EPA United States
Environmental Protection
Agency

Ozone Population Exposure Analysis for Selected Urban Areas

EPA 452/R-07-004
January 2007

Ozone Population Exposure Analysis for Selected Urban Areas

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Health and Environmental Impacts Division
Ambient Standards Group
Research Triangle Park, North Carolina

DISCLAIMER

This document has been prepared by staff from the Ambient Standards Group, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, in conjunction with ICF Consulting (through Contract No. 68-D-01-052) and Alion Science and Technology, Inc. (through Contract No. 68-D-00-206). Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the views of the EPA, ICF Consulting, or Alion Science and Technology, Inc. A previous draft of this document was circulated to obtain review and comment from the Clean Air Scientific Advisory Committee (CASAC) and the general public. Any questions concerning this document should be addressed to John E. Langstaff, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, C504-06, Research Triangle Park, North Carolina 27711 (email: langstaff.john@epa.gov).

Subsequent to completion of the analyses described in this document, EPA analysis of uncertainty of the exposure modeling results uncovered an error in how children are characterized as active. This error resulted in an overestimate of the number of active children included in the exposure estimates for active children. Thus, the reader should not rely on the exposure estimates for active children provided in this document.

Acknowledgements

In addition to EPA staff, the following people contributed to writing this document.

ICF Consulting, San Francisco, CA and RTP, NC
Arlene Rosenbaum
Jonathan Cohen
Dan Bowman

Alion Science and Technology Inc, RTP, NC
Kristin Isaacs
Graham Glen
Luther Smith

This page intentionally left blank.

Table of Contents

List of Tables

1. INTRODUCTION

The Clean Air Act, which was last amended in 1990, requires EPA to set National Ambient Air Quality Standards (NAAQS) for widespread pollutants from numerous and diverse sources considered harmful to public health and the environment. EPA has set NAAQS for the following pollutants, which are called "criteria" pollutants: ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen oxides, and lead. The Clean Air Act requires periodic review of the science upon which the standards are based and the standards themselves to (1) ensure that they provide adequate health and environmental protection and (2) update those standards as necessary.

Under the NAAQS review process, EPA's Office of Research and Development (ORD) develops an "air quality criteria document" – a compilation and evaluation by EPA scientific staff and other expert authors of the latest scientific knowledge useful in assessing the health and welfare effects of the air pollutant. In August 2005, the second external review draft of the Air Quality Criteria for Ozone and Related Photochemical Oxidants was released for public comment and review by EPA's Clean Air Scientific Advisory Committee (CASAC), and a final document was released in 2006 (Ozone Criteria Document, US EPA, 2006a). The Ozone Criteria Document presents the latest available pertinent information on atmospheric science, air quality, exposure, dosimetry, health effects, and environmental effects of ozone and other related photochemical oxidants.

This report documents the methodology and input data used in the inhalation exposure assessment for ozone conducted in support of the current review of the ozone NAAQS. Specifically, this report includes the following:

- Summary of the overall inhalation exposure assessment methodology;
- Description of the inhalation exposure model used in this assessment;
- Description of the input data used for the 12 selected urban areas;
- Assessment of the quality and limitations of the input data for supporting the goals of the ozone NAAQS exposure analysis; and
- Sensitivity analyses.

The results of the exposure modeling are presented and discussed in the Ozone Staff Paper (US EPA, 2007); only selected results are presented in this report.

1.1 Selection of Urban Areas

The selection of urban areas to include in the exposure analysis takes into consideration the location of ozone field and epidemiology studies, the availability of ambient monitoring data for ozone, and the desire to represent a range of geographic areas, population demographics, and ozone climatology. These selection criteria are discussed further in the Ozone Staff Paper.

Based on these criteria, EPA selected the 12 urban areas in Table 1 for inclusion in the exposure analysis:

1.2 Exposure Periods

The exposure periods modeled were the ozone monitoring seasons for three years, 2002, 2003, and 2004. The seasons modeled for each area are listed in Table 1-1.

Table 2-1. Urban areas and time periods modeled

Urban Area (CSA)	Period modeled
Atlanta-Sandy Springs-Gainesville, GA-AL	March 1 to Oct. 31
Boston-Worcester-Manchester, MA-NH	April 1 to Sept. 30
Chicago-Naperville-Michigan City, IL-IN-WI	April 1 to Sept. 30
Cleveland-Akron-Elyria, OH	April 1 to Oct. 31
Detroit-Warren-Flint, MI	April 1 to Sept. 30
Houston-Baytown-Huntsville, TX	Jan. 1 to Dec. 30
Los Angeles-Long Beach-Riverside, CA	Jan. 1 to Dec. 30
New York-Newark-Bridgeport, NY-NJ-CT-PA	April 1 to Sept. 30
Philadelphia-Camden-Vineland, PA-NJ-DE-MD	April 1 to Oct. 31
Sacramento--Arden-Arcade--Truckee, CA-NV	Jan. 1 to Dec. 30
St. Louis-St. Charles-Farmington, MO-IL	April 1 to Oct. 31
Washington-Baltimore-N. Virginia, DC-MD-VA-WV	April 1 to Oct. 31

1.3 Populations Analyzed

Exposure modeling was conducted for the general population residing in each area modeled, as well as for school-age children (ages 5 to 18) and asthmatic school-age children. Due to the increased amount of time spent outdoors engaged in relatively high levels of physical activity, school-age children as a group are particularly at risk for experiencing ozone-related health effects.

2. DESCRIPTION OF THE APEX MODEL

The Air Pollutants Exposure model (APEX) is a personal computer (PC)-based program designed to estimate human exposure to criteria and air toxic pollutants at the local, urban, and consolidated metropolitan levels. APEX, also known as TRIM.Expo, is the human inhalation exposure module of EPA's Total Risk Integrated Methodology (TRIM) model framework (EPA, 1999), a modeling system with multimedia capabilities for assessing human health and ecological risks from hazardous and criteria air pollutants. It is being developed to support evaluations with a scientifically sound, flexible, and user-friendly methodology. Additional information on the TRIM modeling system, as well as downloads of the APEX Model, user's guide, and other supporting documentation, can be found on EPA's Technology Transfer Network (TTN) at http://www.epa.gov/ttn/fera.

2.1 History of APEX

APEX was derived from the National Ambient Air Quality Standards (NAAQS) Exposure Model (NEM) series of models. The NEM series was developed to estimate exposure to the criteria pollutants (e.g., CO, ozone). In 1979, EPA began to develop NEM by assembling a database of human activity patterns that could be used to estimate exposures to indoor and outdoor pollutants (Roddin et al., 1979). The data were then combined with measured outdoor concentrations in NEM to estimate exposures to CO (Biller et al., 1981; Johnson and Paul, 1983). In 1988, OAQPS began to incorporate probabilistic elements into the NEM methodology and use activity pattern data based on various human activity diary studies to create an early version of probabilistic NEM for ozone (i.e., pNEM/O3). In 1991, a probabilistic version of NEM was developed for CO (pNEM/CO) that included a one-compartment mass-balance model to estimate CO concentrations in indoor microenvironments. The application of this model to Denver, Colorado has been documented in Johnson et al. (1992). Several newer versions of pNEM/O3 were developed in the early- to mid-1990's, including versions developed for applications to nine urban areas for the general population, outdoor children, and outdoor workers (Johnson et al., 1996a,b,c). Between 1999 and 2001, updated versions of pNEM/CO (versions 2.0 and 2.1) were developed that rely on activity diary data from EPA's Consolidated Human Activities Database (CHAD) and enhanced algorithms for simulating gas stove usage, estimating alveolar ventilation rate (a measure of human respiration), and modeling home-to-work commuting patterns.

The first version of APEX was essentially identical to pNEM/CO (version 2.0) except that it ran on a PC instead of a mainframe. The next version, APEX2, was substantially different, particularly in the use of a personal profile approach rather than a cohort simulation approach. APEX3 introduced a number of new features including automatic site selection from national databases, a series of new output tables providing summary exposure and dose statistics, and a thoroughly reorganized method of describing microenvironments and their parameters. Most of the spatial and temporal constraints of pNEM and APEX1 were removed or relaxed by version 3.

The version of APEX used in this modeling analysis is APEX4, described in the APEX User's Guide and the APEX Technical Support Document (EPA, 2006c,d), henceforth referred to as the APEX User's Guide and TSD.

2.2 Theoretical Basis and Limitations of APEX

APEX estimates human exposure to criteria and toxic air pollutants at the local, urban, or consolidated metropolitan area levels using a stochastic, "microenvironmental" approach. The model randomly selects data for a sample of hypothetical individuals from an actual population database and simulates each hypothetical individual's movements through time and space (e.g., at home, in vehicles) to estimate their exposure to the subject pollutant. APEX models commuting and thus exposures at both home and work locations for individuals who work in different areas than they live.

A **microenvironment** is a three-dimensional space in which human contact with an environmental pollutant takes place and which can be treated as a well-characterized, relatively homogeneous location with respect to pollutant concentrations for a specified time period.

APEX can be conceptualized as a simulated field study that would involve selecting an actual sample of specific individuals who live in (or work and live in) a geographic area and then continuously monitoring their activities and subsequent inhalation exposure to a specific air pollutant during a specific period of time.

The main differences between APEX and an actual field study are that in APEX:

- The sample of individuals is a "virtual" sample, created by the model according to various demographic variables and census data of relative frequencies, in order to obtain a representative sample (to the extent possible) of the actual people in the study area;

- The activity patterns of the sampled individuals (e.g., the specification of indoor and other microenvironments visited and the time spent in each) are assumed by the model to be comparable to individuals with similar demographic characteristics, according to activity data such as diaries compiled in EPA's CHAD (EPA, 2002; McCurdy et al., 2000);

- The pollutant exposure concentrations are estimated by the model using a set of user-input ambient outdoor concentrations and information on the behavior of the pollutant in various microenvironments;

- Various reductions in ambient air quality levels can be simulated by either adjusting air quality concentrations to attain alternative ambient standards under consideration or by reducing source emissions and obtaining resulting air quality modeling outputs that reflect these potential emission reductions, and

- The model attempts to account for the most significant factors contributing to inhalation exposure – the temporal and spatial distribution of people and pollutant concentrations throughout the study area and among the microenvironments – while also allowing the flexibility to adjust some of these factors for regulatory assessment and other reasons.

All models have limitations that require the use of assumptions. Limitations of APEX lie primarily in the uncertainties associated with data distributions input to the model (e.g., human activity patterns). Primary uncertainties and assumptions associated with these distributions include the following:

- The population activity pattern data supplied with APEX (i.e., CHAD activity data) are compiled from a number of studies in different areas, and for different seasons and years. Therefore, the combined data set may not constitute a representative sample for a particular study scenario. Nevertheless, a large portion of CHAD is from a study of national scope (which could be extracted by the user if desired to create a representative sample).

- Commuting pattern data were derived from the 2000 U.S. Census. The commuting data address only home-to-work travel. The population not employed outside the home is assumed to always remain in the residential census tract. Furthermore, although several of the APEX microenvironments account for time spent in travel, the travel is assumed to always occur in basically a composite of the home and work tract. No other provision is made for the possibility of passing through other tracts during travel.

- APEX creates seasonal or annual sequences of daily activities for a simulated individual by sampling human activity data from more than one subject. Each simulated person essentially becomes a composite of several actual people in the underlying activity data.

- The model currently does not capture certain correlations among human activities that can impact microenvironmental concentrations (for example, cigarette smoking leading to an individual opening a window, which in turn affects the amount of outdoor air penetrating the residence).

- Certain aspects of the personal profiles are held constant, though in reality they change as individuals age. This is generally only an issue for simulations with long timeframes.

These and other uncertainties are discussed in section 4.

2.3 Overview of Model

APEX is designed to simulate population exposure to criteria and air toxic pollutants at local, urban, and regional scales. The user specifies the geographic area to be modeled and the number of individuals to be simulated to represent this population. APEX then generates a personal profile for each simulated person that specifies various parameter values required by the model. The model next uses diary-derived time/activity data matched to each personal profile to generate an exposure event sequence (also referred to as “activity pattern” or “composite diary”) for the modeled individual that spans a specified time period, such as one year. Each event in the sequence specifies a start time, exposure duration, geographic location, microenvironment, and activity. Probabilistic algorithms are used to estimate the pollutant concentration and ventilation (respiration) rate associated with each exposure event. The estimated pollutant concentrations account for the effects of ambient (outdoor) pollutant concentration, penetration factors, air exchange rates, decay/deposition rates, and proximity to emission sources, depending on the microenvironment, available data, and estimation method selected by the user. The ventilation rate is derived from an energy expenditure rate estimated for the specified activity. Because the modeled individuals represent a random sample of the population of interest, the distribution of modeled individual exposures can be extrapolated to the larger population. The model simulation includes five steps, each of which is described in the sections indicated below:

1. **Characterize the study area**. APEX selects census tracts within a study area – and thus identifies the potentially exposed population – based on user-defined criteria and availability of air quality and meteorological data for the area. (Section 2.3.1)

2. **Generate simulated individuals**. APEX stochastically generates a sample of hypothetical individuals based on the census data for the study area and human profile distribution data (such as age-specific employment probabilities). The user must specify the size of the sample. The larger the sample, the more representative it is of the population in the study area and the more stable the model results are (but also the longer the computing time). (Section 2.3.2)

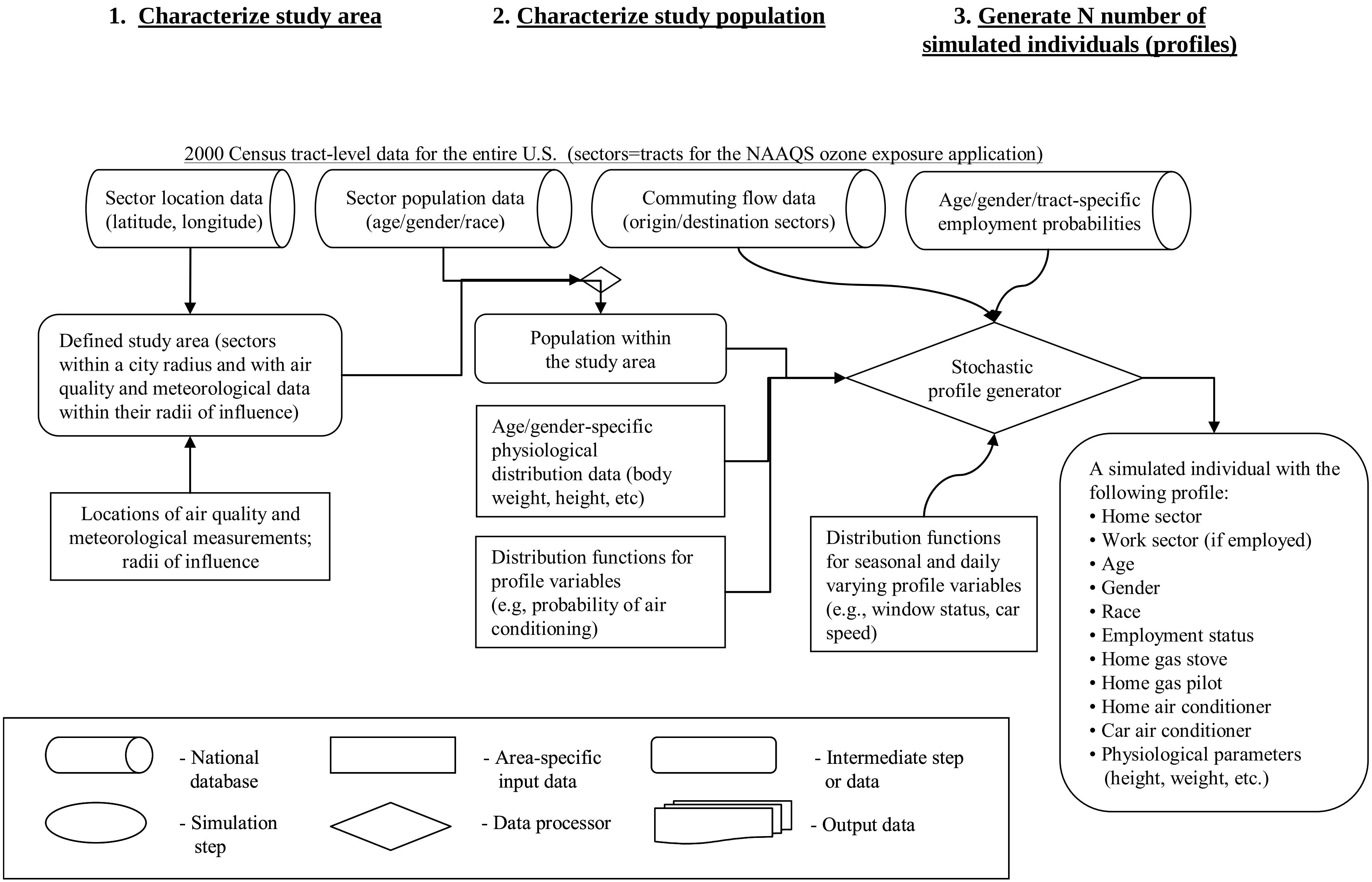
3. **Construct a sequence of activity events**. APEX constructs an exposure event sequence (activity pattern) spanning the period of the simulation for each of the hypothetical individuals (based on the supplied CHAD data, although other data could be used). (Section 2.3.3)

4. **Calculate hourly concentrations in microenvironments**. APEX users must define microenvironments that people in the study area would visit by assigning location codes in the supplied CHAD database to the user-specified microenvironments. The model then calculates hourly concentrations of a pollutant in each of these microenvironments for the period of simulation, based on the user-provided microenvironment descriptions and hourly ambient air quality data. All the hourly concentrations in the microenvironments are re-calculated for each of simulated individuals. (Section 2.4)

5. **Determine exposures**. APEX assigns a concentration to each exposure event based on the microenvironment occupied during the event and the person's activity. These values are averaged by clock hour to produce a sequence of hourly average exposures spanning the specified exposure period (typically one year). These hourly values may be further aggregated to produce daily, monthly, and annual average exposure values. (Section 2.5)

The model simulation continues until exposures are determined for entire modeling period for the user-specified number of simulated individuals. Figure 2-1 presents these steps within a schematic of the APEX model design. Subsections that follow provide addition detail on the key algorithms used in Steps 1 through 5.

Figure 2-1. Overview of the APEX Model



1. Characterize study area
2. Characterize study population
3. Generate N number of simulated individuals (profiles)
2000 Census tract-level data for the entire U.S. (sectors=tracts for the NAAQS ozone exposure application)
Sector location data (latitude, longitude)
Sector population data (age/gender/race)
Commuting flow data (origin/destination sectors)
Age/gender/tract-specific employment probabilities
Defined study area (sectors within a city radius and with air quality and meteorological data within their radii of influence)
Population within the study area
Stochastic profile generator
Locations of air quality and meteorological measurements; radii of influence
Age/gender-specific physiological distribution data (body weight, height, etc)
Distribution functions for profile variables (e.g, probability of air conditioning)
Distribution functions for seasonal and daily varying profile variables (e.g., window status, car speed)
A simulated individual with the following profile:
• Home sector
• Work sector (if employed)
• Age
• Gender
• Race
• Employment status
• Home gas stove
• Home gas pilot
• Home air conditioner
• Car air conditioner
• Physiological parameters (height, weight, etc.)
- National database
- Area-specific input data
- Intermediate step or data
- Simulation step
- Data processor
- Output data

