residents. Final Report. Survey Research Center, UCB. Prepared for CARB, Contract No. A6-177-33, May 1991.

Appendix 1: Data sources for use in development of PDFs

An extensive literature review was performed to identify sources of raw data that would be appropriate for constructing distributions for each exposure factor. The attached tables include all sources located during this search along with a brief description of each data set and contact person when available. The data is also labeled as to importance or usefulness for this report. The attributes of the data are coded in the last column of each table and the attributes are provided below.

Table A: Description of data attributes used to code data in following tables table

Label	Data attribute
1	primary raw data sources used to generate distributions in this report.
2	secondary raw data sources, available but not used to generate distributions in this report.
3	auxiliary data sources and referenced scientific reports or papers used for supporting and cross validation purposes.
4	attempts made to locate and obtain, but not available.
5	not obtained sources, i.e., raw data not presently available to be analyzed.

Table A.1. Original Data Sources for Body Weight

Original Source	Data Description	Referenced in	Contact	Attribute (*)
NHANES III (US DHHS, 1996). CD-ROM	plain drinking water as well as total water in foods reported for ~33,994 persons (~31,310 used) total tap water not included		CD-ROM available from NTIS.	(1)
USDA's Continuing Survey of Food Intakes by Individuals and Diet and Health Knowledge Survey (CSFII/DHKS) 1994-96	10^{th} national food consumption survey. Representative samples from 48 contiguous US states. Same variable as NHANES III.		CD-ROM available from NTIS	(1)
Wiley JA et al (1991b) Study of Childrens Activity Patterns,.	body weight data given for 1200 children (0-11 yrs).		available from CARB	(2)
NIST (National Institute of Standards and Technology), Anthrometric Data of Children (1977).	n=3,899 subjects aged 2-20 yrs with body weight data.		data accessible via web at http://ovrt.nist.gov/projects/anthrokids/orig77/individuals.csv	(2)
US Veterans Administration's "Normative Aging Study" in Boston, MA	long term longitudinal cohort body weight data for men between ages 50-80 yrs as reported by medical staff of the US Veterans	Burmaster and Murray (1998).	Corresponding author, David. E Burmaster: deb@Alceon.com	(5)

(*)

Table A.1. Original Data Sources for Body Weight (continued)

Original Source	Data Description	Referenced in	Contact	Attribute (*)
Burmaster, DE (1998). Lognormal Distributions for Skin Area as a Function of Body Weight.	Recreated database relied upon by USEPA EFH (1990, 1995, 1996). Measurements for 401 individuals (161 males, 140 females, 100 not identified by gender) aged >1 month to 66 yrs and 2 months.		Corresponding author, David. E Burmaster: deb@Alceon.com	(5)
Najjar, M.F. And Rowland, M. Anthropometric reference data and prevalence of overweight: US 1976-1980 (1987).	body weight; 18000 subjects (6months-74 yrs)	1. Layton, 1993 2. Burmaster and Crouch (1994) 3. USEPA EF (1996) 4. Brainard and Burmaster (1992) [Note: AIHC (1994) used their distributions and, Finley et al (1994) summarized their distributions too] 5. CalEPA, ATHSP (1996) 6. Burmaster and Couch (1997) 8. ODEQ (1998)	DHHS/CDC/NCHS/OVH, MD Najjar Health Statistician (301)436-7072, FAX (301) 436-3431 mfn1@CDC.GOV Mike Rowland HRSA, DHHS/HRSA/BPHG, MD (301) 594-4243, Mrowland@HRSA.DHHS.GOV	(3)
Fels Research Institute, Yellow Springs, OH	children 0-36 mnths.	1. Hamill et al (1979) 2. USEPA EFH (1996)		(3)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.2. Original Data Sources for Exposure Duration

Original Source	Data Description	Referenced in	Contact	Attribute (*)
BoC (1995) American Housing Survey for 1995.	55,000 sample interviews. National in scope. Number of years rented or owned- current residence time	1. Previous years AHS used by Israeli and Nelson (1992) then Finley et al (1994) fit more percentiles 2. USEPA EFH (1996)		(1)
National Association of Realtors (1993), The Homebuying and Selling Process	questionnaires to 15,000 homes (12% response rate)	USEPA EFH (1996) relevant study		(4)
Transfer of property title- sample of tax records (e.g.: Multnomah County, Oregon)		Sedman et al (1998)	Richard Sedman, PhD, 3158 Fairmount, Portland, OR 97201. Tel: (503)-229-6773, Fax: (503)-229-6945	(3)
Price, PS et al (1998). An Empirical Approach for Deriving Information on Total Duration of Exposure from Information on Historical Exposure	empirical approach to determining distribution of total duration from past durations			(3)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.3. Original Data Sources for Exposure Frequency