Figure 2-1. Overview of the APEX Model, continued

4. Construct sequence of activity events for each simulated individual

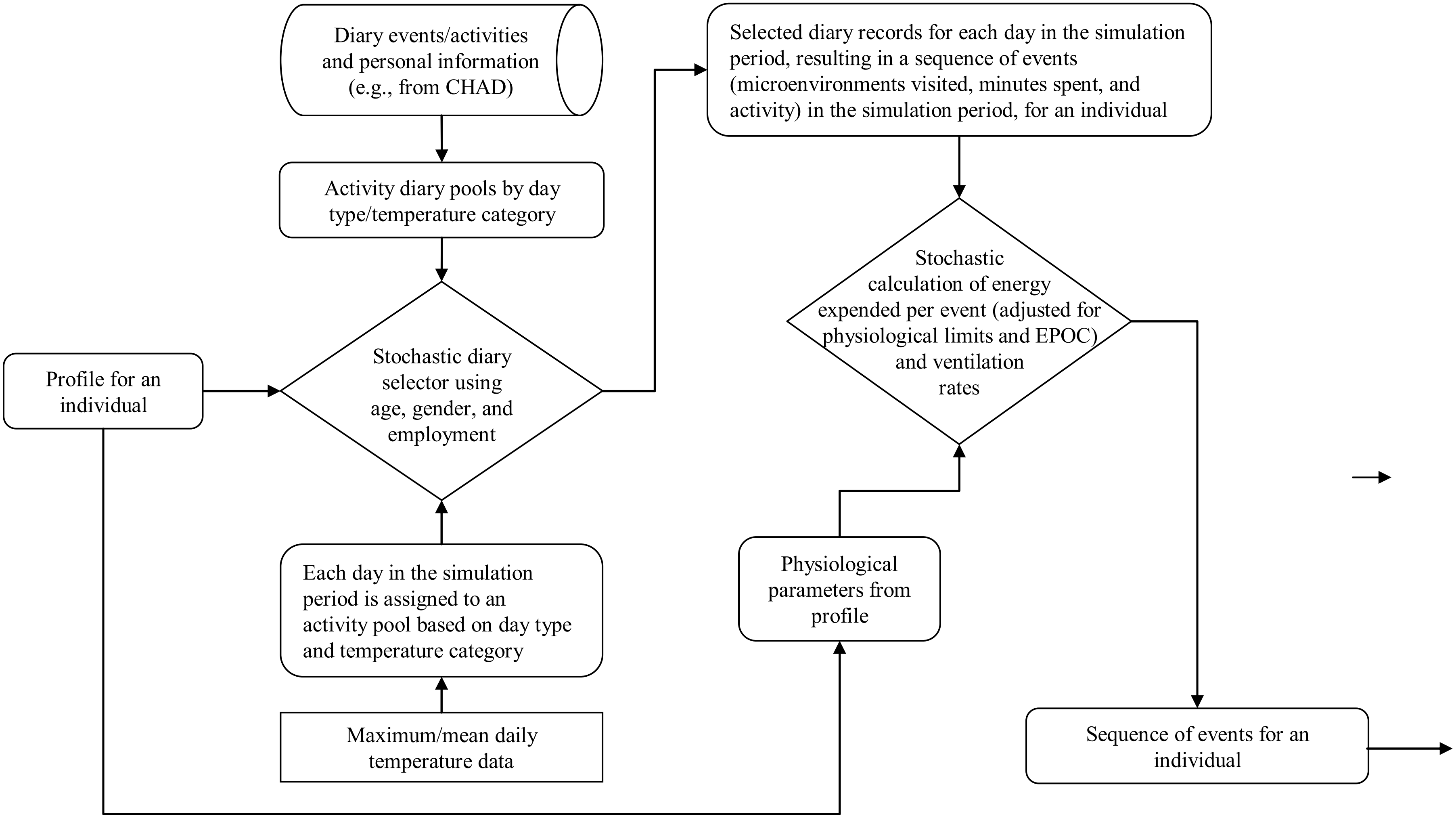
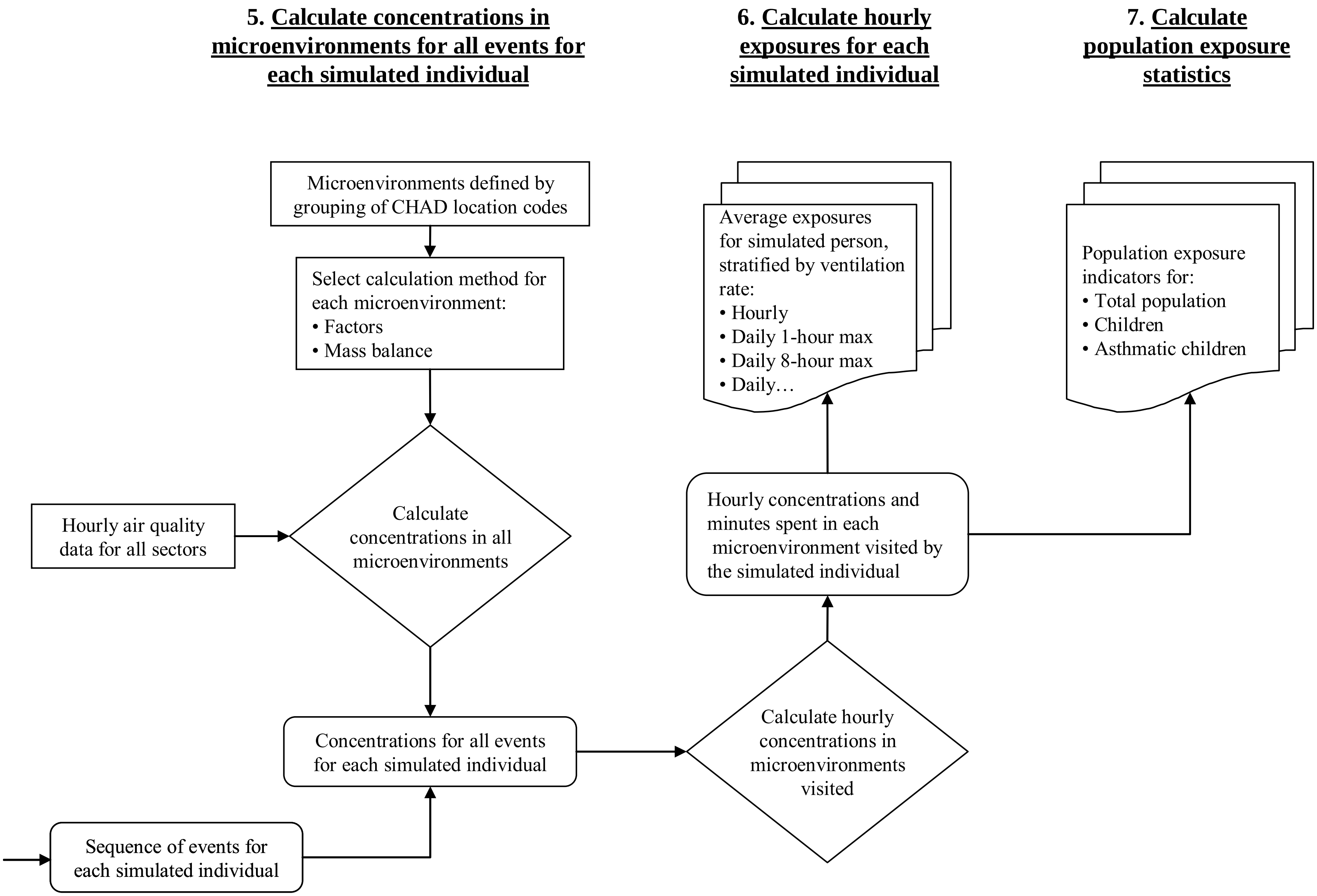

Diary events/activities and personal information (e.g., from CHAD)
Activity diary pools by day type/temperature category
Profile for an individual
Stochastic diary selector using age, gender, and employment
Each day in the simulation period is assigned to an activity pool based on day type and temperature category
Maximum/mean daily temperature data
Selected diary records for each day in the simulation period, resulting in a sequence of events (microenvironments visited, minutes spent, and activity) in the simulation period, for an individual
Stochastic calculation of energy expended per event (adjusted for physiological limits and EPOC) and ventilation rates
Physiological parameters from profile
Sequence of events for an individual

Figure 2-1. Overview of the APEX Model, concluded


5. Calculate concentrations in microenvironments for all events for each simulated individual
6. Calculate hourly exposures for each simulated individual
7. Calculate population exposure statistics
Microenvironments defined by grouping of CHAD location codes
Select calculation method for each microenvironment:
• Factors
• Mass balance
Hourly air quality data for all sectors
Calculate concentrations in all microenvironments
Concentrations for all events for each simulated individual
Sequence of events for each simulated individual
Calculate hourly concentrations in microenvironments visited
Hourly concentrations and minutes spent in each microenvironment visited by the simulated individual
Average exposures for simulated person, stratified by ventilation rate:
• Hourly
• Daily 1-hour max
• Daily 8-hour max
• Daily…
Population exposure indicators for:
• Total population
• Children
• Asthmatic children

2.3.1 Characterize the Study Area

The APEX study area has traditionally been on the scale of a city or slightly larger metropolitan area, although it is now possible to model larger areas, depending primarily on computing capabilities, available data, and the desired precision of the run.

In this analysis the study area is defined by a list of counties. The demographic data used by the model to create personal profiles is provided at the tract level. For each tract the model requires demographic information representing the distribution of age, gender, race, and work status within the study population. Each tract has a location specified by latitude and longitude for some representative point (e.g., geographic center). The current release of APEX includes input files that already contain this demographic and location data for all census tracts in the 50 United States, based on the 2000 Census.

The ambient air quality data are assigned to geographic areas called districts. The districts are used to assign pollutant concentrations to the tracts and microenvironments being modeled. The ambient air quality data are provided by the user as hourly time series for each district. As with tracts, each district has a representative location (latitude and longitude). Districts can extend outside of the study area.

APEX calculates the distance from each tract to each district center, and assigns the tract to the nearest district, provided the tract's representative location point (e.g., geographic center) is in the district. Each tract is assigned to only one district.

Ambient temperatures are input to APEX for different sites (locations). As with districts, APEX calculates the distance from each tract to each temperature site and assigns each tract to the nearest site.

2.3.2 Generate Simulated Individuals

APEX stochastically generates a user-specified number of simulated (hypothetical) persons to represent the population in the study area. Each simulated person is represented by a "personal profile." APEX generates the simulated person or profile by probabilistically selecting values for a set of profile variables (Table 2-1). The profile variables include:

- Demographic variables, which are generated based on the census data;
- Residential variables, which are generated based on sets of distribution data;
- Physiological variables, which are generated based on age- and gender-specific distribution data; and
- Daily varying variables, which are generated based on distribution data that change daily during the simulation period.

APEX first selects and calculates demographic, residential, and physiological variables (except for daily values) for all the specified number of simulated individuals, and then follows each individual over time and calculates exposures (and optionally doses) for each simulated person. The following subsections describe these variables in more detail.

Table 2-1. Profile Variables in APEX

Variable Type	Profile Variables	Description
Demographic variables	Age	Age (years)
	Gender	Male or Female
	Race	White, Black, Native American, Asian, and Other
	Home tract	Tract in which a simulated person lives
	Work tract	Tract in which a simulated person works
	Employment status	Indicates employment outside home
Residential variables	Air conditioner	Indicates presence of air conditioning at home
In-vehicle variables	Daily average car speed	Daily average car speed
	Car air conditioner	Indicates presence of air conditioning in the vehicle
Physiological variables	Height	Height of a simulated person (in)
	Weight	Body weight of a simulated person (lbs)
	Resting metabolic rate	Resting metabolic activity rate (kcal/min)
	Energy conversion factor	Oxygen uptake per unit of energy expanded (liters/kcal)
	Maximum permitted metabolic value	Maximum metabolic activity level that can be sustained for about five minutes (dimensionless)

Demographic Variables

The values of the demographic variables for a simulated profile are selected probabilistically according to their joint distribution in the input population files, which are derived from the 2000 U.S. Census.

Residential Variables

The residential variables are categorical variables that are used to indicate whether a residence or a car associated with a simulated person has the specified characteristic. These are randomly selected based on user-specified probabilities. For example, a user could specify probabilities of 0.3 for not having an air conditioner and 0.7 for having an air conditioner in their home.

Physiological Profile Variables

The physiological variables are used for calculating ventilation rates. Input data to APEX provide gender- and age-specific distributions for these variables.

2.3.3 Construction of Activity Sequences

APEX probabilistically creates a composite diary for each of the simulated persons by selecting a 24-hour diary record – or diary day – from an activity database for each day of the simulation period. CHAD data have been supplied with APEX for this purpose. A composite diary is a sequence of events that simulates the movement of a modeled person through geographical locations and microenvironments during the simulation period. Each event is defined by geographic location, start time, duration, microenvironment visited, and an activity performed. The activity database input to APEX contains the following information for each person for each day in each person's diary: age, gender, race, employment status, occupation, day of week, daily maximum hourly average temperature, the location, start time, duration, and type of each activity during the day.

APEX develops a composite diary for each of the simulated individuals according to the following steps:

1. Divide diary days in the CHAD database into user-defined activity pools, based on day type and temperature.
2. Assign an activity pool number to each day of the simulation period, based on the user-provided daily maximum/average temperature data.
3. Calculate a selection probability for each of the diary days in each of the activity pools, based on age/gender/employment similarity of a simulated person to a diary day.
4. Probabilistically select a diary day from available diary days in the activity pool assigned to each day of the simulation period.
5. Evaluate a metabolic value for each activity performed while in a CHAD location, based on the activity-specific metabolic distribution data. This is used to calculate a ventilation rate for the simulated person performing the activity.
6. Map the CHAD locations in the selected diary to the user-defined modeled microenvironments.
7. Concatenate the selected diary days into a sequential longitudinal diary for a simulated individual covering all days in the simulated period.

The method in APEX for creating longitudinal diaries that reflect the tendency of individuals to repeat activities is based on reproducing realistic variation in a user-selected key diary variable. APEX reads the values of the key variable from an external file. Currently, files have been constructed for both outdoor time and vehicle time for all CHAD diaries by summing the total time associated with "outdoor" and "vehicle" CHAD location codes for each diary. The actual diary construction method targets two statistics, D and A. The D statistic reflects the relative importance of within-person variance and between-person variance in the key variable. The A statistic quantifies the lag-one (day-to-day) variable autocorrelation. Desired D and A values for the key variable are selected by the user and set in the APEX parameters file, and the method algorithm constructs a longitudinal diary that preserves these parameters. Longitudinal diary data from a field study in children (Geyh et al., 2000), and subsequent analyses (Xue et al., 2004) suggest that D and A are stable over time (and perhaps over cohorts as well). Based on these studies, appropriate target values for the two statistics for outdoor time are estimated to be D=0.22 and A=0.19. A value of 0.2 is used for both of these parameters in the ozone exposure modeling, since precision beyond 0.1 is not warranted for these statistics. It turns out that the model results are insensitive to small changes in these values. The longitudinal diary methodology is described further in Appendix C.

2.4 Algorithms for Calculating Microenvironmental Concentrations

Probabilistic algorithms are used to estimate the pollutant concentration associated with each exposure event. The estimated pollutant concentrations account for the effects of ambient (outdoor) pollutant concentration, penetration factor, air exchange rate, decay/deposition rate, and proximity to emission sources, depending on the microenvironment, available data, and the estimation method selected by the user.

APEX calculates air concentrations in the various microenvironments visited by the simulated person by using the ambient air data for the relevant tracts and the user-specified method and parameters that are specific to each microenvironment. APEX calculates hourly concentrations in all the microenvironments at each hour of the simulation for each of the simulated individuals, based on the hourly ambient air quality data specific to the geographic locations visited by the individual. APEX provides two methods for calculating microenvironmental concentrations: the mass balance method and the transfer factors method (described in Sections 2.4.1 and 2.4.2, respectively). The user is required to specify a calculation method for each of the microenvironments; there are no restrictions on the method specified for each microenvironment (e.g., some microenvironments can use the transfer factors method while the others use the mass balance method).

2.4.1 Mass Balance Model

The mass balance method models an enclosed microenvironment as a well-mixed volume in which the air concentration is spatially uniform at any specific time. The concentration of an air pollutant in such a microenvironment is estimated using the following four processes:

- Inflow of air into the microenvironment;
- Outflow of air from the microenvironment;
- Removal of a pollutant from the microenvironment due to deposition, filtration, and chemical degradation; and
- Emissions from sources of a pollutant inside the microenvironment.

Table 2-2 lists the parameters required by the mass balance method to calculate concentrations in a microenvironment. The **proximity factor ($f_{proximity}$)** is used to account for differences in ambient concentrations between the geographic location represented by the ambient air quality data (e.g., a regional fixed-site monitor) and the geographic location of the microenvironment (e.g., near a roadway). This factor could take a value either greater than or less than 1. **Emission source (ES)** represents the emission rate for the emission source and **concentration source (CS)** is the mean air concentration resulting from the source. $\mathbf{R}_{removal}$ is defined as the removal rate of a pollutant from a microenvironment due to deposition, filtration, and chemical reaction. The **air exchange rate** ($\mathbf{R}_{air\ exchange}$) is expressed in air changes per hour. This analysis of ozone exposures does not consider sources of ozone, and these terms are set to zero.

Table 2-2. Mass Balance Model Parameters

Variable	Definition	Units	Value Range
$f_{proximity}$	Proximity factor	unitless	$f_{proximity} > 0$
CS	Concentration source	ppm	$CS \geq 0$
ES	Emission source	µg/hr	$ES \geq 0$
$R_{removal}$	Removal rate due to deposition, filtration, and chemical reaction	1/hr	$R_{removal} \geq 0$
$R_{air\ exchange}$	Air exchange rate	1/hr	$R_{air\ exchange} \geq 0$
V	Volume of microenvironment	m^3	$V > 0$

The mass balance equation for a pollutant in a microenvironment is described by:

$$\frac{\mathrm{dC}_{\mathrm{ME}}(\mathrm{t})}{\mathrm{dt}} = \Delta C_{in} - \Delta C_{out} - \Delta C_{removal} + \Delta C_{source} \qquad (2\text{-}1)$$

where:

$\mathrm{d}C_{\mathrm{ME}}(\mathrm{t})$ = Change in concentration in a microenvironment at time *t* (ppm),

ΔC_{in} = Rate of change in microenvironmental concentration due to influx of air (ppm/hour),

ΔC_{out} = Rate of change in microenvironmental concentration due to outflux of air (ppm/hour),

$\Delta C_{removal}$ = Rate of change in microenvironmental concentration due to removal processes (ppm/hour), and

ΔC_{source} = Rate of change in microenvironmental concentration due to an emission source inside the microenvironment (ppm/hour).

Within the time period of an hour each of the rates of change, ΔC_{in}, ΔC_{out}, $\Delta C_{removal}$, and $\Delta C_{sourcel}$, is assumed to be constant.

The change in microenvironmental concentration due to influx of air is represented by the following equation:

$$\Delta C_{in} = \frac{dC_{in}(t)}{dt} = C_{ambient}\ x\ f_{proximity}\ x\ f_{penetration}\ x\ R_{air\ exchange} \tag{2-2}$$

where:

$C_{ambient}$ = Ambient hourly outdoor concentration (ppm)
$f_{proximity}$ = Proximity factor (unitless)
$f_{penetration}$ = Penetration factor (unitless)
$R_{air\ exchange}$ = Air exchange rate (1/hour)

The change in microenvironmental concentration due to outflux of air is described by:

$$\Delta C_{out} = \frac{dC_{out}(t)}{dt} = R_{air\ exchange} \times C_{ME}(t) \tag{2-3}$$

The change in concentration due to deposition, filtration, and chemical degradation in a microenvironment is simulated based on the first-order equation:

$$\Delta C_{removal} = \frac{dC_{removal}(t)}{dt} = (R_{deposition} + R_{filtration} + R_{chemical})C_{ME}(t) = R_{removal}\ x\ C_{ME}(t) \tag{2-4}$$

where:

$R_{deposition}$ = Removal rate of a pollutant from a microenvironment due to deposition (1/hour)
$R_{filtration}$ = Removal rate of a pollutant from a microenvironment due to filtration (1/hour)
$R_{chemical}$ = Removal rate of a pollutant from a microenvironment due to chemical degradation (1/hour)
$R_{removal}$ = Removal rate of a pollutant from a microenvironment due to overall removal (1/hour)

We are not modeling indoor emissions of ozone, so the optional term ΔC_{source} will be uniformly equal to 0.0 for this study.

Equation 2-1 combined with Equations 2-2, 2-3, and 2-4 leads to:

$$\frac{\mathrm{dC}_{\mathrm{ME}}(\mathrm{t})}{\mathrm{dt}} = \Delta C_{in} - R_{air\ exchange}\ C_{ME}(t) - R_{removal}\ C_{ME}(t) \tag{2-5}$$

Solving the differential equation in Equation 2-5 leads to:

$$C_{ME}(t) = \frac{\Delta C_{in}}{R_{combined}} + (C_{ME}(0) - \frac{\Delta C_{in}}{R_{combined}}) \exp(-R_{combined} t) \qquad (2\text{-}6)$$

where:

$C_{ME}(0)$ = Concentration of a pollutant in a microenvironment at the beginning of a hour (ppm)

$C_{ME}(t)$ = Concentration of a pollutant in a microenvironment at time *t* within the time period of a hour (ppm)

$R_{combined}$ = $R_{air\ exchange} + R_{removal}$ (1/hour)

Based on Equation 2-6, the following three hourly concentrations in a microenvironment are calculated:

$$C_{ME}^{equil} = C_{ME}(t \to \infty) = \frac{\Delta C_{in}}{R_{combined}} \qquad (2\text{-}7)$$

$$C_{ME}^{hourly\ end} = C_{ME}^{equil} + (C_{ME}(0) - C_{ME}^{equil}) \exp(-R_{combined}) \qquad (2\text{-}8)$$

$$C_{ME}^{hourlymean} = \frac{\int_0^1 C(t)dt}{\int_0^1 dt} = C_{ME}^{equil} + (C_{ME}(0) - C_{ME}^{equil}) \frac{1 - \exp(-R_{combined})}{R_{combined}} \qquad (2\text{-}9)$$

where:

C_{ME}^{equil} = Equilibrium concentration in a microenvironment (ppm)

$C_{ME}(0)$ = Concentration in a microenvironment at the beginning of an hour (ppm)

$C_{ME}^{hourly\ end}$ = Concentration in a microenvironment at the end of an hour (ppm)

$C_{ME}^{hourlymean}$ = Hourly mean concentration in a microenvironment (ppm)

At each hour time step of the simulation period, APEX uses Equations 2-7, 2-8, and 2-9 to calculate the hourly equilibrium, hourly ending, and hourly mean concentrations. APEX reports hourly mean concentration as hourly concentration for a specific hour. The calculation continues to the next hour by using $C_{ME}^{hourly\ end}$ for the previous hour as $C_{ME}(0)$.

2.4.2 Factors Model

The factors method is simpler than the mass balance method. It does not calculate concentration in a microenvironment from the concentration in the previous hour and it has fewer parameters. Table 2-3 lists the parameters required by the factors method to calculate concentrations in a microenvironment without emissions sources.

Table 2-3. Factors Model Parameters

Variable	Definition	Units	Value Range
$f_{proximity}$	Proximity factor	unitless	$f_{proximity} > 0$
$f_{penetration}$	Penetration factor	unitless	$0 \leq f_{penetration} \leq 1$

The factors method uses the following equation to calculate hourly mean concentration in a microenvironment from the user-provided hourly air quality data:

$$C_{ME}^{hourlymean} = C_{ambient} \; x \; f_{proximity} \; x \; f_{penetration} \qquad (2\text{-}10)$$

where:

$C_{ME}^{hourlymean}$ = Hourly concentration in a microenvironment (ppm)
$C_{ambient}$ = Hourly concentration in ambient environment (ppm)
$f_{proximity}$ = Proximity factor (unitless)
$f_{penetration}$ = Penetration factor (unitless)

2.4.3 Commuting Outside of the Study Area

APEX allows for some flexibility in the treatment of persons in the modeled population who commute to destinations outside the study area. By specifying “KeepLeavers = No” in the simulation control parameters file (see Section 3.1), people who work inside the study area but live outside of it are not modeled, nor are people who live in the study area but work outside of it. By specifying “KeepLeavers = Yes,” these commuters are modeled. This triggers the use of two additional parameters, called LeaverMult and LeaverAdd. While a commuter is at work, if the workplace is outside the study area, then the ambient concentration is assumed to be related to the average concentration over all air districts at the same point in time, and is calculated as:

$$Ambient\ Concentrat\ ion = LeaverMult\ \times avg(t) + LeaverAdd \qquad (2\text{-}11)$$

where:

Ambient Concentration = Calculated ambient air concentrations for locations outside of the study area (ppm or ppm)
LeaverMult = Multiplicative factor for city-wide average concentration, applied when working outside study area
avg(t) = Average ambient air concentration over all air districts in study area, for time *t* (ppm or ppm)

LeaverAdd = Additive term applied when working outside study area

All microenvironmental concentrations for locations outside of the study area are determined from this ambient concentration by the same function as applies inside the study area.

2.5 Algorithms for Calculating Dose

Probabilistic algorithms are used to estimate the ventilation (respiration) rate associated with each exposure event. Ventilation (as discussed in Section 2.5.1 below) is a measure of human respiration, which is activity and physiology dependent. It is used in APEX to simulate human activities in order to estimate, more realistically, inhalation exposure and dose. The ventilation rate is derived from an energy expenditure rate estimated for the specified activity.

2.5.1 Ventilation

Ventilation is a general term for the movement of air into and out of the lungs. Minute or total ventilation is the amount of air moved in or out of the lungs per minute. Quantitatively, the amount of air breathed in per minute (V_I) is slightly greater than the amount expired per minute (V_E). Clinically, however, this difference is not important, and by convention minute ventilation is always measured on an expired sample, V_E.

The oxygen ventilation rate V_{O2} (l of O_2/min) is related to the energy expenditure rate for the given event activity and the given profile's physiology in terms of oxygen ventilation per unit energy expenditure, or:

$$V_{O2} = EE \, x \, ECF \tag{2-12}$$

where:

EE = Energy expenditure (kcal/min)
ECF = Energy conversion factor (l of O_2/kcal).

ECF is based on the physiology of the individual being modeled. EE is related to the activity-specific energy expenditure rate and the basal or resting energy expenditure (metabolic) rate of the given profile, or:

$$EE = MET \, x \, RMR \tag{2-13}$$

where:

MET = Metabolic equivalent of work (the ratio of the rate of energy consumption for non-rest activities to the resting rate of energy consumption) (dimensionless)
RMR = Resting metabolic rate (kcal/min).

RMR is based on the physiology of the individual being modeled. MET is the ratio of the activity-specific energy expenditure rate to the basal or resting energy expenditure rate. While different people have very different basal metabolic rates, it is generally found that the metabolic ratios do not exhibit as much variability. Thus, standing still might require two times the basal energy expenditure, or two MET, for most people, with relatively little variation. Since the basal rate is constant for each profile, it only has to be determined once and the activity-specific metabolic ratio can be used to determine the absolute energy expenditure rate, EE, for each activity.

Dividing equation 2-12 by body mass (BM) and using equation 2-13, one obtains:

$$V_{O2} / BM = RMR\ x\ ECF\ \ x\ MET / BM \qquad (2\text{-}14)$$

Graham and McCurdy (2005) describe an approach to estimate V_E directly from VO_2 using a series of regression-based equations. Using data compiled from 32 clinical exercise studies collected over a 25-year period by Dr. William C. Adams of the University of California at Davis, they developed an algorithm for four age groups and both genders. The algorithm accounts for differences in ventilation rate due to activity level, variability within age groups, and variation both between and within individuals. Their model is implemented in APEX as:

$$\ln(VE / BM)_i = b_0 + (b_1 * \ln(V_{O2} / BM_i)) + (b_2 * \ln(1 + age_i)) + (b_3 * gender_i) + eb_i + ew_i \qquad (2\text{-}15)$$

where:

the V_{O2}/BM term is given in terms of the APEX variables by equation 2-14,
age is the age of the individual in years, and
gender is a flag with value -1 for males and +1 for females.

Random error (*ε*) is allocated to two variance components used to estimate the between-person (inter-individual variability) residuals distribution (e_b) and within-person (intra-individual variability) residuals distribution (e_w). The regression parameters b_0, b_1, b_2, b_3, and e_b are assumed to be constant over time for a given simulated person, whereas e_w varies from event to event. These parameters are randomly drawn from normal distributions with means and standard deviations given in Table 2-4. e_b and e_w have mean zero.

Table 2-4. Ventilation Regression Parameters

Age range	mean b_0	stdev b_0	mean b_1	stdev b_1	mean b_2	stdev b_2	mean b_3	stdev b_3	stdev e_b	stdev e_w
0-19	4.4329	0.0579	1.0864	0.0097	-0.2829	0.0124	0.0513	0.0045	0.0955	0.1117
20-33	3.5718	0.0792	1.1702	0.0067	0.1138	0.0243	0.045	0.0031	0.1217	0.1296
34-60	3.1876	0.1271	1.1224	0.012	0.1762	0.0335	0.0415	0.0095	0.126	0.1152
>60	2.4487	0.3646	1.0437	0.0195	0.2681	0.0834	-0.0298	0.01	0.1064	0.0676

2.5.2 Excess Post-Exercise Oxygen Consumption

At the beginning of exercise, there is a lag between work expended and oxygen consumption. During this work/ventilation mismatch, an individual's energy needs are met by anaerobic processes. The magnitude of the mismatch between expenditure and consumption is termed the oxygen deficit. During heavy exercise, further oxygen deficit (in addition to that associated with the start of exercise) may be accumulated. At some point, oxygen deficit reaches a maximum value, and performance and energy expenditure deteriorate. After exercise ceases, ventilation and oxygen consumption will remain elevated above baseline levels. This increased oxygen consumption was historically labeled the "oxygen debt" or "recovery oxygen consumption." However, recently the term "excess post-exercise oxygen consumption" (EPOC) has been adopted for the phenomenon. APEX has an algorithm for adjusting the MET values to account for EPOC. This algorithm is described in Appendix B.

2.5.3 Body Surface Area

The algorithm for calculating body surface area (BSA) in APEX was developed by Burmaster (1998), and uses a univariate model for total skin area as a function of body weight. Through regression analysis, Burmaster determined that weight alone does as well as weight and height together in predicting total skin area, with the advantage of requiring only a single explanatory variable. Total skin area was found to follow a lognormal distribution that is a function of body weight according to:

$$BSA = e^{-2.2781} \, BM^{0.6821} \tag{2-16}$$

where:

BSA = body surface area (m^2)
BM = body mass (kg).

2.6 Exposure Calculations

APEX calculates exposure as a time series of exposure concentrations that a simulated individual experiences during the simulation period. APEX determines the exposure using hourly ambient air concentrations, calculated concentrations in each microenvironment based on these ambient air concentrations, and the minutes spent in a sequence of microenvironments visited according to the composite diary. The hourly exposure concentration at any clock hour during the simulation period is determined using the following equation:

$$C_i = \frac{\sum_{j=1}^{N} C_{ME(j)}^{hourlymean} \quad t_{(j)}}{T} \tag{2-17}$$

where:

C_i = Hourly exposure concentration at clock hour *I* of the simulation period (ppm)

N = Number of events (i.e., microenvironments visited) in clock hour *I* of the simulation period.

$C_{ME(j)}^{hourlymean}$ = Hourly mean concentration in microenvironment *j* (ppm)

$t_{(j)}$ = Time spent in microenvironment *j* (minutes)

T = 60 minutes

From the hourly exposures, APEX calculates time series of 8-hour and daily average exposure concentrations that a simulated individual would experience during the simulation period. APEX then statistically summarizes and tabulates the hourly, 8-hour, and daily exposures.

2.7 Model Output

All of the output files written by APEX are ASCII text files. Table 2-5 lists each of the output data files written for these simulations and provides descriptions of their content. Additional output files that can produced by APEX are given in Table 5-1 of the APEX User's Guide, and include hourly exposure, ventilation, and energy expenditures, and even detailed event-level information, if desired. The names and locations, as well as the output table levels (e.g., output percentiles, cut-points), for these output files are specified by the user in the simulation control parameters file. Specific output generated for the purposes of this document are discussed in Section 3.1.

Table 2-5. APEX Output Files

Output File Type	Description
Log	The *Log* file contains the record of the APEX model simulation as it progresses. If the simulation completes successfully, the log file indicates the input files and parameter settings used for the simulation and reports on a number of different factors. If the simulation ends prematurely, the log file contains error messages describing the critical errors that caused the simulation to end.
Profile Summary	The *Profile Summary* file provides a summary of each individual modeled in the simulation.
Microenvironment Summary	The *Microenvironment Summary* file provides a summary of the time and exposure by microenvironment for each individual modeled in the simulation.
Sites	The *Sites* file lists the tracts, districts, and zones in the study area, and identifies the mapping between them.
Output Tables	The *Output Tables* file contains a series of tables summarizing the results of the simulation. The percentiles and cut-off points used in these tables are defined in the simulation control parameters file.

3. PREPARATION OF MODEL INPUTS

The APEX model inputs require extensive analysis and preparation in order to ensure the model run gives valid and relevant results. This chapter begins with a description of the selected model options and discusses their significance. Following this introduction is a discussion of the model input files and other critical parameters. The chapter goes on to describe the sources of data for the APEX input files. File formats and physical file structures are not discussed in detail, as this information is presented in the APEX User's Guide (EPA, 2006c).

3.1 Model Options

Many of the important characteristics of a model run in APEX are set in the simulation control parameters file. In this file the user specifies the input and output files and their associated directories, as well as the basic parameters that characterize the run. The settings used for the model runs are described here.

The number of simulated persons in each model run was set to 60,000, an amount that tests indicated would be a large enough sample size to provide stable model results. The parameters controlling the location and size of the simulated area were set to include all counties in the study area CSA.

The settings that allow for replacement of CHAD data that are missing gender, employment or age values were all set to preclude replacing missing data. The width of the age window was set to 20 percent to increase the pool of diaries available for selection. The variable that controls the use of additional ages outside the target age window was set to 0.1 to further enhance variability in diary selection. See the APEX User's Guide for an explanation of these parameters.

The diary activity contributing the most to variability in exposure to ozone is the time spent outdoors, and we have selected that as the key predictor of exposure for the assembly of longitudinal diaries (see Appendix C). For school-age children, we take the diversity statistic D to be 0.2 and the autocorrelation to be 0.2. These values were derived from the Southern California Children's Study. We do not have data to base estimates of these parameters on for younger children and for adults, and we use the school-age children values for all ages.

Levels of physical activity were categorized by the Physical Activity Index (PAI), which is discussed in Appendix B. Children were characterized as active if their median daily PAI over the period modeled is 1.75 or higher, a level characterized by exercise physiologists as being "moderately active" or "active" (McCurdy, 2000).

3.2 Air Quality

APEX requires hourly ambient ozone concentrations at a set of sites in the study area. These data were obtained from the EPA AIRS Air Quality Subsystem for the years 2002, 2003, and 2004. These years were modeled since they are recent years that exhibit the year-to-year variability that is characteristic of ozone formation. All of the sites in AIRS within the boundaries of each CSA were used in this analysis. APEX uses the concentrations from the closest site to represent ambient concentrations at different locations in the study area. Table 3-1 lists the number of ozone monitors in each of the modeled CSAs.

Table 3-1. The Numbers of Ozone Monitors in the Study Areas

Urban area (CSA)	Number of monitors		
	2002	2003	2004
Atlanta-Sandy Springs-Gainesville, GA-AL	13	12	12
Boston-Worcester-Manchester, MA-NH	17	19	15
Chicago-Naperville-Michigan City, IL-IN-WI	32	30	27
Cleveland-Akron-Elyria, OH	11	11	11
Detroit-Warren-Flint, MI	10	10	10
Houston-Baytown-Huntsville, TX	21	23	21
Los Angeles-Long Beach-Riverside, CA	45	43	44
New York-Newark-Bridgeport, NY-NJ-CT-PA	30	30	29
Philadelphia-Camden-Vineland, PA-NJ-DE-MD	18	17	16
Sacramento--Arden-Arcade--Truckee, CA-NV	21	22	22
St. Louis-St. Charles-Farmington, MO-IL	18	18	17
Washington-Baltimore-N. Virginia, DC-MD-VA-WV	28	28	26

3.2.1 Missing Data Replacement

Missing air quality data were estimated by the following procedure. Where there were consecutive strings of missing values (data gaps of less than 6 hours, missing values were estimated by linear interpolation between the valid values at the ends of the gap. Remaining missing values at a monitor were estimated by fitting linear regression models for each hour of the day, with each of the other monitors, and choosing the model which maximizes R^2 for each hour of the day, subject to the constraints that R^2 be greater than 0.5 and the number of regression data values is at least 50. If there were any remaining missing values at this point, for gaps of less than 9 hours, missing values were estimated by linear interpolation between the valid values at the ends of the gap. Any remaining missing values were replaced with the regionwide mean for that hour.

3.3 Air Quality Projections for Alternative Standards Scenarios

In addition to modeling exposures based on historical air quality, an analysis was conducted using air quality representative of just meeting the current 8-hour O_3 NAAQS of 0.08 ppm, as well as seven alternative standards. This was done using a quadratic rollback approach to adjust the hourly O_3 concentrations observed in 2002-2004 to yield a design value corresponding to the standard being modeled. Design values for the current 8-hour O_3 NAAQS are calculated as the 3-year averages of the annual 4th daily maximum 8-hr average concentration based on the maximum monitor within an urban area. The quadratic rollback technique combines both linear and quadratic elements to reduce higher concentrations more than lower concentrations near ambient background levels. Table 3-2 shows the alternative standards, their corresponding attainment thresholds and the form of the standard used for each scenario. Additional details about rollback and the alternative standard scenarios can be found in the Ozone Staff Paper.

Table 3-2. List of the Current and Alternative 8-hour Ozone Standard Scenarios used in the Exposure Analysis

Attainment Threshold	Form of Standard	Labels for graphs
0.084 ppm	3rd-highest form 4th-highest form	85/3 85/4
0.080 ppm	4th-highest form	81/4
0.074 ppm	3rd-highest form 4th-highest form 5th-highest form	75/3 75/4 75/5
0.070 ppm	4th-highest form	71/4
0.064 ppm	4th-highest form	65/4

3.4 Meteorological Data

Hourly temperature data are from the National Climatic Data Center Surface Airways Hourly TD-3280 dataset (NCDC Surface Weather Observations). Daily average and 1-hour maxima are computed from these hourly data.

There are two files that are used to provide meteorological data to APEX. One file, the meteorological station location file, contains the locations of meteorological data recordings, expressed in latitude and longitude coordinates. This file also contains start and end dates for the data recording periods. The temperature data file contains the data from the locations in the temperature zone location file. This file contains daily maximum and daily average temperature readings for the period being modeled for the meteorological stations in and around the study area. Table 3-3 lists the meteorological stations used for each modeled area.

Table 3-3 The Meteorological Stations for Each Study Area

Urban area (CSA)	NWS station ID (WBAN)
Atlanta-Sandy Springs-Gainesville, GA-AL	03813, 13873, 13874, 13882, 93842
Boston-Worcester-Manchester, MA-NH	14739, 14745, 14765, 94746
Chicago-Naperville-Michigan City, IL-IN-WI	14839, 14848, 94822, 94846
Cleveland-Akron-Elyria, OH	14820, 14852, 14860, 14891, 14895
Detroit-Warren-Flint, MI	14822, 14826, 14836, 94830, 94847
Houston-Baytown-Huntsville, TX	12912, 12917, 12960, 93987
Los Angeles-Long Beach-Riverside, CA	23129, 23155, 23161, 23174, 23188, 23190
New York-Newark-Bridgeport, NY-NJ-CT-PA	04725, 04781, 13739, 14732, 14734, 14737, 14740, 14765, 14777, 93730, 94702, 94728, 94789
Philadelphia-Camden-Vineland, PA-NJ-DE-MD	13739, 13781, 14737, 93730
Sacramento--Arden-Arcade--Truckee, CA-NV	23185, 23232, 23237
St. Louis-St. Charles-Farmington, MO-IL	13994, 93822
Washington-Baltimore-N. Virginia, DC-MD-VA-WV	13740, 13743, 13781, 14711, 93721, 93738, 93739

3.5 Population Demographics

APEX takes population characteristics into account to develop accurate representations of study area demographics. Specifically, population counts by area and employment probability estimates are used to develop representative profiles of hypothetical individuals for the simulation.

APEX is very flexible in the resolution of population data provided. As long as the data are available, any resolution can be used (e.g., county, census tract, census block). For this application of the model, we used census tract level data.

Tract-level population counts come from the 2000 Census of Population and Housing Summary File 1. Summary File 1 (SF 1) contains the 100-percent data, which is the information compiled from the questions asked of all people and about every housing unit. The first level of official Census race categories and their abbreviations are:

- White (W)
- Black or African American (B)

- American Indian or Alaska native (N)
- Asian (A)
- Native Hawaiian or other Pacific Islander (OH)
- Other single race (OO)
- Two or more races combined (O2)

The categories OH, OO, and O2 were combined into a single “Other” class (“O”) for modeling purposes. Hispanics are not separated, as the Census Bureau did not consider Hispanic to be a race.

In the 2000 U.S. Census, estimates of employment were developed by census tract. Employment data from the 2000 census can be found on the U.S. census web site at the address http://www.census.gov/population/www/cen2000/phc-t28.html (Employment Status: 2000-Supplemental Tables). The file input to APEX is broken down by gender and age group, so that each gender/age group combination is given an employment probability fraction (ranging from 0 to 1) within each census tract. The age groupings in this file are: 16-19, 20-21, 22-24, 25-29, 30-34, 35-44, 45-54, 55-59, 60-61, 62-64, 65-69, 70-74, and >75. Children under 16 years of age are assumed to be not employed.

3.6 Asthma Prevalence Rates

One of the important population subgroups for the exposure assessment is asthmatic children. Evaluation of the exposure of this group with APEX requires the estimation of children’s asthma prevalence rates, detailed in Appendix G. The estimates are based on children’s asthma prevalence data from the National Health Interview Survey (NHIS) for 2003. Asthma prevalence rates for children aged 0 to 17 years were calculated for each age, gender, and region. The regions defined by NHIS are the Census Regions: “Midwest,” “Northeast,” “South,” and “West.” The reported survey responses were weighted to take into account the complex survey design of the NHIS survey. Standard errors and confidence intervals for the prevalence rates were calculated using a logistic model, taking into account the survey design. Logistic analysis of the prevalence relationships to age showed statistically significant differences between the prevalence functions for the two genders and for the four regions. Therefore the relationships of prevalence to age were estimated separately for each gender and region. A scatterplot smoothing technique using the LOESS smoother was applied to smooth the prevalence curves and compute the standard errors and confidence intervals for the smoothed prevalence estimates.

3.7 Commuting Database

As part of the population demographics inputs, it is important to integrate working patterns into the assessment. In addition to using estimates of employment by tract, APEX also incorporates home-to-work commuting data.

Commuting data were originally derived from the 2000 Census and were collected as part of the Census Transportation Planning Package (CTPP). These data are available from the U.S. DOT Bureau of Transportation Statistics (BTS) at the web site http://transtats.bts.gov/. The data used to generate APEX inputs were taken from the “Part 3-The Journey To Work” files. These files contain counts of individuals commuting from home to work locations at a number of geographic scales.

These data were processed to calculate fractions for each tract-to-tract flow to create the national commuting data distributed with APEX. This database contains commuting data for each of the 50 states and Washington, D.C.

Commuting within the Home Tract

The APEX data set does not differentiate people that work at home from those that commute within their home tract.

Commuting Distance Cutoff

A preliminary data analysis of the home-work counts showed that a graph of log(flows) versus log(distance) had a near-constant slope out to a distance of around 120 kilometers. Beyond that distance, the relationship also had a fairly constant slope but it was flatter, meaning that flows were not as sensitive to distance. A simple interpretation of this result is that up to 120 km, the majority of the flow was due to persons traveling back and forth daily, and the numbers of such persons decrease fairly rapidly with increasing distance. Beyond 120 km, the majority of the flow is made up of persons who stay at the workplace for extended times, in which case the separation distance is not as crucial in determining the flow.

To apply the home-work data to commuting patterns in APEX, a simple rule was chosen. It was assumed that all persons in home-work flows up to 120 km are daily commuters, and no persons in more widely separated flows commute daily. This meant that the list of destinations for each home tract was restricted to only those work tracts that are within 120 km of the home tract. When the same cutoff was performed on the 1990 census data, it resulted in 4.75% of the home-work pairs in the nationwide database being eliminated, representing 1.3% of the workers. The assumption is that this 1.3% of workers do not commute from home to work on a daily basis. It is expected that the cutoff reduced the 2000 data by similar amounts.

Eliminated Records

A number of tract-to-tract pairs were eliminated from the database for various reasons. A fair number of tract-to-tract pairs represented workers who either worked outside of the U.S. (9,631 tract pairs with 107,595 workers) or worked in an unknown location (120,830 tract pairs with 8,940,163 workers). An additional 515 workers in the commuting database whose data were missing from the original files, possibly due to privacy concerns or errors, were also deleted.

3.8 Activity Patterns – CHAD

Exposure models use human activity pattern data to predict and estimate exposure to pollutants. Different human activities, such as outdoor exercise, indoor reading, or driving, have different pollutant exposure characteristics. In addition, different human activities require different metabolic rates, and higher rates lead to higher doses. To accurately model individuals and their exposure to pollutants, it is critical to have a firm understanding of their daily activities.

The Consolidated Human Activity Database (CHAD) provides data on human activities through a database system of collected human diaries, or daily activity logs. The purpose of CHAD is to provide a basis for conducting multi-route, multi-media exposure assessments (McCurdy et al., 2000).

The data contained within CHAD come from multiple surveys with varied structures (Table 3-3). In general, the surveys have a data foundation based on daily diaries of human activity. This is the foundation from which CHAD was created. Individuals filled out diaries of their daily activities and this information was input and stored in CHAD. Relevant data for these individuals, such as age, are included as well. In addition, CHAD contains activity-specific metabolic distributions developed from literature-derived data, which are used to provide an estimate of metabolic rates of respondents through their various activities.

There are four CHAD-related input files used in the APEX system. Two of these files are downloaded directly from the "Query Questionnaire" link on the CHADNet (http://www.epa.gov/chadnet1) page, and then manipulated to fit into the APEX framework. These are the human activity diaries file and the personal data file.

The third input file contains metabolic information for different activities listed in the diary file. These metabolic activity levels are in the form of distributions. Some activities are specified as a single point value (for instance, sleep), while others, such as athletic endeavors or manual labor, are normally, lognormally, or otherwise statistically distributed. APEX samples from these distributions and calculates values to simulate the variable nature of activity levels among different people.

Table 3-3. Summary of Studies Used in CHAD

Study name	Geographic coverage	Study time period	Subject ages	Number of persons	Number of person-days	Diary type and study design (random or not)	Reference
Baltimore	A single building in Baltimore	01/1997-02/1997, 07/1998-08/1998	72-93	26	391	Diary	Williams et al, 2000
California Adolescents and Adults (CARB)	California	10/1987-09/1988	12-17 18-94	183 1,579	183 1,579	Recall; Random	Robinson et al. (1989), Wiley et al. (1991a)
California Children (CARB)	California	04/1989- 02/1990	0-11	1,200	1,200	Recall; Random	Wiley et al. (1991b)
Cincinnati (EPRI)	Cincinnati metropolitan area	03/1985-04/1985, 08/1985	0-86	888	2,614	Diary; Random	Johnson (1989)
Denver (EPA)	Denver metropolitan area	11/1982- 02/1983	18-70	432	805	Diary; Random	Johnson (1984), Akland et al. (1985)
Los Angeles: Elementary School Children	Los Angeles	10/1989	10-12	17	51	Diary	Spier et al. (1992)

Los Angeles: High School Adolescents	Los Angeles	09/1990-10/1990	13-17	19	43	Diary	Spier et al. (1992)
National: NHAPS-Air	National	09/1992-10/1994	0-93	4,723	4,723	Recall; Random	Klepeis et al. (1995), Tsang and Klepeis (1996)
National: NHAPS-Water	National	09/1992-10/1994	0-93	4,663	4,663	Recall; Random	Klepeis et al. (1995), Tsang and Klepeis (1996)
Washington, D.C. (EPA)	Wash., D.C. metropolitan area	11/1982-02/1983	18-98	699	699	Diary; Random	Hartwell et al. (1984), Akland et al. (1985)

The fourth input file maps five-digit location codes used in the diary file to APEX microenvironments. Because each simulation may contain different numbers and types of microenvironments, it is important to ensure that the codes map properly to the appropriate microenvironment. If this file does not represent a reasonable mapping, the model will not accurately simulate exposure related to daily activities.