Original Source	Data Description	Referenced in	Contact	Attribute (*)
National Human Activity Patterns Survey (NHAPS) CD-ROM.	"largest and most current human activity pattern study available"; 82 possible locations, 91 different activities.(EF). Self reported diary data	1. USEPA EFH (1996) 2. Tsang and Klepeis (1996). 3. Klepeis, NE, AM Tsang, and JV Behar (1996). 4. CHAD v1.0 (1997)	available on CD-ROM from NTIS. Data used for this report obtained from Neil Klepeis: klepeis@uclink4.berkeley.edu or by tel: (510) 848-5827)	(1)
Wiley JA et al. (1991a) Activity Patterns of CA Residents.	1762 CA residents, aged 12 and over. Telephone interviews based on previous 24 hour activities	1. CalEPA, ATHSP (PDR 1996) 2. Funk, et al (1998) 3. Silvers, et al (1994) 4. CHAD v1.0 (1997)	Dr. Wiley, Asst Dir SRS, (510)642-3086, email: jwiley@uclink3.berkeley.edu	(2)
Wiley JA et al (1991b) Study of Children's Activity Patterns	1200 CA children, 0-11 yrs. Previous 24 hours activities recorded by telephone interview	1. CalEPA, ATHSP (PDR 1996) 2. Funk, et al (1998) 3. CHAD v1.0 (1997)	see Wiley JA et al. (1991a)	(2)
Data from Denver, Washington, DC, Cincinnati, and Valdez Activity Patterns Studies	All studies from self-reported diary data.	Included in CHAD v1.0 (1997) (Denver) and/or slated to be included in next CHAD version (Washington, DC, Cincinnati, and Valdez)		(2)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.3. Original Data Sources for Exposure Frequency (continued)

Original Source	Data Description	Referenced in	Contact	Attribute (*)
NHANES III, Phase I (1988-1990)	Adult, aged 20 and over as part of the 6 year NHANES III study from 1988 to 1994	Crespo, CJ, et al (1996).		(2)
Sell (1989). The Use of Children's Activity Patterns in Development of a Strategy for Soil Sampling in West Central Phoenix	preschool and tot school pop./ outdoors	US EPA EFH (1996)	contact Arizona DEQ: (602)207-300; env library: (602)207-2217	(4)
Robinson, J.P, and J. Thomas (1991). Time Spent in Activities, Locations, and Microenvironments: A CA-National Comparison Project Report.	CARB Time Activity Study + 1985 National Study (Americans Use of Time)	USEPA EFH (1996)		(5)
Kanawha County Health Study	Daily diary data for 90 children as part of a respiratory health status and gender stratified sample taken during both fall and spring.	Schwab, M et al (1992).		(5)
Field, et al (1998). Retrospetive Temporal and Spatial Mobility of Adult Iowa Women	619 Iowa females spatial and temporal mobility within and outside the home and in other buildings		Field, RW (lead author). Department of Preventative Medicine and Environmental Health, College of Medicine, University of Iowa, N222 Oakdale Hall, Iowa City, Iowa 52242	(3)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.3. Original Data Sources for Exposure Frequency (continued)

Original Source	Data Description	Referenced in	Contact	Attribute (*)
USEPA (1992) Dermal Exposure Assessment: Principles and Applications.	Exposure frequency data			(5)
Hill, M.S. Patterns of time use. In: Juster, F.T. Et al, Time, goods, and well being.	activity patterns	1. Layton, 1993		(3)
Timmer, S.G. Et al. How children use time. (1985)	activity patterns	1. Layton, 1993 2. USEPA EFH (1996)		(3)
Carey, M (1990). Occupational Tenure, Employer Tenure and Occupational Mobility.	occupational activity pattterns	USEPA EFH (1996)		(3)
Carey, M. (1988) Occuaptional Tenure in 1987: Many Workers Have Remained in their Fields.	occupational activity pattterns	USEPA EFH (1996)		(3)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.4. Original Data Sources for Inhalation Rate

Original Source	Data Description	Referenced in:	Contact	Attribute (*)
Adams, WC. Prof. Emeriti of Exercise Science, Human Performance Laboratory, UCD.	160 subjects, adults and children. Excel spreadsheets containing raw data on active, resting and field measured inhalation rates for adults, children. Subjects body weights also given.	1. Beals, J.A. Et al (1996); 2. CARB (1993)	Dr. Adams (UCD Prof Exercise Science) wcadams@ucdavis.edu or tel: (916) 752-0645	(1)
Layton, D. (1993)	children, M's and F's daily inhalation rates using three methods (no raw data- used data from USDA (NFCS), USDHHS (dietary data), USDA(NFCS))	US EPA, 1996 (EFH)		(3)
Roy, M and Courtay, C. Daily activitities and breathing parameters for use in respiratory tract dosimetry.	no raw data		Roy, Monique (NH) DHHS/NIH/NEI tel: (301) 496-5846 email: mr53f@NIH.GOV	(5)
Linn, W.S. Et al (1993) Activity Patterns in Ozone exposured construction workers.	self estimated inhalation rates of 19 construction workers before and during typical work day	US EPA EFH (1996).		(5)
Linn, W.S et al (1992) Documentation of activity patterns in "high risk" groups exposed to ozone in the LA area.	different groups (7 panels) = outdoor workers, school faculty + students, asthmatic adults and children, and male construction workers. Total 151 participants.	USEPA, EFH (1996)		(5)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.4. Original Data Sources for Inhalation Rate (continued)