Personal Information file. Personal data are contained in the CHAD questionnaire file that is distributed with APEX This file also has information for each day individuals have diaries. The different variables in this file are:

- The study, person, and diary day identifiers
- Day of week
- Gender
- Race
- Employment status
- Age in years
- Maximum temperature in degrees Celsius for this diary day
- Mean temperature in degrees Celsius for this diary day
- Occupation code
- Time, in minutes, during this diary day for which no data are included in the database

Diary Events file. The human activity diary data are contained in a file that is distributed with APEX. This is a large file because it contains diaries for about 23,000 people broken out at intervals ranging from one minute to one hour. These diaries vary in length from one to 15 days. This file contains the following variables:

- The study, person, and diary day identifiers
- Start time of this activity
- Number of minutes for this activity
- Activity code
- Location code

Activity Specific Metabolic file. The third CHAD file is also distributed with APEX and contains the metabolic parameters for each of the CHAD activities.

3.9 Physiological Distributions

APEX requires physiological parameters for subjects in order to accurately model their pollutant intake via metabolic processes. This is because physiological differences may cause people with the same exposure and activity scenarios to have different pollutant intake levels. The physiological parameters file distributed with APEX is described in the APEX User's Guide.

3.10 Microenvironment Specifications

The microenvironments in APEX provide the specific locations within an air quality district where modeled individuals are exposed to pollutants. Microenvironments are used to capture the differences between exposure concentrations in different types of environments (e.g., indoors, in cars, outdoors) within an area with the same estimated ambient air concentration. There are two basic methods for calculating concentrations in microenvironments: the transfer factors method and the mass balance method. The parameters for both factors and mass balance calculations used in this simulation are listed in Table 3-4.

Table 3-4. Microenvironment Parameter Information

Calculation Method	Parameter Type with Abbreviation	Units	Distribution
Transfer Factors	Proximity (PR)	unitless	Normal distribution
	Penetration (PE)	unitless	Normal distribution
Mass Balance	Proximity (PR)	unitless	Normal distribution
	Decay Rate (DE)	1/hr	Lognormal distribution
	Air exchange rate (AER)	Air changes/hr	Lognormal distribution

The factors method is used to model simple environments, like outdoor areas, that do not contain pollutant sources. The ambient ozone concentrations are from the air quality data input file. There are two parameters that affect the pollutant concentration calculation in the factors method, the proximity and penetration factors. The proximity factor is a unitless parameter that represents the proximity of the microenvironment to a monitoring station. The penetration factor is a unitless parameter that represents the fraction of pollutant entering a microenvironment from outside the microenvironment via air exchange. The development of the proximity factors and penetration factors used in this analysis is discussed in Appendix A.

The mass balance method is more appropriate for complex environments. In addition to proximity factors and penetration factors, this method supports parameters for emissions sources, decay rate, air exchange rate, volume, and the average removal rate. Each of these parameters can be modeled within the microenvironment or left out of the simulation. Both decay rate and emissions sources have a default value of zero, which gives them no effect on the simulation. The air exchange rate and volume have no default values. They only effect the microenvironment calculation if they are specifically included in the definition of the microenvironment. Several microenvironments using the mass balance method utilize one or more of these additional parameters. See Appendix A for a full description of the values used for the development of these parameters.

3.10.1 Microenvironments Modeled

In APEX, microenvironments provide the exposure locations for modeled individuals. For exposures to be estimated accurately, it is important to have realistic microenvironments that match closely to the locations where actual people spend time on a daily basis.

As discussed above, the two methods available in APEX for calculating pollutant levels within microenvironments are: 1) factors and 2) mass balance. A list of microenvironments used in this study, the calculation method used, and the parameters used to calculate the microenvironment concentrations can be found in Table 3-5.

Table 3-5. List of Microenvironments and Calculation Methods Used

Microenvironment	Calculation Method	Parameter Types used [1]
Indoors – Residence	Mass balance	AER and DE
Indoors – Bars and restaurants	Mass balance	AER and DE
Indoors – Schools	Mass balance	AER and DE
Indoors – Day-care centers	Mass balance	AER and DE
Indoors – Office	Mass balance	AER and DE
Indoors – Shopping	Mass balance	AER and DE
Indoors – Other	Mass balance	AER and DE
Outdoors – Near road	Factors	PR
Outdoors – Public garage - parking lot	Factors	PR
Outdoors – Other	Factors	None
In-vehicle – Cars and Trucks	Factors	PE and PR
In-vehicle - Mass Transit (bus, subway, train)	Factors	PE and PR

[1] AER=air exchange rate, DE=decay-deposition rate, PR=proximity factor, PE=penetration factor

Each of the microenvironments is designed to simulate an environment in which people spend time during the day. CHAD locations are linked to the different microenvironments in the *Microenvironment Mapping* File (see Section 3.8.4). There are many more CHAD locations than microenvironment locations (there are 113 CHAD codes versus 12 microenvironments in this assessment) and thus most of the microenvironments have multiple CHAD locations mapped to them.

The mass balance microenvironments have two parameters defined, the air exchange rate and the decay rate. The air exchange rate models the exchange of outside air with the microenvironment, while the decay rate models the rate of ozone breakdown or removal within the microenvironment. The development of air exchange rate values for this analysis is discussed in Section 3.9.2 and Appendix A. The development of the decay rate distribution is described in Section 3.9.3.

3.10.2 Microenvironment Descriptions

Microenvironment #1: Indoors-Residence. The Indoors-Residence Microenvironment accounts for three variables that affect ozone exposure: whether or not air conditioning is present, the average outdoor temperature, and the ozone decay rate. The first two of these variables affect the air exchange rate. An excerpt from the input file describing this microenvironment appears after this paragraph.

```
Micro number     = 1        !    Indoors - residence
Parameter Type   = AER
Condition # 1    = AvgTempCat
Condition # 2    = AC_Home
ResampHours      = NO
ResampDays       = YES
ResampWork       = YES
Block DType Season Area C1  C2  C3  Shape      Min   Max   Par1   Par2
1     1     1      1    1   1   1   Lognormal  .1    10    0.956  1.962
1     1     1      1    2   1   1   Lognormal  .1    10    0.517  2.017
1     1     1      1    3   1   1   Lognormal  .1    10    0.524  2.189
1     1     1      1    4   1   1   Lognormal  .1    10    0.392  2.076
1     1     1      1    5   1   1   Lognormal  .1    10    0.392  2.076
1     1     1      1    1   2   1   Lognormal  .1    10    0.754  2.317
1     1     1      1    2   2   1   Lognormal  .1    10    0.698  2.180
1     1     1      1    3   2   1   Lognormal  .1    10    1.367  2.292
1     1     1      1    4   2   1   Lognormal  .1    10    1.067  1.989
1     1     1      1    5   2   1   Lognormal  .1    10    1.067  1.989

Micro number     = 1
Parameter Type   = DE
ResampHours      = NO
ResampDays       = NO
ResampWork       = YES
Block DType Season Area C1  C2  C3  Shape      Min   Max   Par1   Par2
1     1     1      1    1   1   1   LogNormal  0.95  8.05  2.51   1.53
```

The first section of the excerpt specifies the air exchange rate distributions for the microenvironment. Average temperature and air conditioning presence, which are city-specific, were coded into air exchange rate conditional variables C1 and C2, respectively. Average temperatures were broken into five categories: less than 50 degrees F, 50 to 68, 68 to 77, 77 to 86, and 86 and above. Using data from several studies, air exchange rate estimates in the form of lognormal distributions were generated. These functions are specific to the cities in the model run. For cities with similar climatic and other relevant characteristics, the same distributions were used (e.g., New York, Philadelphia, and Boston use the same distributions). The data sources used and the development of these distributions are discussed in detail in Appendix A. The ozone decay rate is characterized as a lognormal distribution (as shown in the last section of the excerpt). The development of the decay rate is discussed in Section 3.9.3.

In the input file excerpt, there is a large block of numbers (which totals ten rows) in the air exchange rate portion of the file. In this block, the fifth number across, which falls under "C1" in the excerpt, represents the temperature. The code "C1" represents "Conditional Variable 1." In this range, the numeral one represents temperatures below 50 Fahrenheit, two represents temperatures from 50 to 68, three represents 68 to 77, four represents 77 to 86, and five represents 86 and above. The sixth number in this block, which falls under "C2" and ranges from one to two, represents air conditioning status, with the numeral one representing having an air conditioning, and two not having it. There are five distributions listed for each value, for a total of ten distributions. In the above example, there are actually four different distributions for each air conditioning setting; the last two distributions for each air conditioning setting (which represent temperature ranges from 77 to 86, and 86 and above) are the same.

An example of how this microenvironment would function may help to elucidate the code. For the city of Atlanta, it is estimated that 85 percent of the population has air conditioning in the home, and 15 percent does not (see Appendix A for more information on the origin of these data). These percentages are included in the Profile Functions file, which is discussed in Section 3.9. Using these percentages, APEX can stochastically generate air conditioning status for a profiled individual. In addition, APEX takes as input the daily average temperature in Atlanta. Based on the air conditioning status and the temperature, the appropriate one of the ten distributions listed is chosen for a particular profile. For example, if the profile had air conditioning and the average temperature was 70, the third row would be chosen to characterize the air exchange rate. If the profile had no air conditioning, and the average temperature was 90, the tenth row would be chosen.

Microenvironments 2-7: All other indoor microenvironments. The remaining five indoor microenvironments, which represent Bars and Restaurants, Schools, Day Care Centers, Office, Shopping, and Other environments, are all modeled using the same data and functions. The data and methodology for developing these functions are detailed in Appendix A. An excerpt from the input file describing one of these microenvironments is given on the next page.

As with the Indoor-Residence microenvironment, these microenvironments use both air exchange rates and decay rates to calculate exposures within the microenvironment. The air exchange rate distribution was developed based on an indoor air quality study (Persily et al, 2005). This research indicated that the lognormal distributions should provide effective modeling of ozone exposure. The decay rate is the same as used in the Indoor-Residence microenvironment, and is discussed in Section 3.9.3.

```
Micro number     = 2        !    Bars & restaurants
Parameter Type   = AER
ResampHours      = NO
ResampDays       = YES
ResampWork       = YES
Block DType Season Area C1  C2  C3   Shape      Min    Max    Par1   Par2
1      1          1          1     1     1    1  LogNormal  0.07   13.8   1.109  3.015

Micro number     = 2
Parameter Type   = DE
ResampHours      = NO
ResampDays       = YES
ResampWork       = YES
Block DType Season Area C1  C2  C3   Shape      Min    Max    Par1   Par2
1      1          1          1     1     1    1  LogNormal   0.95   8.05   2.51   1.53
```

Microenvironments 8 and 9: Outdoor microenvironments. Two outdoor microenvironments, the Near Road and Public Garage/Parking Lot environments, are different from the indoor microenvironments in that they use the factors method to calculate pollutant exposure. Proximity factors were developed to estimate exposures in these microenvironments. Penetration factors are not applicable to outdoor environments. An excerpt from the file describing this microenvironment follows this paragraph.

```
Micro number     = 8        !    Outdoor near road
Parameter Type   = PR
ResampHours      = YES
ResampDays       = YES
ResampWork       = YES
Block DType Season Area C1  C2  C3   Shape      Min    Max    Par1   Par2
1      1          1          1     1     1    1   Normal     0.422   1.0    0.755  0.203
```

The distribution for the proximity factor was developed from an ozone study (Johnson et al, 1995) conducted in the greater Cincinnati metropolitan area in August and September, 1994 (see Appendix A for details on this study). Vehicle tests were conducted according to an experimental design specifying the vehicle type, road type, vehicle speed, and ventilation mode.

Microenvironment 10: Outdoors-General. The general outdoor environment concentrations should be well represented by the ambient monitors. Therefore the penetration factor and proximity factor for this microenvironment were set to 1.

Microenvironments 11 and 12: In Vehicle- Cars and Trucks, and Mass Transit. Both of these microenvironments were calculated using the same values. These microenvironments use the factors method to calculate pollutant exposure. Both proximity factors and penetration factors were developed to estimate exposures in the microenvironments. Again, the optional concentration source variable is not relevant to ozone studies and was not used. An excerpt from the file describing this microenvironment follows this discussion.

The penetration factor distribution was developed using the inside-vehicle to outside-vehicle ratios from the Cincinnati ozone study previously mentioned (Johnson et al, 1995). Three proximity factor distributions were developed, one for local roads, one for urban roads, and one for interstates. The proportion of vehicle miles traveled in each city was estimated and used to weight the selection of the distributions. These weightings are included in the Profile Functions file, which is discussed in Section 3.11. Again, these distributions were developed based on the Cincinnati ozone study.

```
Micro number     = 11       !    Cars & trucks
Parameter Type   = PE
ResampHours      = YES
ResampDays       = YES
ResampWork       = YES
Block DType Season Area C1  C2  C3   Shape     Min    Max    Par1   Par2
1     1     1      1    1   1   1    Normal    0.1    1.0    0.300  0.232

Micro number     = 11
Parameter Type   = PR
Condition # 1    = Conditional1
ResampHours      = YES
ResampDays       = YES
ResampWork       = YES
Block DType Season Area C1  C2  C3   Shape     Min    Max    Par1   Par2
1     1     1      1    1   1   1    Normal    0.422  1.0    0.755  0.203
1     1     1      1    2   1   1    Normal    0.355  1.0    0.754  0.243
1     1     1      1    3   1   1    Normal    0.093  1.0    0.364  0.165
```

3.10.3 Ozone Decay and Deposition Rates

For this analysis, the same ozone decay rate distribution was used for all microenvironments that use the mass balance method. This distribution is based on data from an ozone decay study (Lee et al., 1999). This study measured decay rates in the living rooms of 43 residences in Southern California. Measurements of decay rates in a second room were made in 24 of these residences. The 67 decay rates range from 0.95 to 8.05 hour^{-1}. A lognormal distribution was fit to the measurements from this study, yielding a geometric mean of 2.5 and a geometric standard deviation of 1.5. These values are constrained to lie between 0.95 and 8.05 hour^{-1}. The specification of the decay rate (parameter type = DE) is shown in the file excerpts for the residential microenvironment and the other indoor microenvironments above.

3.10.4 Microenvironment Mapping

The *Microenvironment Mapping* file matches the APEX Microenvironments to CHAD Location codes. Table 3-6 gives the mapping used for the APEX simulations.

Table 3-6. Mapping of CHAD activity locations to APEX microenvironments

```
CHAD Loc.  Description                                 APEX micro
---------  -------------------------------------------------
U          Uncertain of correct code            =   -1  Unknown
X          No data                              =   -1  Unknown
30000      Residence, general                   =    1  Indoors-Residence
30010      Your residence                       =    1  Indoors-Residence
30020      Other residence                      =    1  Indoors-Residence
30100      Residence, indoor                    =    1  Indoors-Residence
30120      Your residence, indoor               =    1  Indoors-Residence
30121      ..., kitchen                         =    1  Indoors-Residence
30122      ..., living room or family room      =    1  Indoors-Residence
30123      ..., dining room                     =    1  Indoors-Residence
30124      ..., bathroom                        =    1  Indoors-Residence
30125      ..., bedroom                         =    1  Indoors-Residence
30126      ..., study or office                 =    1  Indoors-Residence
30127      ..., basement                        =    1  Indoors-Residence
30128      ..., utility or laundry room         =    1  Indoors-Residence
30129      ..., other indoor                    =    1  Indoors-Residence
30130      Other residence, indoor              =    1  Indoors-Residence
30131      ..., kitchen                         =    1  Indoors-Residence
30132      ..., living room or family room      =    1  Indoors-Residence
30133      ..., dining room                     =    1  Indoors-Residence
30134      ..., bathroom                        =    1  Indoors-Residence
30135      ..., bedroom                         =    1  Indoors-Residence
30136      ..., study or office                 =    1  Indoors-Residence
30137      ..., basement                        =    1  Indoors-Residence
30138      ..., utility or laundry room         =    1  Indoors-Residence
30139      ..., other indoor                    =    1  Indoors-Residence
30200      Residence, outdoor                   =   10  Outdoors-Other
30210      Your residence, outdoor              =   10  Outdoors-Other
30211      ..., pool or spa                     =   10  Outdoors-Other
30219      ..., other outdoor                   =   10  Outdoors-Other
30220      Other residence, outdoor             =   10  Outdoors-Other
30221      ..., pool or spa                     =   10  Outdoors-Other
30229      ..., other outdoor                   =   10  Outdoors-Other
30300      Residential garage or carport        =    7  Indoors-Other
30310      ..., indoor                          =    7  Indoors-Other
30320      ..., outdoor                         =   10  Outdoors-Other
30330      Your garage or carport               =    1  Indoors-Residence
30331      ..., indoor                          =    1  Indoors-Residence
30332      ..., outdoor                         =   10  Outdoors-Other
30340      Other residential garage or carport  =    1  Indoors-Residence
30341      ..., indoor                          =    1  Indoors-Residence
30342      ..., outdoor                         =   10  Outdoors-Other
30400      Residence, none of the above         =    1  Indoors-Residence
31000      Travel, general                      =   11  In Vehicle-Cars_and_Trucks
31100      Motorized travel                     =   11  In Vehicle-Cars_and_Trucks
31110      Car                                  =   11  In Vehicle-Cars_and_Trucks
31120      Truck                                =   11  In Vehicle-Cars_and_Trucks
31121      Truck (pickup or van)                =   11  In Vehicle-Cars_and_Trucks
31122      Truck (not pickup or van)            =   11  In Vehicle-Cars_and_Trucks
31130      Motorcycle or moped                  =    8  Outdoors-Near_Road
31140      Bus                                  =   12  In Vehicle-Mass_Transit
31150      Train or subway                      =   12  In Vehicle-Mass_Transit
```

```
31160      Airplane                              =    0  Zero_concentration
31170      Boat                                  =   10  Outdoors-Other
31171      Boat, motorized                       =   10  Outdoors-Other
31172      Boat, other                           =   10  Outdoors-Other
31200      Non-motorized travel                  =   10  Outdoors-Other
31210      Walk                                  =   10  Outdoors-Other
31220      Bicycle or inline skates/skateboard   =   10  Outdoors-Other
31230      In stroller or carried by adult       =   10  Outdoors-Other
31300      Waiting for travel                    =   10  Outdoors-Other
31310      ..., bus or train stop                =    8  Outdoors-Near_Road
31320      ..., indoors                          =    7  Indoors-Other
31900      Travel, other                         =   11  In Vehicle-Cars_and_Trucks
31910      ..., other vehicle                    =   11  In Vehicle-Cars_and_Trucks
32000      Non-residence indoor, general         =    7  Indoors-Other
32100      Office building/ bank/ post office    =    5  Indoors-Office
32200      Industrial/ factory/ warehouse        =    5  Indoors-Office
32300      Grocery store/ convenience store      =    6  Indoors-Shopping
32400      Shopping mall/ non-grocery store      =    6  Indoors-Shopping
32500      Bar/ night club/ bowling alley        =    2  Indoors-Bars_and_Restaurants
32510      Bar or night club                     =    2  Indoors-Bars_and_Restaurants
32520      Bowling alley                         =    2  Indoors-Bars_and_Restaurants
32600      Repair shop                           =    7  Indoors-Other
32610      Auto repair shop/ gas station         =    7  Indoors-Other
32620      Other repair shop                     =    7  Indoors-Other
32700      Indoor gym /health club               =    7  Indoors-Other
32800      Childcare facility                    =    4  Indoors-Day_Care_Centers
32810      ..., house                            =    1  Indoors-Residence
32820      ..., commercial                       =    4  Indoors-Day_Care_Centers
32900      Large public building                 =    7  Indoors-Other
32910      Auditorium/ arena/ concert hall       =    7  Indoors-Other
32920      Library/ courtroom/ museum/ theater   =    7  Indoors-Other
33100      Laundromat                            =    7  Indoors-Other
33200      Hospital/ medical care facility       =    7  Indoors-Other
33300      Barber/ hair dresser/ beauty parlor   =    7  Indoors-Other
33400      Indoors, moving among locations       =    7  Indoors-Other
33500      School                                =    3  Indoors-Schools
33600      Restaurant                            =    2  Indoors-Bars_and_Restaurants
33700      Church                                =    7  Indoors-Other
33800      Hotel/ motel                          =    7  Indoors-Other
33900      Dry cleaners                          =    7  Indoors-Other
34100      Indoor parking garage                 =    7  Indoors-Other
34200      Laboratory                            =    7  Indoors-Other
34300      Indoor, none of the above             =    7  Indoors-Other
35000      Non-residence outdoor, general        =   10  Outdoors-Other
35100      Sidewalk, street                      =    8  Outdoors-Near_Road
35110      Within 10 yards of street             =    8  Outdoors-Near_Road
35200      Outdoor public parking lot /garage    =    9  Outdoors-Public_Garage-Parking
35210      ..., public garage                    =    9  Outdoors-Public_Garage-Parking
35220      ..., parking lot                      =    9  Outdoors-Public_Garage-Parking
35300      Service station/ gas station          =   10  Outdoors-Other
35400      Construction site                     =   10  Outdoors-Other
35500      Amusement park                        =   10  Outdoors-Other
35600      Playground                            =   10  Outdoors-Other
35610      ..., school grounds                   =   10  Outdoors-Other
35620      ..., public or park                   =   10  Outdoors-Other
35700      Stadium or amphitheater               =   10  Outdoors-Other
35800      Park/ golf course                     =   10  Outdoors-Other
35810      Park                                  =   10  Outdoors-Other
35820      Golf course                           =   10  Outdoors-Other
35900      Pool/ river/ lake                     =   10  Outdoors-Other
36100      Outdoor restaurant/ picnic            =   10  Outdoors-Other
36200      Farm                                  =   10  Outdoors-Other
36300      Outdoor, none of the above            =   10  Outdoors-Other
```

3.11 Profile Functions

The *Profile Functions* file contains settings used to generate results for variables related to simulated individuals. While certain settings for individuals are generated automatically by APEX based on other input files, including demographic characteristics, others can be specified using this file. For example, the file may contain settings for determining whether the profiled individual has a car air conditioner, a gas stove, etc. The details and mechanics of this process are discussed in Section 2.3.2.

As discussed in Section 3.8.2, the *Profile Functions* file contains fractions indicating the prevalence of air conditioning in the cities modeled in this experiment. APEX uses these fractions to stochastically generate air conditioning status for profiled individuals. The derivation of this data is discussed in Appendix A. An excerpt from the file describing this microenvironment follows this paragraph.

```
AC_Home
! Has air conditioning at home
TABLE
INPUT1 PROBABILITY 2     “A/C probabilities”
0.85 0.15
RESULT INTEGER 2         “Yes/No”
1 2
#
```

One user-defined function was included in the *Profile Functions* file in order to reflect regional driving characteristics. The Conditional1 function is used to simulate In-vehicle penetration factors for modeled individuals. An excerpt from the file describing this microenvironment follows this paragraph.

```
Conditional1
! Penetration values for vehicles ME 11 and 12
TABLE
INPUT1 PROBABILITY 3
0.14 0.55 0.31
RESULT INTEGER 3
1 2 3
#
```

The function contains different distributions for three road types: urban, local, and interstate. These distributions model how the road type affects pollutant level penetration into the microenvironment. For each of the 12 locations modeled, the percentage of vehicle miles traveled on each road type was generated from 2003 Federal Highway Administration data (FHWA, 2004). These percentages are listed as fractions on the fifth line of the above excerpt. Using these percentages, the function allowed each of the distributions, which are defined in the microenvironment file, to be selected based on the amount of vehicle miles traveled in the area. See Appendix A for more information on development of the distributions in the microenvironments.

4. PRINCIPAL LIMITATIONS AND UNCERTAINTIES OF THE MODELING APPROACH

Inhalation exposure and risk modeling attempts to simulate real world conditions in order to accurately estimate exposures to pollutants and their resulting risk. In general, the methods and the model used in this assessment conform to the most contemporary modeling methodologies available. APEX is a powerful, highly customizable modeling system that allows for the realistic estimation of air pollutant exposure to individuals. Since it is based on human activity diaries and accounts for all important variables known to affect exposure, it has the ability to effectively approximate actual conditions. In addition, the data used to run the system were chosen because they were the best available to ensure realistic and defensible results. However, there are constraints and uncertainties with the modeling approach and the input data that limit the realism and accuracy of the model results.

4.1 Methodology

As described in Appendix A, several ozone and air pollution studies were reviewed, and data from these studies were used to develop the parameters and factors that were used to build the microenvironments in this assessment. A constraint on this effort is that there are limited ozone exposure studies. In addition, there are geographical limitations of the studies used to develop factors for this assessment. While these studies were generally performed in the geographical areas modeled in this assessment or in similar areas, there were differences that could lend uncertainty. For example, the ozone study (Johnson et al, 1995) that was used to develop proximity factors for in-vehicle microenvironments for all 12 cities was performed in Cincinnati. In addition, the air exchange rate distributions used for Boston, Chicago, Cleveland, and Philadelphia were developed from a study conducted in New York City. It is possible that climatic and other differences among these cities would produce different results. Scientific judgments were made in choosing appropriate data and information sources to best model ozone exposures. However, it is possible that despite best efforts there could be different interpretations about which data sources and methodologies are appropriate. Evaluation of several components of uncertainty in specification of CSA-specific air exchange rates are discussed below.

There are other areas of the modeling approach that have either assumptions or estimates that could affect results. For example, the microenvironments that are used in the program are matched to CHAD data. Because there are fewer microenvironments than CHAD locations, there is some information lost in this translation.

4.2 Input Data

Modeling results are heavily dependent on the quality of the data that are input to the system. The data for this analysis were selected in order to give the best opportunity to simulate actual conditions. One benefit of using well characterized data as inputs to the model is that limitations and other problems with the data are well understood. Still, the limitations and uncertainties of each of the data streams affect the overall quality of the model output. These issues and how they specifically affect each data stream are discussed in this section. The highest quality data streams are discussed first.

4.2.1 Meteorological Data

The least problematic of the data input to APEX are likely the meteorological data. These data are taken directly from monitoring stations in the assessment areas. One strength of these data is that it is relatively easy to see significant errors if they appear in the data. Because general climactic conditions are known for each area simulation, it would have been apparent upon review if there were outliers in the dataset. However, there are limitations in the use of these data. Because APEX only uses one temperature value per day, the model does not represent hour-to-hour variations in meteorological conditions throughout the day that may affect both ozone formation and exposure estimates within microenvironments.

4.2.2 Air Quality Data

The air quality data are taken directly from monitoring sites within each of the study areas, and thus the data are reliable and of high quality. Some data issues specific to air quality data result from the nature of pollutant formation and dispersion. Because many variables affect pollutant fate and transport, it is difficult to determine exactly how concentrations in the vicinity of a monitoring station may differ from the concentrations at other locations. Pollutant levels are highly dependent on weather and wind, and other unknowns may effect how well the data represent pollutant concentrations in the area. However, because APEX uses hourly average ozone concentrations, the model employs a temporally refined pollutant concentration record, which increases the accuracy of both ozone concentration and exposure estimates.

4.2.3 Population and Commuting Data

The population and commuting data are drawn from U.S. Census data from the year 2000. This is a high quality data source for nationwide population data in the U.S. However, the data do have limitations. The Census used random sampling techniques instead of attempting to reach all households in the U.S., as it has in the past. While the sampling techniques are well established and trusted, they introduce some uncertainty to the system. The Census has a quality section (http://www.census.gov/quality/) that discusses these and other issues with Census data.

In addition to these data quality issues, certain simplifying assumptions were made in order to better match reality or to make the data match APEX input specifications. For example, the APEX dataset does not differentiate people that work at home from those that commute within their home tract, and individuals that commute over 120 km a day were assumed to not commute daily. In addition to emphasizing some of the limitations of the input data, these assumptions introduce some uncertainty to the results. These issues were discussed in Sections 3.5 and 3.6.

4.2.4 Physiological Data

Because the physiological data were drawn from a sample, it is possible that they do not accurately mirror national physiological characteristics. Furthermore, on a larger scale, it is possible that national physiological characteristics have drifted somewhat since the publication of these data. For example, both the marked rise in obesity and ongoing national demographic shifts could result in some inaccuracies.

4.2.5 Activity Pattern Data

It is probable that the CHAD data used in the system is the most subject to limitations and uncertainty of all the data used in the system. Much of the data used to generate the daily diaries are over 20 years old. Table 3-3 indicates the ages of the CHAD diaries used in this modeling analysis. While the specifics of people's daily activities may not have changed much over the years, it is certainly possible that some differences do exist. In addition, the CHAD data are taken from numerous surveys that were performed for different purposes. Some of these surveys lasted only a day while others went on for weeks. Some of the studies were specifically designed not to be representative of the population at large in order to fulfill their specific mission when they were conducted. These issues affect the overall quality of the data that now resides in CHAD. An investigation on the sensitivity of the APEX results to the activity pattern database is discussed below.

4.2.6 Air Exchange Rates

There are several components of uncertainty in the CSA-specific residential air exchange rate distributions used for this analysis. Appendix D details an analyses of uncertainty due to extrapolation of air exchange rate distributions among CSAs , and of within-CSA uncertainty due to sampling variation. In Appendix E we describe an analysis of the uncertainty due to estimating daily air exchange rate distributions from air exchange rate measurements with varying averaging times. The results of those investigations are summarized here.

Extrapolation among cities CSA-specific distributions for use with the APEX ozone model were developed for 12 target CSAs, as detailed in Appendix A. Because we did not have CSA-specific data for all the 12 CSAs targeted in this analysis, for many we used data from another CSA or a combinations of other CSAs thought to have similar characteristics with respect to factors that might influence air exchange rates (see Table 4-1). Such factors include age composition of housing stock, construction methods, and other meteorological variables not explicitly treated in the analysis, such as humidity and wind speed patterns. In order to assess the uncertainty associated with this extrapolation, we investigated the between-CSA uncertainty by examining the variation of the geometric means and standard deviations across cities and studies.

The analysis showed a relatively wide variation across different cities in the air exchange rate geometric mean and standard deviation, stratified by air-conditioning status and temperature range. This implies that the air exchange rate modeling results would be very different if the matching of modeled CSAs to study CSAs was changed, although a sensitivity study using the APEX model would be needed to assess the impact on the ozone exposure estimates. For example, the ozone exposure estimates may be sensitive to the assumption that the St. Louis air exchange rate distributions can be represented by the combined non-California air exchange rate data. One way to address this would to perform a Monte Carlo analysis where the first stage is to randomly select a CSA outside of California, the second stage picks the air conditioning status, and the third stage picks the air exchange rate value from the assigned distribution for the CSA, air conditioning status and temperature range. Note that this will result in a very different distribution to the current approach for St. Louis that fits a single log-normal distribution to all the non-California data for a given temperature range and air conditioning status. The current approach weights each data point equally, so that CSAs like New York with most of the data values get the greatest statistical weight. The Monte Carlo approach gives the same total statistical weight for each CSA and fits a mixture of log-normal distributions rather than a single distribution.

Within CSA uncertainty In general, there is also some variation within studies for the same CSA, but this is much smaller than the variation across CSAs. This finding tends to support the approach of combining different studies for a CSA.

In addition, we assessed the within-city uncertainty by using a bootstrap distribution to estimate the effects of sampling variation on the fitted geometric means and standard deviations for each CSA. The bootstrap distributions assess the uncertainty due to random sampling variation but do not address uncertainties due to the lack of representativeness of the available study data or the variation in the lengths of the AER monitoring periods. The analysis showed that the geometric standard deviation uncertainty for a given CSA/air-conditioning-status/temperature-range combination tended to have a range of at most from "fitted GSD-1.0 hr^{-1}" to "fitted GSD+1.0 hr^{-1}", but the intervals based on larger AER sample sizes were frequently much narrower. The ranges for the geometric means tended to be approximately from "fitted GM-0.5 hr^{-1}" to "fitted GM+0.5 hr^{-1}", but in some cases were much smaller.

Table 4-1. Assignment of residential air exchange rate distributions to modeled CSAs

Modeled CSA	Air exchange rate distribution
Atlanta, GA, A/C	Research Triangle Park, A/C only
Atlanta, GA, no A/C	All non-California, no A/C ("Outside California")
Boston, MA	New York
Chicago, IL	New York
Cleveland, OH	New York
Detroit, MI	New York
Houston, TX	Houston
Los Angeles, CA	Los Angeles
New York, NY	New York
Philadelphia, PA	New York
Sacramento	Inland parts of Los Angeles ("Inland California")
St. Louis	All non-California
Washington, DC, A/C	Research Triangle Park, A/C only
Washington, DC, no A/C	All non-California, no A/C

Varying measurement averaging times Although the averaging periods for the air exchange rates in the study databases varied from one day to seven days, our analyses did not take the measurement duration into account and treated the data as if they were a set of statistically independent daily averages. To investigate the uncertainty of this assumption, we investigated the correlations between consecutive 24-hour air exchange rates measured at the same house from the Research Triangle Park Panel Study. The results showed extremely strong correlations, providing support for the simplified approach of treating multi-day averaging periods as if they were 24-hour averages.

However, this finding raises another issue. In the current version of the APEX model, there are several options for stratification of time periods with respect to air exchange rates distributions, and for when to re-sample from a distribution for a given stratum. The options selected for this current set of simulations resulted in a uniform air exchange rates for each 24-hour period and re-sampling of the 24-hour air exchange rates for each simulated day. This re-sampling for each simulated day implies that the simulated air exchange rates on consecutive days in the same microenvironment are statistically independent. Although we have not identified sufficient data to test the assumption of uniform air exchange rates throughout a 24-hour period, the analyses described in Appendix D suggest that air exchange rates on consecutive days are highly correlated. Therefore, we performed sensitivity simulations to assess the impact of the assumption of temporally independent air exchange rates, but found little difference between APEX predictions for the two scenarios (i.e., temporally independent and autocorrelated air exchange rates).

4.2.7 Air Conditioning Prevalence

Because the selection of an air exchange rate distribution is conditioned on the presence or absence of an air-conditioner, for each modeled CSA, the air conditioning status of the residential microenvironments is simulated randomly using the probability that a residence has an air conditioner, i.e., the residential air conditioner prevalence rate. For this study we used CSA-specific data from the American Housing Survey of 2003. Appendix F details the specification of uncertainty estimates in the form of confidence intervals for the air conditioner prevalence rate, and compares these with prevalence rates and confidence intervals developed from the Energy Information Administration's Residential Energy Consumption Survey of 2001 for more aggregate geographic subdivision (e.g., states, multi-state Census regions).

Air–conditioning prevalence rates for the 12 target CSAs from the American Housing Survey ranged from 55% for Los Angeles to 97% for Atlanta. Reported standard errors were relatively small, ranging from less than 1% for Houston to 3.4% for Cleveland. The corresponding 95% confidence interval spans range from approximately 4% to 14%.

4.2.8 Evaporative Coolers

Some residences use evaporative coolers, also known as "swamp coolers," for cooling. In our estimation of air exchange rate distributions from measurement data, we did not take into account the presence or absence of an evaporative cooler.

Although both the housing surveys discussed in section 4.2.7 specifically exclude evaporative coolers from their definitions of an air conditioner, it is plausible that the air exchange rate distributions might also depend upon the presence of an evaporative cooler. To evaluate this issue, Appendix F also details a comparison of the air exchange rate distributions estimated with and without accounting for the presence or absence of an evaporative cooler, using the available data from three air exchange rate measurement studies. The analysis showed no improvement in the statistical air exchange model when the data were also stratified by evaporative cooler presence or absence, given that they are already stratified by CSA, air conditioner presence or absence, and temperature range.

5. RESULTS OF EXPOSURE MODELING

5.1 Base case

In this section we present APEX results for population subgroups of interest in Boston and Houston for 2002 and 2004 as examples. Other CSAs show similar results. The population subgroups are:

- Children
- Active children
- Asthmatic children

Figures 5-1 and 5-2 show the APEX estimates of the number of person-days of exposure to exceedances of various 8-hour average ozone exposure concentrations during moderate exertion in Boston for 2002 and 2004 conditions respectively. Comparison of the figures indicate generally higher exposure levels in 2002 with a 95th percentile value of approximately 0.055 ppm-8hr and a maximum of 0.13 ppm-8hr. The corresponding values for 2004 are approximately 0.045 ppm-8hr and 0.09 ppm-8hr, respectively.

Figures 5-3 and 5-4 show the same APEX estimates for Houston, where exposure levels for 2002 are only slightly higher than for 2004. For example, the 95th percentile value for both years is approximately 0.045 ppm-8hr and the maxima are 0.13 and 0.12 ppm-8hr, respectively.

The remainder of the discussion of APEX estimates will be confined to 2002 only.

Figures 5-5 through 5-7 show the number of persons in Boston with 8-hour average ozone exposure concentrations above the indicated level and the number of exceedances per person, during moderate exertion for 2002 conditions. Figures 5-8 through 5-10 show the same APEX estimates for Houston.

More than 40% of each child sub-population in Boston (i.e., children, active children, asthmatic children) is estimated to be exposed to at least one exceedance of an 8-hour average ozone concentration of 0.07 ppm during moderate exertion, and more than 7% are estimated to be exposed to at least 3 exceedances during moderate exertion.

The proportions are somewhat lower in Houston. For the children's subpopulations the proportion with at least one exceedance of 0.07 ppm during moderate exertion is estimated to be between 25% and 30%, and with at least 3 exceedances between 1% and 2%.

Figure 5-1

Person-Days of Exposure to 8-Hour Average Concentrations During Moderate Exertion
--Boston 2002--

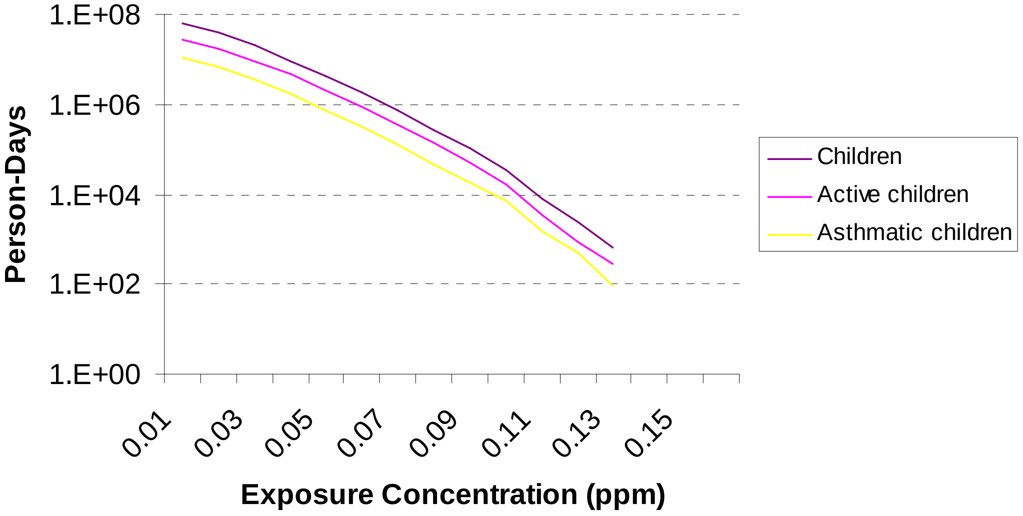


Figure 5-2

Person-Days of Exposure to 8-Hour Average Concentrations During Moderate Exertion
--Boston 2004--

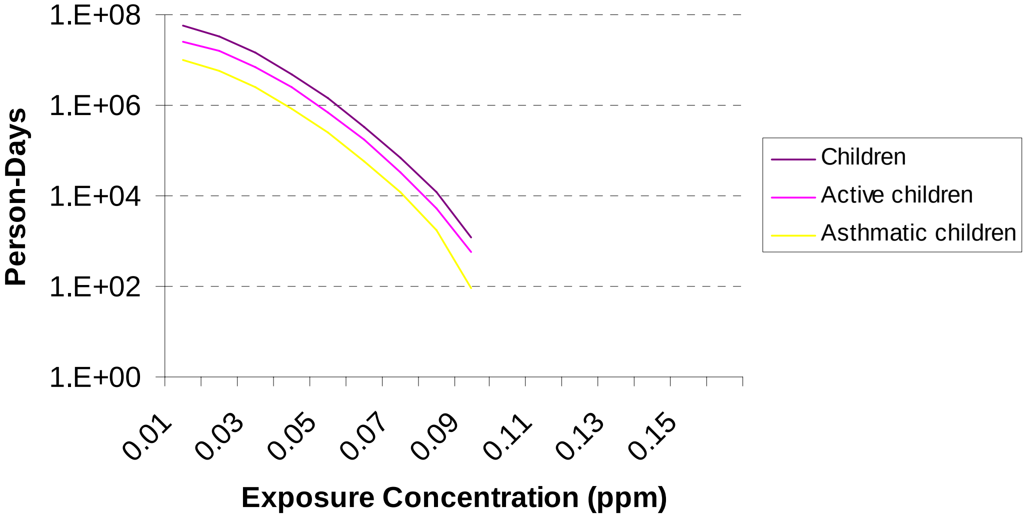


Figure 5-3

Person-Days of Exposure to 8-Hour Average Concentrations During Moderate Exertion
--Houston 2002--

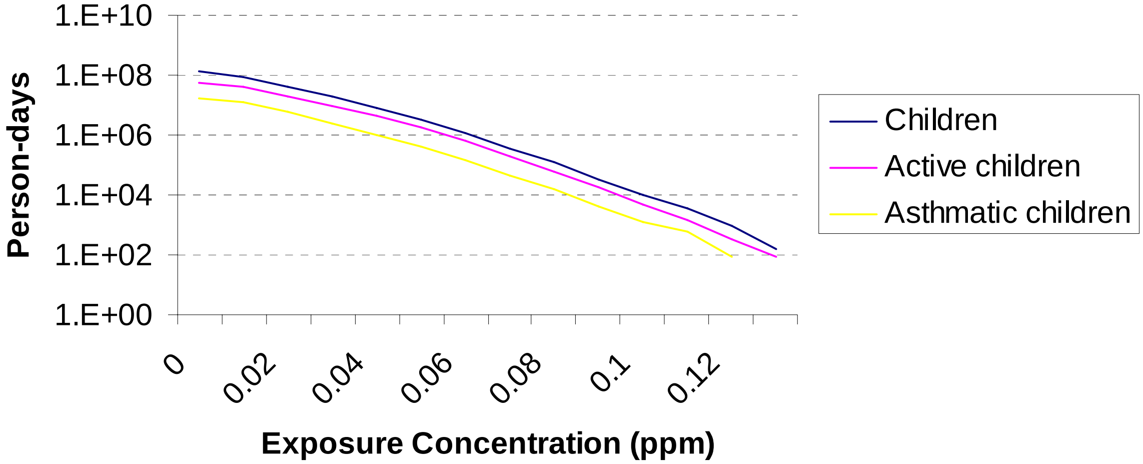


Figure 5-4

Person-Days of Exposure to 8-Hour Average Concentrations During Moderate Exertion
--Houston 2004--

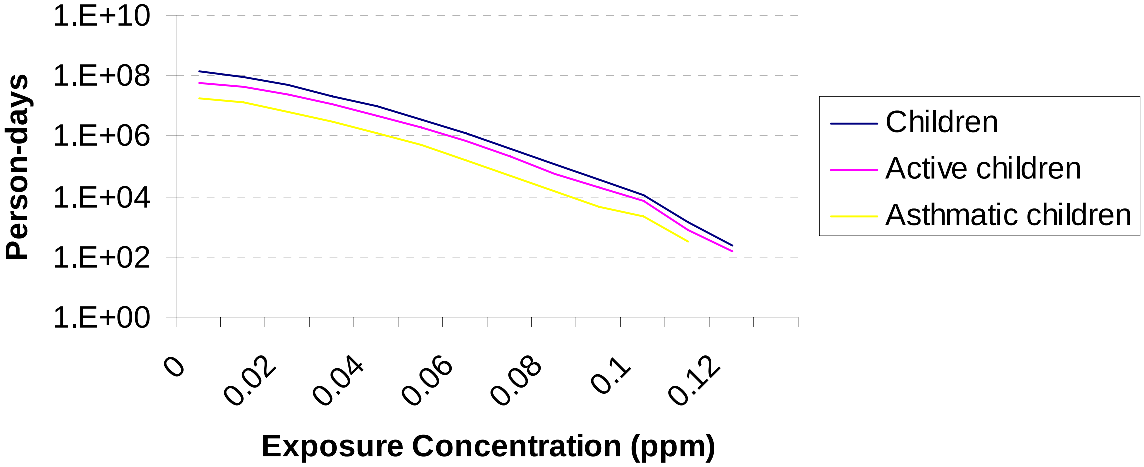


Figure 5-5

Exceedances of 8-Hour Average Exposure Concentrations During Moderate Exertion --Children, Boston, 2002--

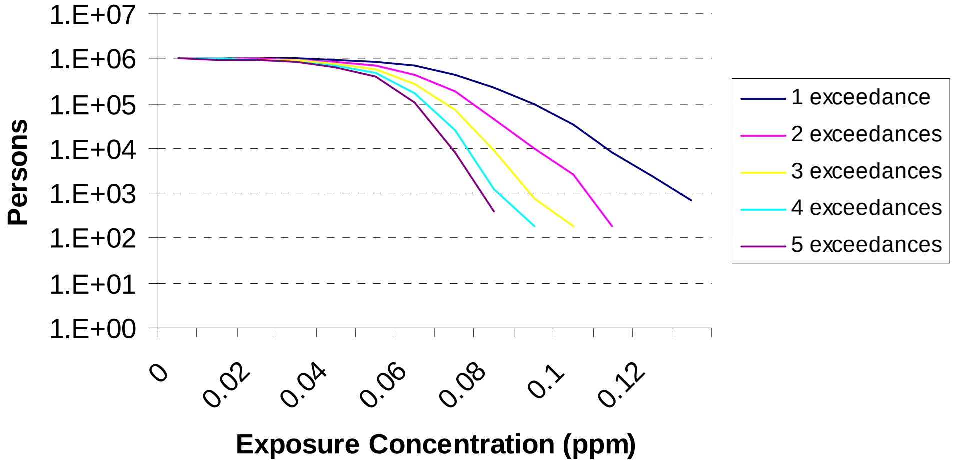


Figure 5-6

Exceedances of 8-Hour Average Exposure Concentrations During Moderate Exertion --Active Children, Boston, 2002--