Original Source	Data Description	Referenced in:	Contact	Attribute (*)
Spier, C.E et al. (1992) Activity Patterns in elementary and high school students exposed to oxidant pollution..	17 elem students + 19 HS students from suburban LA	USEPA, EFH (1996)		(5)
Shamoo, D.A et al (1991) Activity Patterns in a panel of outdoor workers exposed to oxidant pollution.	20 adult volunteers--> summer activity pattern (15 M's and 5 F's)	US EPA, EFH (1996)		(5)
Snyder, W.S. et al (1975) and ICRP (1981).	adult M, F, children (10 yrs), infant (1 yr) and newborn estimates.	US EPA, EFH (1981)		(5)
Sallis, J. et al (1985). Physical activity Assessment Methodology in the Five-City project.	1120 F's and 1006 M's (aged 20-74 yrs) from four communities in CA (activity)	Layton, 1993		(5)

(*) see footnote under Table 5 for explanation of attributes of the data source and their relevance to this study.

Table A.5. Original Data Sources for Water Ingestion

Original Source	Data Description	Referenced in	Contact	Attribute (*)
NHANES III (US DHHS, 1996). CD-ROM	plain drinking water as well as total water in foods reported for ~33,994 persons (~31,310 used). Total tap water not reported.		CD-ROM available from NTIS.	(1)
USDA's Continuing Survey of Food Intakes by Individuals and Diet and Health Knowledge Survey (CSFII/DHKS) 1994-96	10th national food consumption survey. Representative samples from 48 contiguous US states. Same variables as NHANES III.	1977-78 USDA NFCS data used by Ershow, et al (1991).	CD-ROM available from NTIS	(1)
Total Water and Tap Water Intake in the US: population based estimates of quantities and sources, Ershow & Cantor (1989):	26,000 sample subjects. Data from the USDA (1977-78) NFCS (National Food Consumption Survey). Database created to estimate tap water intake.	1. ODEQ (1998) 2. USEPA EFH (1996) 3. data used in Roseberry and Burmaster (1992) 4. Ershow et al (1991). 5. USEPA (1984) 6. CalTEPA ATHSP (1996)	Abbey Ershow at NHLBI (Md): ershow@gwgate.nhlbi.nih.gov or tel: (301) 435-0540	(3)
Water Consumption by man in a warm environment: a statistical analysis. Greenleaf JE, et al (1966).	87 male subjects. Used regression analysis to determine six variables associated with water intake.			(3)
Canadian Ministry of National Health and Welfare (1981). Tap Water Consumption in Canada.	970 Canadian individuals	1. USEPA EFH (1996) 2. CalEPA ATHSP (1996)		(3)

(*) see footnote under Table 5 for explanation of attributes of the data source and the relevance to this study.

Table A.5. Original Data Sources for Water Ingestion (continued)

Original Source	Data Description	Referenced in	Contact	Attribute (*)
Levallois, P, et al (1998). New pattens of drinking water consumption: results of a pilot study	139 rural and urban subjects from Quebec city. 24 hr recall plus 2 day diary data		Patrick Levallois (corresponding author). Patrick.Levallois@msp.ulaval.ca	(3)/ raw data is (5)
National Cancer Institute (NCI) Study. Bladder Cancer, drinking water source and tapwater consumption: A Case Control Study Cantor, et al (1987).	8000 adults (all white), 100% > 21 yrs, 57% > 65 yrs	USEPA EFH (1996)	Kenneth Cantor (epidemiologist) at NHLBI: cantork@epndce.nci.nih.gov or by tel: (301) 435-4718	(5)
Data from the FDA's Total Diet Study (conducted annually). Includes recall data from NFCS 1977-78 by USDA and NHANES II	no distinguishment between "sources of water" (i.e., tap, bottled, etc.), ~50,000 participants total from all surveys.	Pennington (1983)	Jean Pennington (NIH) DHHS/NIH/niddk-research nutritionist tel: (301) 594-8822, FAX (301) 480-3768 email: jp157d@NIH.GOV	(5)
NAS (1977). Drinking Water and Health. Vol I.	based on 8 previous studies (pre 1975)	USEPA EFH (1996)		(5)
US Army (1983). Water Consumption Planning Factors Study	based on climate and activity levels.	USEPA EFH (1996)		(4)