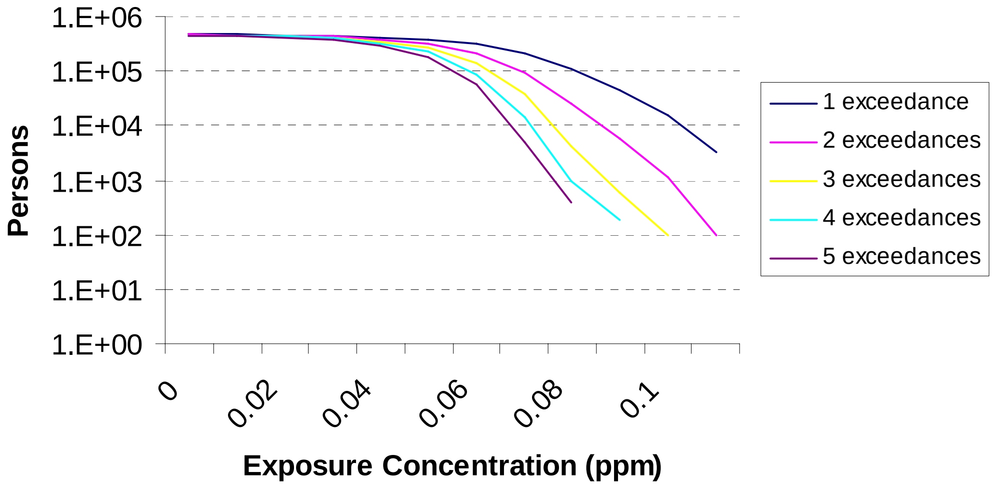


Figure 5-7

Exceedances of 8-Hour Average Exposure Concentrations During Moderate Exertion --Asthmatic Children, Boston, 2002--

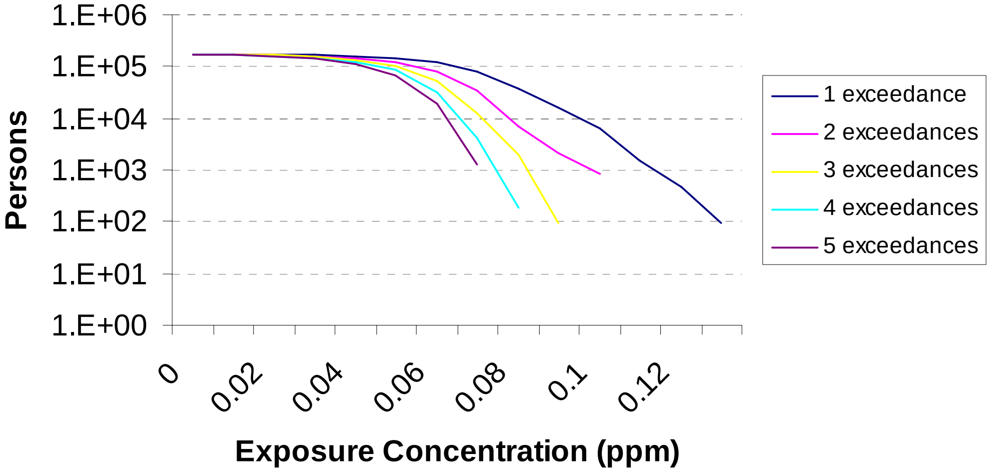


Figure 5-8

Exceedances of 8-Hour Average Exposure Concentrations During Moderate Exertion --Children, Houston, 2002--

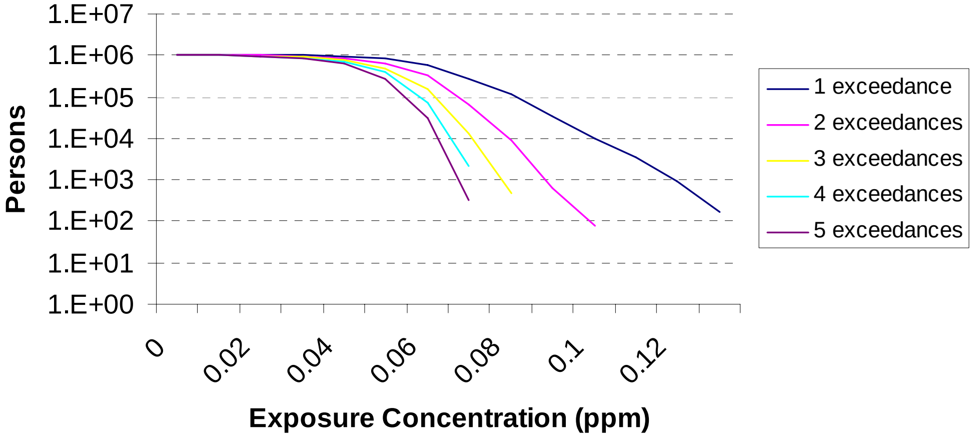


Figure 5-9

Exceedances of 8-Hour Average Exposure Concentrations During Moderate Exertion
--Active Children, Houston, 2002--

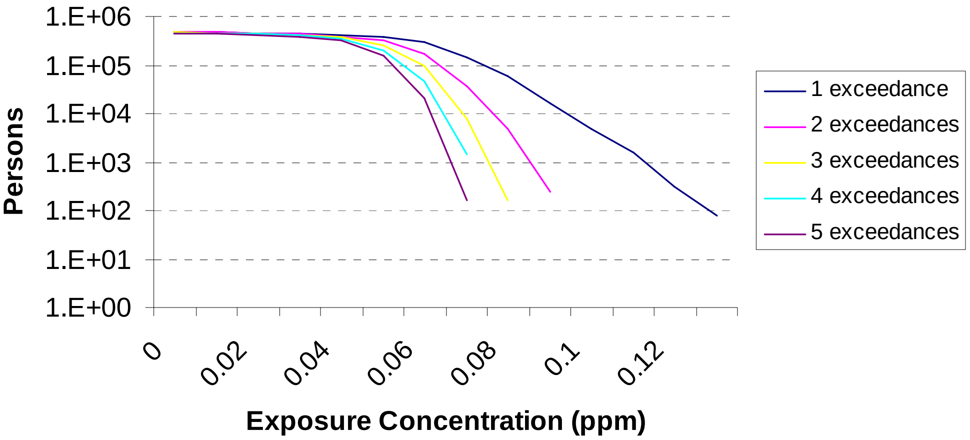
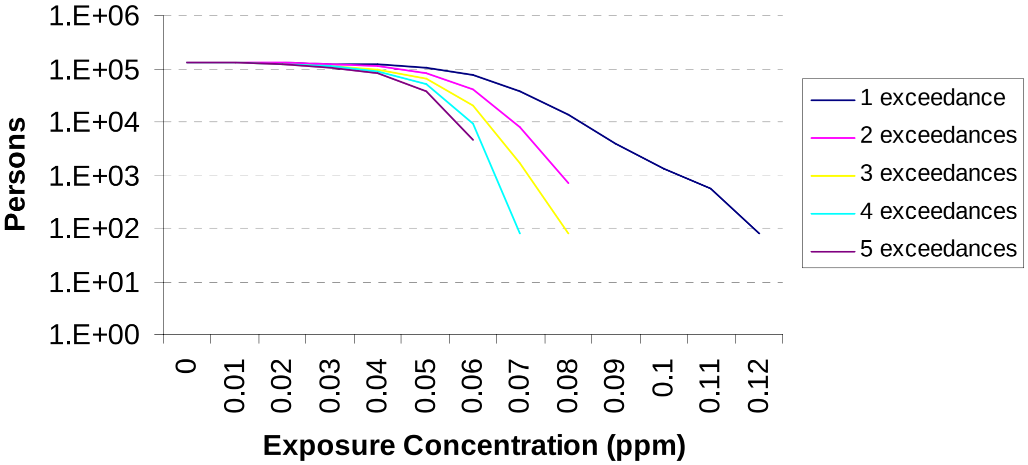


Figure 5-10

Exceedances of 8-Hour Average Exposure Concentrations During Moderate Exertion
--Asthmatic Children, Houston, 2002--



5.2 Attainment scenarios

The exposure projections presented in this section are based on the air quality projections discussed in section 3.3. Due to the uncertainties in the air quality projections, these exposure projections should be viewed as demonstrating the general range of possible changes in population exposure due to implementation of alternative NAAQS standards. They should not be considered as CSA-specific projections.

Figures 5-11 through 5-13 present APEX projections of person-days of exposure for the 3 children's subpopulations in Boston given attainment of various alternative NAAQS standards described in Section 3.3 above and 2002 meteorological conditions. Each figure contains a set of curves corresponding to potential air quality standards specified as various combinations of concentration thresholds and number of exceedances allowed. Table 3-2 explains the meaning of the labels "85/4," etc.

The figures indicate a reduction in exposures under the alternative standards, with the 95th percentile 8-hour average exposure concentration decreasing from approximately 0.055 ppm (2002 base case) to a range of 0.040 to 0.045 ppm, depending on the alternative standard. Similarly, the maximum 8-hour average exposure concentration decreases from 0.13 ppm (2002 base case) to a range of 0.08 to 0.11 ppm.

Figures 5-14 through 5-16 present the corresponding information for Houston. In this case the 95th percentile 8-hour average exposure concentration decreases from approximately 0.045 ppm to approximately 0.035 ppm and the maximum exposure concentrations decreases from 0.13 ppm to a range of 0.07 to 0.09 ppm, depending on the alternative standard.

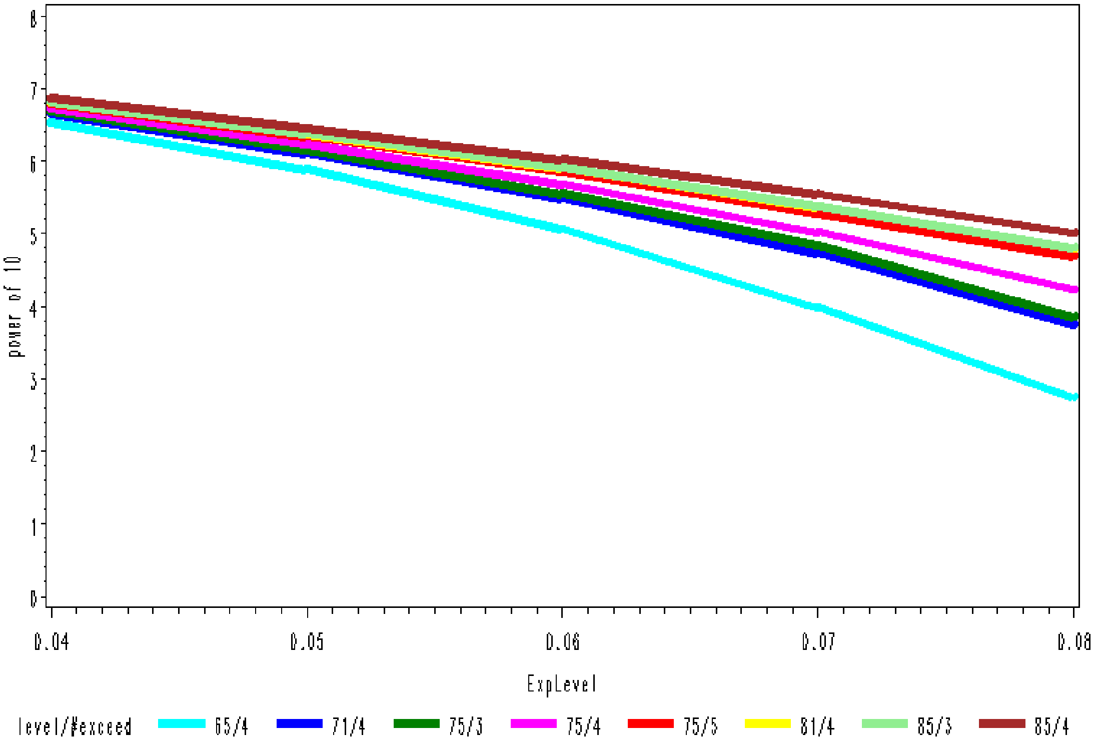
Figures 5-17 through 5-19 present APEX projections for the number of Boston children exposed to at least 3 exceedances of various 8-hour average concentrations given attainment of each of the several alternative NAAQS standards. In general the figures suggest the changing the threshold concentration would have more influence on this measure of population exposure than changing the number of permitted exceedances.

The figures show that each of the alternative standards is estimated to reduce the proportion of the population subgroup exposed at least 3 times to 8-hour average concentrations exceeding 0.07 ppm from about 7% (2002 base case) to between 0% and 2% depending on the alternative standard.

Similar information is presented for Houston is Figures 5-20 through 5-22. These indicate that all of the alternative standards would be expected to reduce the proportion of the population subgroup exposed at least 3 times to 8-hour average concentrations exceeding 0.07 ppm from about 2% (base case) to 0%.

Figure 5-11

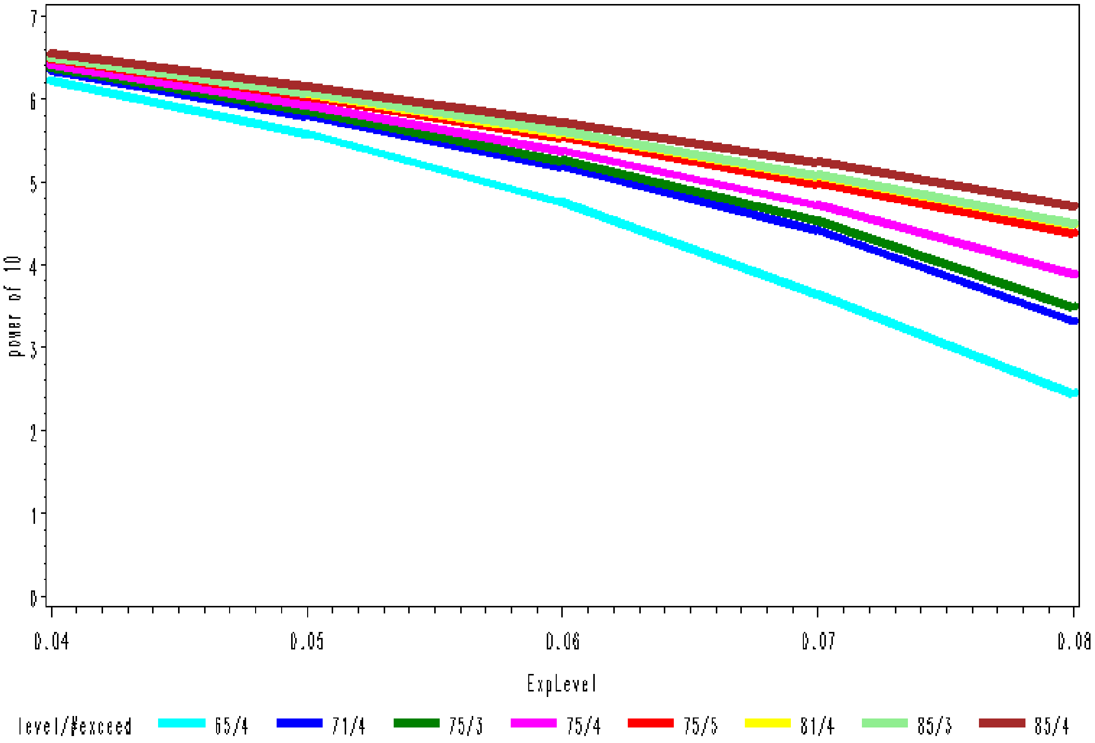
Person-days (occurrences) above daily maximum 8-hour exposure levels
For different alternative standards (ppb)
Group=Children, moderate exertion CSA=Boston year=2002



--

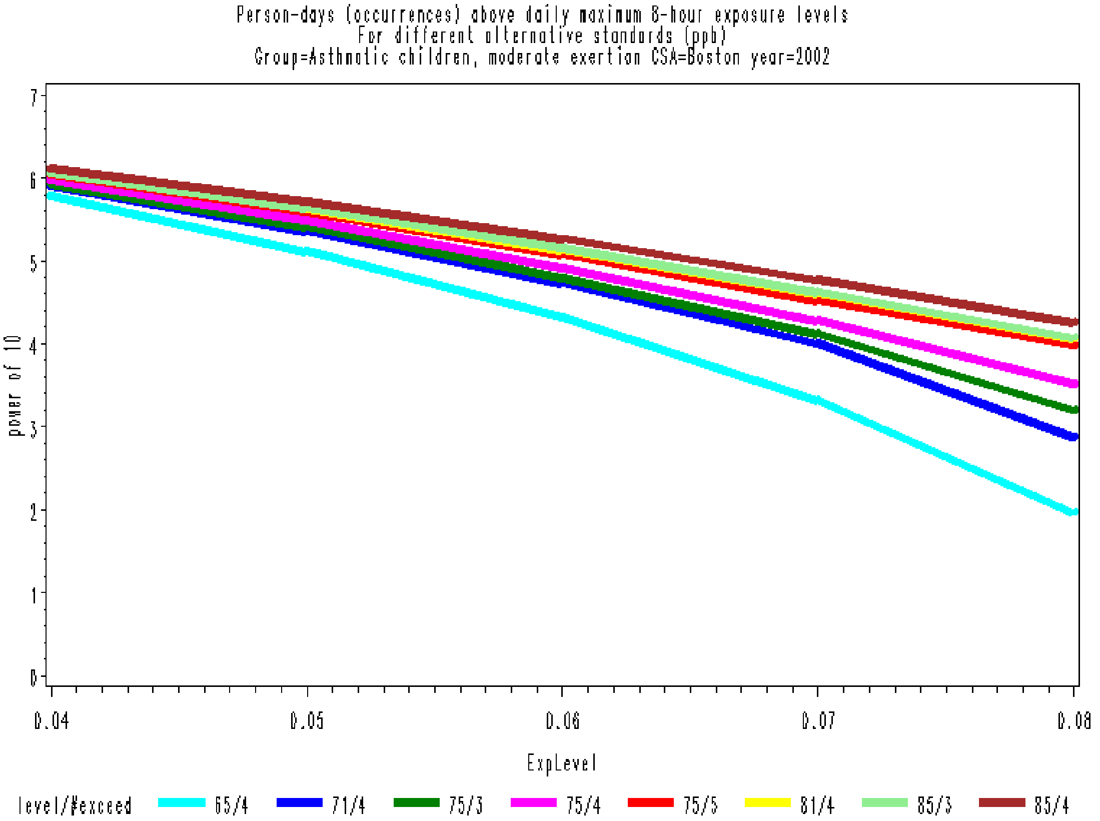
Figure 5-12

Person-days (occurrences) above daily maximum 8-hour exposure levels
For different alternative standards (ppb)
Group=Active children, moderate exertion CSA=Boston year=2002



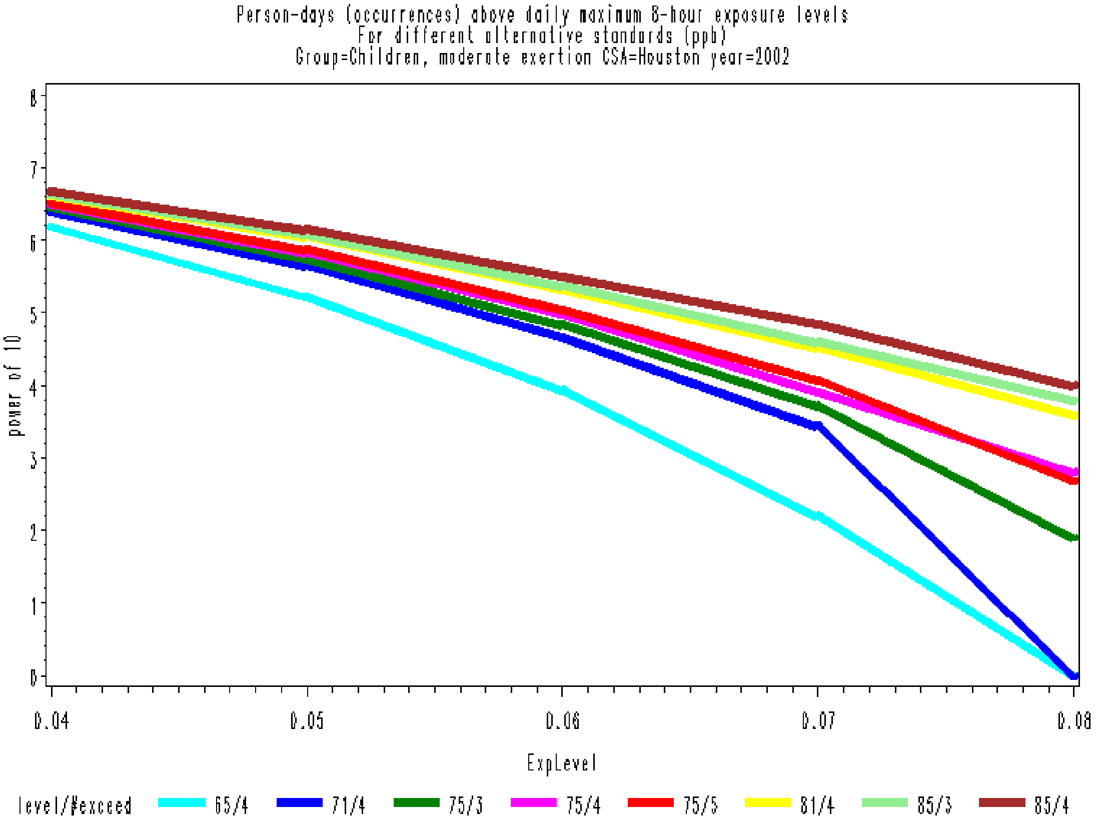
--

Figure 5-13



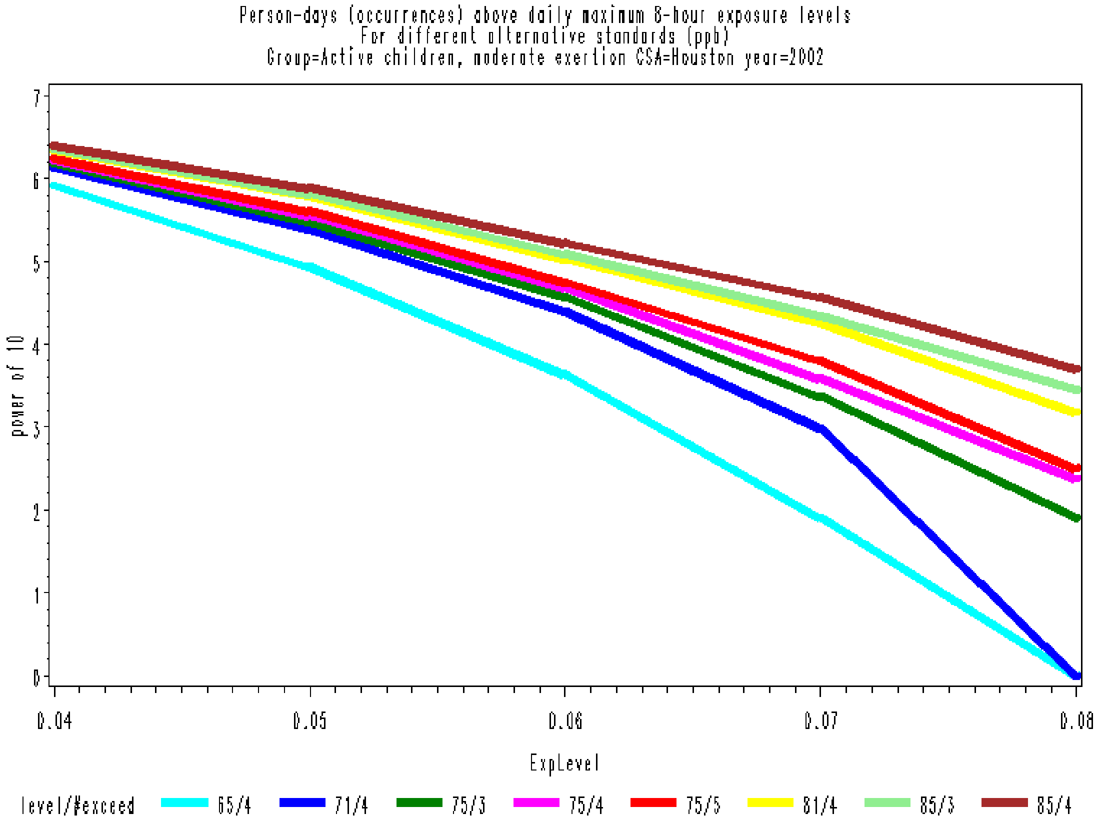
--

Figure 5-14



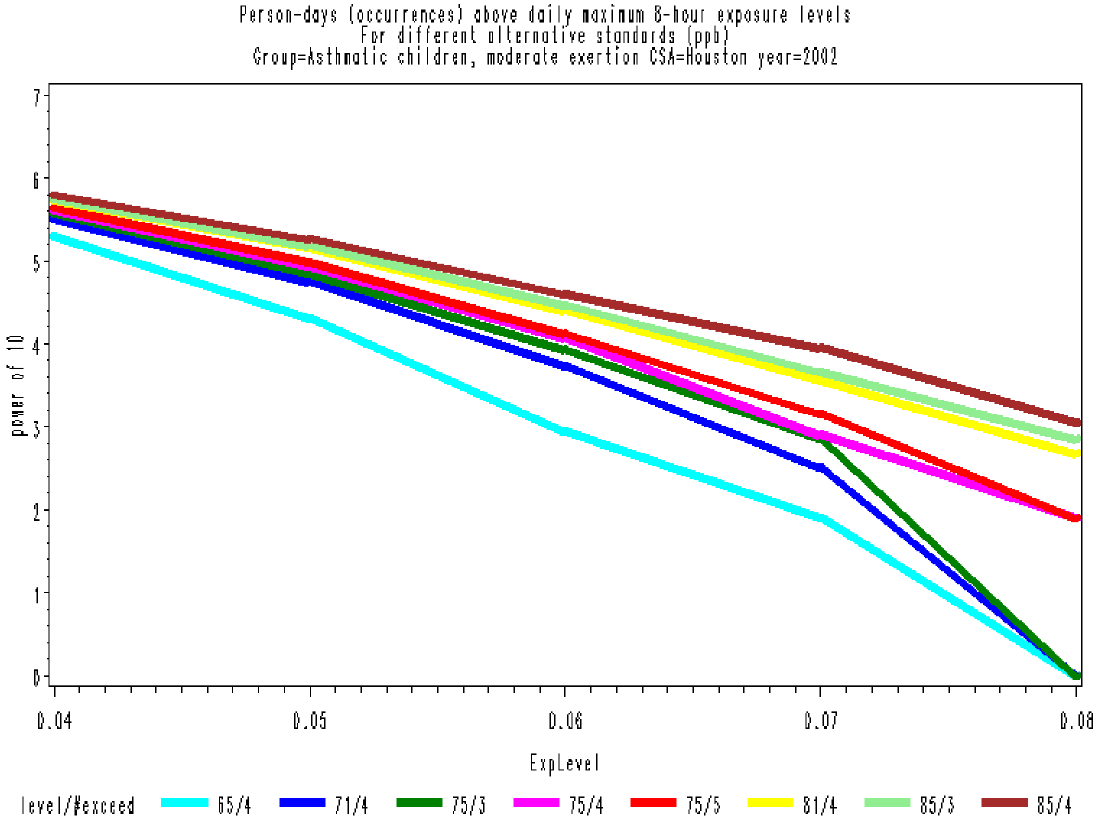
--

Figure 5-15



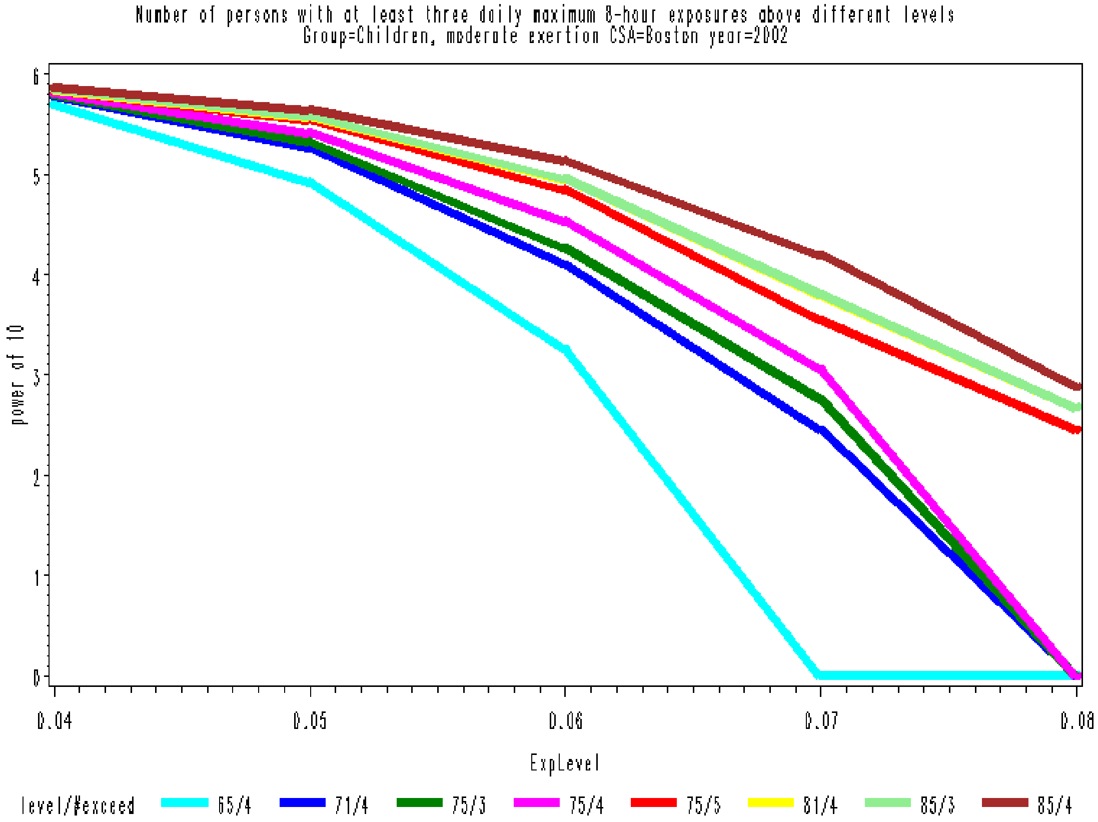
--

Figure 5-16



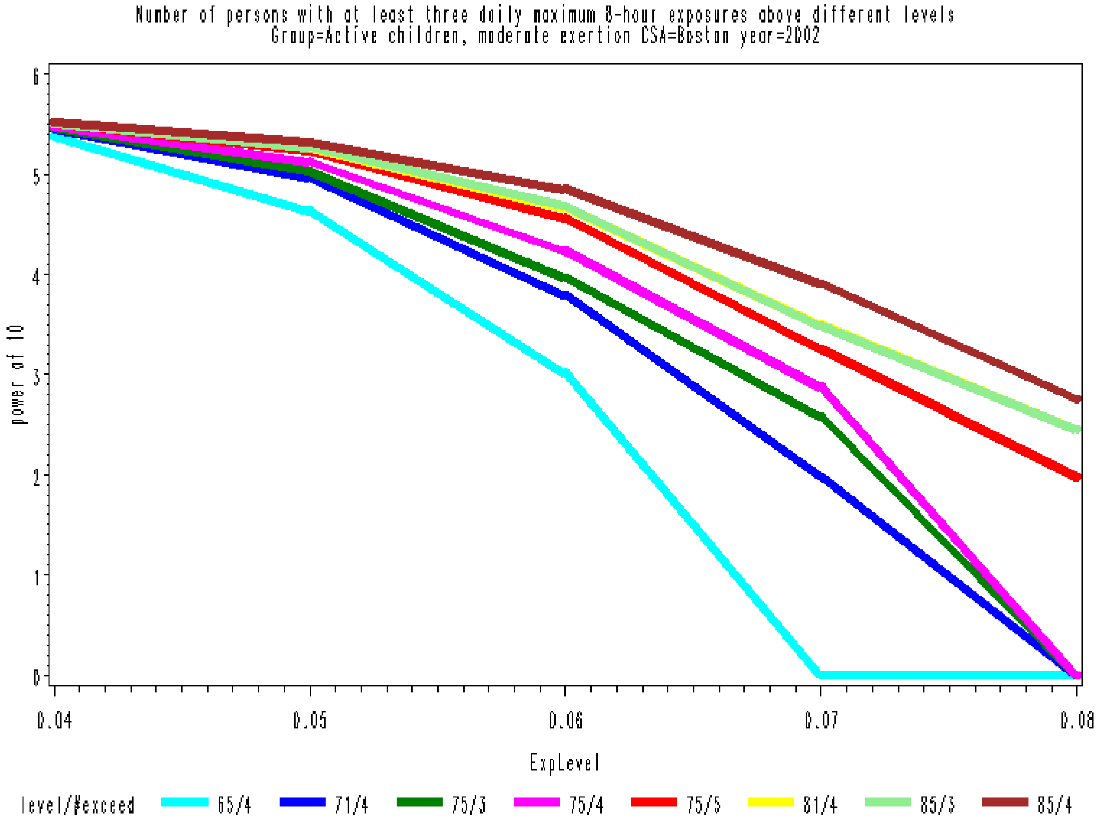
--

Figure 5-17



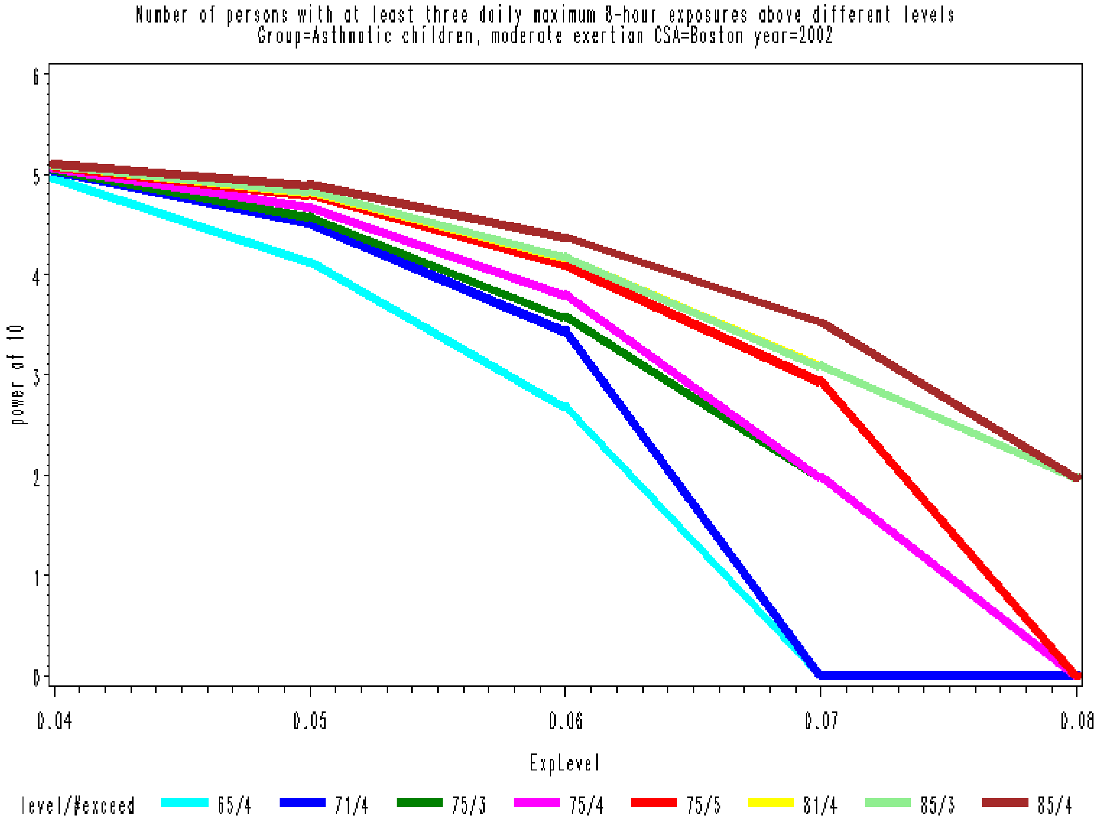
--

Figure 5-18



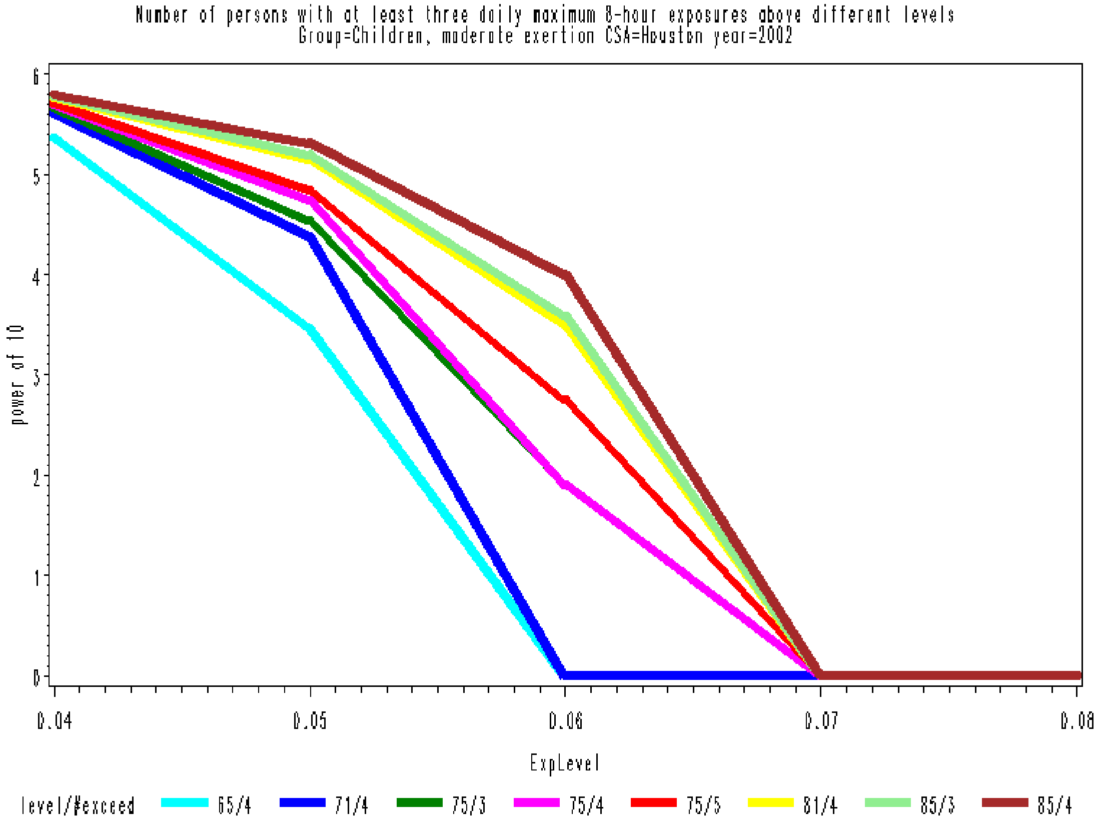
--

Figure 5-19



--

Figure 5-20



--

Figure 5-21

Number of persons with at least three daily maximum 8-hour exposures above different levels
Group=Active children, moderate exertion CSA=Houston year=2002

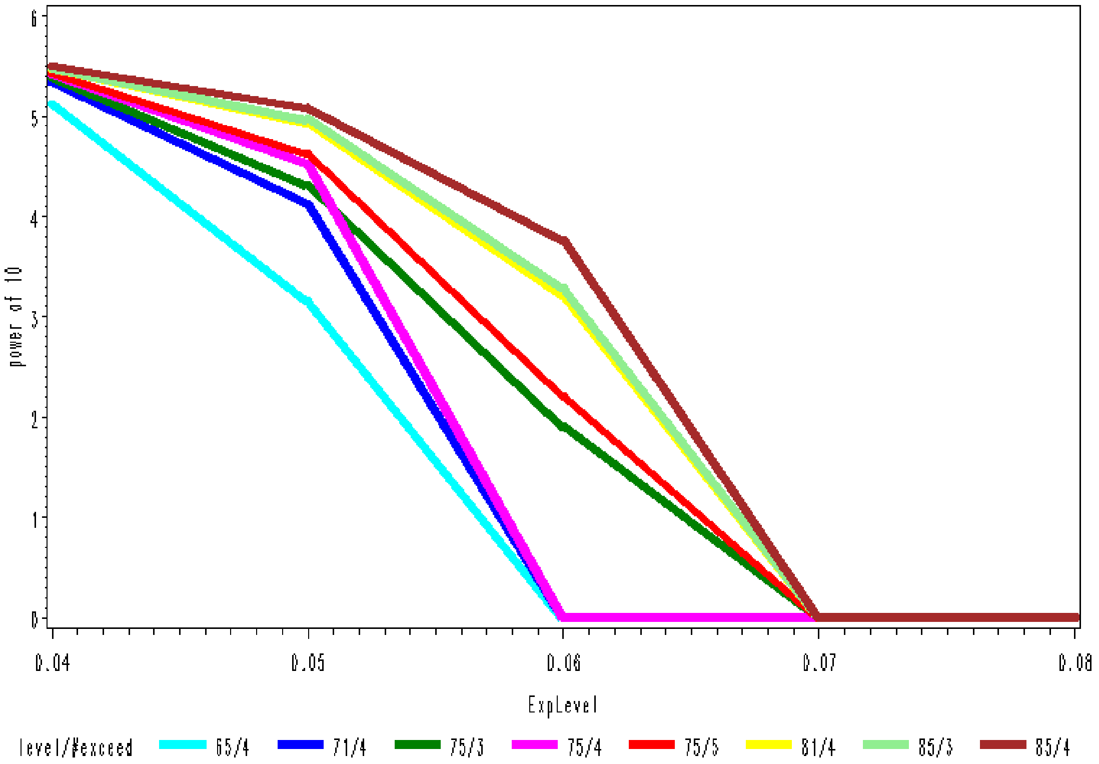
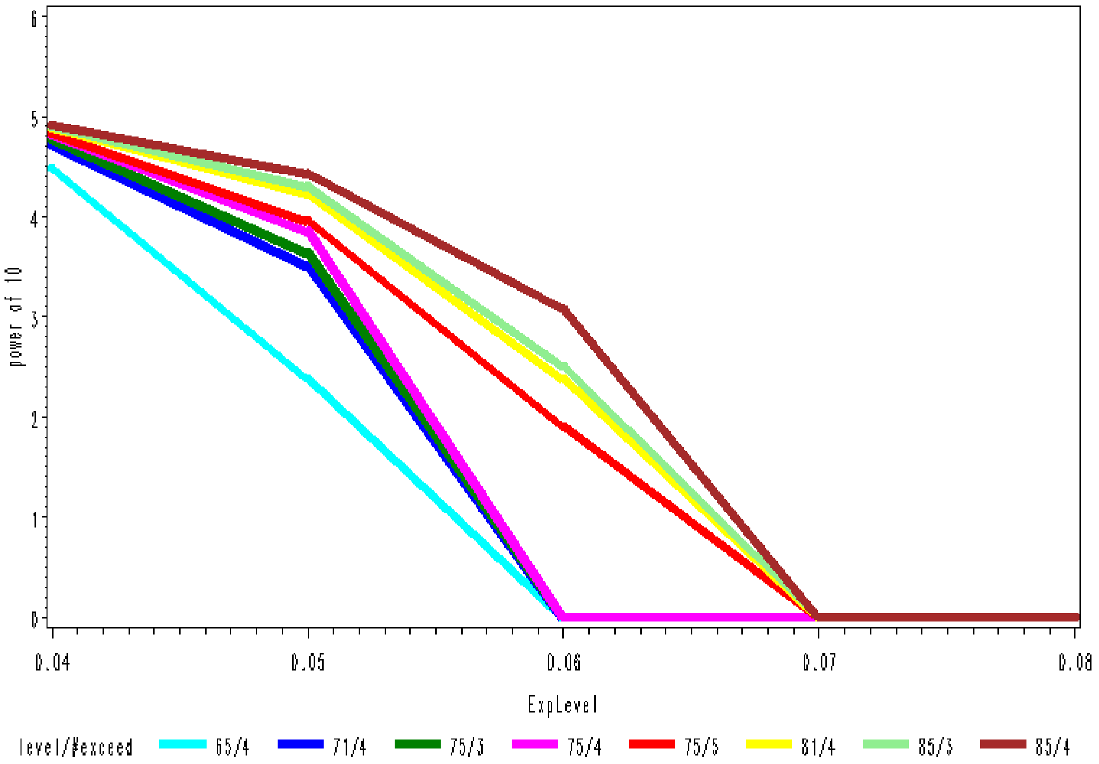


Figure 5-22

Number of persons with at least three daily maximum 8-hour exposures above different levels
Group=Asthmatic children, moderate exertion CSA=Houston year=2002



This page intentionally left blank.

6. SENSITIVITY STUDIES

In order to assess the sensitivity of the APEX predictions to some of the uncertain data inputs, several additional simulations were conducted for 2002 conditions. The inputs analyzed were the activity pattern database, the ozone decay rate, the proximity factor, and the air exchange rate. In addition, we evaluated the impact of the new method for constructing long-term activity patterns from short-term records.

6.1 Activity Pattern Database

Because many of the studies included in the CHAD database are not national in scope, nor do they necessarily correspond to the CSAs targeted here, it would be useful to know how similar the component studies are. Strong similarity would suggest that extrapolation of activity data gathered from one sample population to another population is appropriate.

The most comprehensive individual study is probably the National Human Activity Pattern Study (NHAPS), so we compared the base case exposure results with corresponding results using only the NHAPS data. The results for all persons at all activity levels in Boston are presented in Figure 6-1 and for active persons with moderate activity in Boston are presented in Figure 6-2. The figures present several pairs of cumulative distribution functions for the number of days/person that a given 8-hour average threshold concentration is exceeded. The different curves represent different thresholds, ranging from 0.01 to 0.05 ppm-8hr, as indicated in the figure legends. Figure 6-1 shows little difference between the base case distributions and those for NHAPS activity patterns only, indicating that both the average number of exceedances and the variability among individuals are similar. This suggests that the composite database is similar to NHAPS.

However, Figure 6-2 does show moderate differences, with the NHAPS results systematically lower than the base case. This suggests that the activity patterns for active people in the NHAPS data base may differ somewhat from those in the other component data bases of CHAD.

Tables 6-1 through 6-4 present the results for the number of persons exposed to 8-hour average concentrations exceeding 0.07 ppm in the general population and for children, respectively, with moderate exertion. Simulations were performed both with the base case air quality and with a scenario of attainment of the current NAAQS.

The results show that use of the NHAPS database leads to exposure predictions systematically higher than use of the complete CHAD database. The percentage differences are generally small when there are relatively high numbers of exposures, e.g., for one or more exposures with base case air quality. As the number of exposures decreases the percentage differences increase. The

Table 6-1. Sensitivity to activity database with base case air quality: 2002 counts of the general population with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

CSA	One or more exposures			Three or more exposures		
	All CHAD	NHAPS only	Difference	All CHAD	NHAPS only	Difference
Atlanta	822,480	888,045	+8%	68,900	104,449	+52%
Boston	1,183,742	1,245,743	+5%	154,575	18,9909	+23%
Chicago	1,890,859	2,131,852	+13%	169,145	248,287	+47%
Cleveland	887,472	953,496	+7%	208,430	249,517	+20%
Detroit	1,268,931	1,418,832	+12%	161,150	218,646	+36%
Houston	691,705	776,764	+12%	27,765	48,708	+75%
Los Angeles	3,106,438	3,280,518	+6%	744,072	904,782	+22%
New York	5,439,379	5,710,262	+5%	958,948	1,130,519	+18%
Philadelphia	1,793,197	1,817,305	+1%	464,655	502,857	+8%
Sacramento	389,381	433,736	+11%	54,905	76,036	+38%
St. Louis	752,850	777,542	+3%	112,815	132,000	+17%
Washington DC	1,991,336	2,065,664	+4%	338,578	395,870	+17%

Table 6-2. Sensitivity to activity database with base case air quality: 2002 counts of children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

CSA	One or more exposures			Three or more exposures		
	All CHAD	NHAPS only	Difference	All CHAD	NHAPS only	Difference
Atlanta	324,869	357,007	+10%	33,048	54,877	+66%
Boston	444,963	446,963	+0%	73,811	80,859	+10%
Chicago	773,413	814,536	+5%	90,935	105,677	+16%
Cleveland	340,675	343,178	+1%	106,228	108,977	+3%
Detroit	513,893	542,374	+6%	87,583	100,440	+15%
Houston	284,225	304,447	+7%	13,561	22,549	+66%
Los Angeles	1,285,413	1,273,135	-1%	375,719	410,099	+9%
New York	2,033,582	2,052,804	+1%	494,780	504,391	+2%
Philadelphia	671,708	657,613	-2%	248,755	232,230	-7%
Sacramento	150,240	158,602	+6%	24,799	32,615	+32%
St. Louis	291,676	291,125	-0%	60,079	66,688	+11%
Washington DC	748,707	758,424	+1%	177,302	182,476	+3%

Table 6-3. Sensitivity to activity database with air quality meeting the current standard: 2002 counts of the general population with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

CSA	One or more exposures			Three or more exposures		
	All CHAD	NHAPS only	Difference	All CHAD	NHAPS only	Difference
Atlanta	349,048	415,523	+19%	8,338	14,553	+75%
Boston	678,682	730,874	+8%	33,048	49,525	+50%
Chicago	832,692	950,007	+14%	17,380	24,208	+39%
Cleveland	443,809	513,466	+16%	32,693	49,825	+52%
Detroit	677,454	792,000	+17%	23,302	38,212	+63%
Houston	157,519	194,352	+23%	401	963	+140%
Los Angeles	93,589	133,698	+43%	1,910	4,093	+114%
New York	1,487,900	1,799,719	+21%	44,851	69,768	+56%
Philadelphia	956,236	1,015,144	+6%	87,098	10,707	+24%
Sacramento	82,984	110,935	+34%	1,866	3,892	+109%
St. Louis	538,694	567,150	+5%	40,757	56,040	+38%
Washington DC	961,596	1,058,008	+10%	53,885	71,047	+32%

Table 6-4. Sensitivity to activity database with air quality meeting the current standard: 2002 counts of children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

CSA	One or more exposures			Three or more exposures		
	All CHAD	NHAPS only	Difference	All CHAD	NHAPS only	Difference
Atlanta	138,785	177,821	+28%	2,956	6,215	+110%
Boston	263,244	273,911	+4%	15,810	23,048	+46%
Chicago	354,119	364,826	+3%	10,087	10,863	+8%
Cleveland	184,180	203,718	+11%	15,561	23,219	+49%
Detroit	285,784	315,514	+10%	11,963	17,052	+43%
Houston	66,603	81,127	+22%	0	241	NA
Los Angeles	30,560	46,658	+53%	273	546	+100%
New York	580,566	673,827	+16%	19,578	31,680	+62%
Philadelphia	390,485	398,067	+2%	49,479	56,575	+14%
Sacramento	27,887	39,691	+42%	450	1,351	+200%
St. Louis	213,697	216,175	+1%	21,113	28,410	+34%
Washington DC	381,988	405,713	+6%	24,103	29,151	+21%

An additional set of sensitivity simulations was conducted to test the differences between using the complete CHAD database and using a regional-specific subset corresponding to the CSA of interest. The Los Angeles and Sacramento CSAs were selected for the test, because of the relative abundance of activity pattern data for California (see Table 3-3).

Tables 6-5 and 6-6 present the results for Los Angeles and Sacramento, respectively, for the number of persons exposed to 8-hour average concentrations exceeding 0.07 ppm in the general population and for children with moderate exertion. Again, simulations were performed both with the base case air quality and with a scenario of attainment of the current NAAQS.

As explained in Section 2.3.3, the first step in constructing a multi-day activity sequence for a simulated individual is the stratification of the activity pattern data by day-type and ambient temperature. For the simulations reported here the activity pattern database was stratified into 6 pools, each defined by a combination of day type (weekday or weekend) and ambient temperature (maximum temperature < 55, 55-84, >84 °F). However, the California activity used alone data were insufficient for stratification into these 6 pools. Instead, 2 weekend pools (i.e., temperature = 55–84 °F, and >84 °F) were combined. In order to make a fair comparison with the All-CHAD case, for these simulations we stratified the All CHAD database into the same 5 pools. Tables 6-5 and 6-6 show the results for All CHAD using both the 6 pool and the 5 pool stratifications for reference.

As was the case for Tables 6-1 through 6-4, the percentage differences are generally small when there are relatively high numbers of exposures, but as the number of exposures decreases the percentage differences increase.

Table 6-5. Sensitivity to activity database: 2002 simulated counts of Los Angeles CSA general population and children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

Population Group	One or more exposures			Three or more exposures		
	All CHAD (6 pools)	**All CHAD (5 pools)**	**CA only (5 pools)**	**All CHAD (6 pools)**	**All CHAD (5 pools)**	**CA only (5 pools)**
Base Case						
General Population	3,106,438	2,660,322	2,679,695 (+1%)	744,072	539,704	617,195 (+14%)
Children (5-18)	1,285,413	1,114,334	1,196,736 (+7%)	375,719	274,764	334,518 (+22%)
Current Standard						
General Population	93,589	70,942	45,567 (-36%)	1,910	1,364	1,091 (-20%)
Children (5-18)	30,560	23,738	13,097 (-45%)	273	273	273 (+0%)

Table 6-6. Sensitivity to activity database: 2002 simulated counts of Sacramento CSA general population and children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

Population Group	One or more exposures			Three or more exposures		
	All CHAD (6 pools)	**All CHAD (5 pools)**	**CA only (5 pools)**	**All CHAD (6 pools)**	**All CHAD (5 pools)**	**CA only (5 pools)**
Base Case						
General Population	389,381	353,357	341,199 (-3%)	54,905	43,165	48,632 (+13%)
Children (5-18)	150,240	138,307	137,181 (-1%)	24,799	20,360	24,541 (+21%)
Current Standard						
General Population	82,984	73,528	59,954 (-18%)	1,866	1,158	1,222 (+5%)
Children (5-18)	27,886	24,927	20,907 (-16%)	450	257	579 (+125%)

6.2 Ozone Decay Rate

To test the sensitivity of the APEX predictions to the ozone decay rate distribution, we compared the base case results with corresponding results with the decay rate set uniformly to its 10th percentile value and its 90th percentile value. The results are presented in Figures 6-3 and 6-4 for active persons with moderate activity in Houston and Boston, respectively, for exceedances of an 8-hour average 0.05 ppm threshold. The figures show only small changes in the rate of exceedances for these rather extreme decay rate scenarios.

Tables 6-7 and 6-8 present the results for the number of persons exposed to 8-hour average concentrations exceeding 0.07 ppm in the general population and for children, respectively, with moderate exertion. These results do show differences among the decay rate scenarios for these exposures to elevated concentrations.

Table 6-7. Sensitivity to ozone decay rate: 2002 counts of general population with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

	One or more exposures			Three or more exposures		
Urban Area (CSA)	rate=90th percentile	Base case	Rate=10th percentile	rate=90th percentile	Base case	Rate=10th percentile
Boston	919,069 (-22%)	1,183,742	1,320,697 (+12%)	67,525 (-56%)	154,575	209,624 (+36%)
Houston	620,128 (-10%)	691,705	726,291 (+5%)	19,820 (-29%)	27,765	30,332 (+9%)

Table 6-8. Sensitivity to ozone decay rate: 2002 counts of children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

	One or more exposures			Three or more exposures		
Urban Area (CSA)	rate=90th percentile	Base case	Rate=10th percentile	rate=90th percentile	Base case	Rate=10th percentile
Boston	353,723 (-21%)	444,963	485,916 (+9%)	31,715 (-57%)	73,811	100,098 (+36%)
Houston	253,090 (-11%)	284,225	299,632 (+5%)	8,426 (-38%)	13,561	14,604 (+8%)

6.3 Proximity Factor

To test the sensitivity of the APEX predictions to the proximity factor distribution, we compared the base case results with corresponding results with the proximity factor set uniformly to its 10th percentile value and its 90th percentile value. The results are presented in Figures 6-5 and 6-6 for active persons with moderate activity in Houston and Boston, respectively, for exceedances of an 8-hour average 0.05 ppm threshold. The figures shows only small changes in the rate of exceedances for these rather extreme proximity factor scenarios.

Tables 6-9 and 6-10 present the results for the number of persons exposed to 8-hour average concentrations exceeding 0.07 ppm in the general population and for children, respectively, with moderate exertion. Again, the results show larger differences among the proximity factor scenarios for these exposures to elevated concentrations.

Table 6-9. Sensitivity to proximity factor: 2002 counts of general population with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion

	One or more exposures			Three or more exposures		
Urban Area (CSA)	**Factor = 90th percentile**	**Base case**	**Factor = 10th percentile**	**Factor = 90th percentile**	**Base case**	**Factor = 10th percentile**
Boston	1,133,360 (-4%)	1,183,742	1,283,268 (+8%)	136,098 (-12%)	154,575	196,671 (+27%)
Houston	662,416 (-4%)	691,705	756,944 (+9%)	23,752 (-14%)	27,765	37,795 (+36%)

Table 6-10. Sensitivity to proximity factor: 2002 counts of children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion

	One or more exposures			Three or more exposures		
Urban Area (CSA)	**Factor = 90th percentile**	**Base case**	**Factor = 10th percentile**	**Factor = 90th percentile**	**Base case**	**Factor = 10th percentile**
Boston	421,724 (-5%)	444,963	496,773 (+12%)	61,525 (-17%)	73,811	100,764 (+37%)
Houston	271,547 (-4%)	284,225	323,705 (+14%)	10,833 (-20%)	13,561	19,580 (+44%)

6.4 Air Exchange Rates

To test the sensitivity of the APEX predictions to the air exchange rate distributions, we compared the base case results with corresponding results with the air exchange rates set uniformly to their 10th percentile values and their 90th percentile values. The results are presented in Figure 6-7 for active persons with moderate activity in Houston for exceedances of an 8-hour average 0.05 ppm threshold and Figure 6-8 for active persons with moderate activity in Boston for exceedances of an 8-hour average 0.05 ppm threshold. The figures show a moderate increase in the rate of exceedances in both CSAs for the extremely high air exchange rate scenario.

Tables 6-11 and 6-12 present the results for the number of persons exposed to 8-hour average concentrations exceeding 0.07 ppm in the general population and for children, respectively, with moderate exertion. The results show a large increase in the number of people exposed to elevated concentrations for the scenario of high air exchange rates, especially for the multiple exposure case.

In contrast with the sensitivity simulation for the ozone decay rate discussed above, this increase in the air exchange rates is sufficient to increase concentration exceedances of 0.05 ppm in the indoor microenvironments compared to the base case.

Table 6-11. Sensitivity to air exchange rate: 2002 counts of general population with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion

	One or more exposures			Three or more exposures		
Urban Area (CSA)	**Rate = 10th percentile**	**Base case**	**Rate = 90th percentile**	**Rate = 10th percentile**	**Base case**	**Rate = 90th percentile**
Boston	835,924 (-29%)	1,183,742	2,403,294 (+103%)	53,239 (-66%)	154,575	1,074,596 (+595%)
Houston	595,894 (-14%)	691,705	1,816,088 (+135%)	17,333 (-38%)	27,765	659,527 (+2275%)

Table 6-12. Sensitivity to air exchange rate: 2002 counts of children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion

	One or more exposures			Three or more exposures		
Urban Area (CSA)	**Rate = 10th percentile**	**Base case**	**Rate = 90th percentile**	**Rate = 10th percentile**	**Base case**	**Rate = 90th percentile**
Boston	320,293 (-28%)	444,963	756,875 (+70%)	24,001 (-67%)	73,811	420,296 (+469%)
Houston	243,060 (-14%)	284,225	652,386 (+130%)	7,142 (-47%)	13,561	262,720 (+1837%)

In addition to lack of precision in measurements due to within-residence variations in air exchange rates, some recent studies have suggested that the reported rates may be biased upward from actual rates by as much as a factor of two (Wallace, L., personal communication). Primary reasons for such potential bias include (a) the typical placement of tracer emitters in relatively isolated locations, such as bathrooms and bedrooms, so that the tracer gas is constrained from reaching the collector, and (b) the fact that for apartments a substantial portion of air is exchanged with indoor common areas, such as hallways, rather than to the outdoors.

In order to determine the sensitivity of APEX predictions to such bias, if present, a set of sensitivity simulations was performed for the New York CSA, setting the residential air exchange rates to half the value selected from the air exchange rate distribution. The New York CSA was selected because it is known to contain a high density of apartment buildings.

Table 6-13 presents the results for the number of persons exposed to 8-hour average concentrations exceeding 0.07 ppm in the general population and for children, respectively, with moderate exertion. The results suggest that the results are not sensitive to a potential overestimate of air exchange rates by a factor of two.

Table 6-13. Sensitivity to air exchange rate: 2002 simulated counts of New York CSA general population and children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

Population Group	One or more exposures		Three or more exposures	
	Base case	Residential AER=1/2	Base case	Residential AER=1/2
General Population	5,439,379	5,420,869 (-0.3%)	958,948	953,253 (-0.6%)
Children (5-18)	2,033,582	2,030,023 (-0.2%)	494,780	495,136 (+0.1%)

6.5 Long-term Activity Patterns

The version of APEX used for this analysis includes a new treatment of activity patterns to construct long-term individual activity patterns, as described in section 2.3.3 and Appendix C. To test the sensitivity of the APEX results to this new treatment we compared the base case exposure results with corresponding results where (a) the new treatment was not implemented, and (b) the D statistic was set to 0.75 instead of 0.2. The comparisons of the days/person with exceedances of 8-hour average concentrations for active persons on Boston during moderate exertion (Figure 6-9) shows little difference between APEX estimates for the base case and for the simulation that did not incorporate the new long-term activity pattern treatment. The comparison between the base case and the sensitivity simulation with the diversity statistic, D, set to 0.75 (Figure 6-10) shows a slight elevation of the exposure rate for the highest 5% to 10% or the exposed population compared to the base case for this rather extreme scenario.

Table 6-14 presents the results for the number of persons in Boston population groups with moderate exertion exposed to 8-hour average concentrations exceeding 0.07 ppm. Again, the results show very small differences among the scenarios for these exposures to elevated concentrations.

Table 6-14. Sensitivity to longitudinal activity pattern algorithm: 2002 counts of Boston population groups with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

Population group	One or more exposures			Three or more exposures		
	Base case	Simple re-sampling	Div=0.75	Base case	Simple re-sampling	Div=0.75
General population	1,183,742	1,212,409 (-2%)	1,101,740 (+6%)	154,575	140,765 (-9%)	169,718 (+10%)
Children (ages 5-18)	444,963	458,773 (-3%)	408,962 (+8%)	73,811	70,478 (-5%)	80,097 (+9%)

An alternative algorithm for constructing longitudinal diaries was recently developed for another population exposure model, the Hazardous Air Pollutant Exposure Model (HAPEM). This approach, described in Appendix H, is designed to better represent the variability among individuals by limiting the number of selected activity diaries used to represent an individual.

This approach was adapted to use in APEX as follows. For each simulated individual/diary pool combination, cluster analysis is used to group the corresponding diaries into three clusters of similar patterns. (There are 18 clusters for each selected individual: 3 temperature categories × 2 day-of-week types × 3 clusters = 18.) A single activity pattern is selected from each cluster to represent the set of behaviors of the simulated individual. Next, cluster-to-cluster transition probabilities are defined both within diary pools and across diary pools, and used in a Markov process to select a cluster for each day of the modeling period.

A set of sensitivity simulations was performed for the Atlanta CSA using (a) the alternative clustering algorithm, and (b) simple re-sampling. Table 6-15 presents the results for the number of persons in Atlanta population groups with moderate exertion exposed to 8-hour average concentrations exceeding 0.07 ppm. The results show that the predictions made with alternative algorithm ("cluster") are substantially different from those made with simple re-sampling or with the new APEX algorithm ("base case"). For the cluster algorithm approximately 20% of the individuals with 1 or more exposure have 3 or more exposures. The corresponding value is about 10% or less for the other algorithms.

Table 6-15. Sensitivity to longitudinal diary algorithm: 2002 simulated counts of Atlanta general population and children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

	One or more exposures			Three or more exposures		
Population Group	**Simple re-sampling**	**Base case**	**Cluster**	**Simple re-sampling**	**Base case**	**Cluster**
General Population	849,843	822,480 (-3%)	643,446 (-24%)	52,300	68,900 (+32%)	120,442 (+130%)
Children (5-18)	340,104	324,869 (-4%)	266,580 (-22%)	27,515	33,048 (+20%)	57,076 (+107%)

Table 6-16 presents the results for the mean and standard deviation of number of days/person with 8-hour average exposures exceeding 0.07 ppm with moderate or greater exertion. The results show that although the mean for the cluster algorithm is very similar to the other approaches, the standard deviation is substantially higher, i.e., the cluster algorithm results in substantially higher inter-individual variability. As described in Appendix H, limited evaluation of the cluster algorithm by comparison to measurement data shows reasonably good agreement.

Table 6-16. Sensitivity to longitudinal diary algorithm: 2002 days per person with 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion for Atlanta general population and children (ages 5-18).

Population Group	Mean Days/Person			Standard Deviation		
	Simple re-sampling	Base case	Cluster	Simple re-sampling	Base case	Cluster
General Population	0.250	0.253 (+1%)	0.256 (+2%)	0.592	0.628 (+6%)	0.843 (+42%)
Children (5-18)	0.511	0.509 (-0%)	0.538 (+5%)	0.803	0.843 (+5%)	1.173 (+46%)

6.6 Near roadway locations

Although APEX accounts for depressed ozone concentrations near roadways for activities that occur outdoors, including in vehicles, it does not account for depressed ozone concentrations in indoor microenvironments that are near major roadways. This issue was recently addressed in the HAPEM model by developing a national data base of estimates of the fraction of the population of each census tract that resides in the vicinity of a major roadway, either 0 – 75m or 75m-200m. Although the data specify only the fraction of the residential population of each tract that live in the vicinity of the roadway, it was assumed that the same proportion of non-residential indoor venues are in the vicinity of major roadways for each tract. The development of the data base is described in Appendix I.

In order to test the sensitivity of APEX to the potential overestimate of concentrations in indoor microenvironments near roadways, we used the roadway proximity data developed for HAPEM to estimate the probability of a randomly selected individual living in proximity to either an interstate or a major urban road, as described below. Based on these estimated probabilities, a simulated individual’s residence (and other frequented indoor microenvironments) were assigned roadway-category-specific proximity factor distributions as previously described for in-vehicle microenvironments.

6.6.1 Proximity Probabilities

Given the specifications of the near roadway proximity factors, described in Appendix A, for this application we defined near-roadway proximity as being located within 75 meters of a major roadway, i.e., the first category in the HAPEM database. The probabilities for Boston and Houston were estimated by averaging the HAPEM data across the tracts in the respective modeling domains, and are as follows.

Boston:	27%
Houston:	20%

Next, we allocated these fractions between the categories used to develop the near-roadway proximity factors (i.e., interstate and other major urban road) according to the proportions of roadway miles in the respective categories as specified by the Federal Highway Administration (FHWA 2004) for the urban areas. These proportions were as follows:

Boston:

Interstate	0.05
Other major urban:	0.95

Houston:

Interstate	0.03
Other major urban:	0.97

Combining these data resulted in the following probabilities for residing within 75 meters of an interstate or another major urban road.

Boston:

Interstate	1%
Other major urban:	26%
Not within 75 meters of a major roadway:	73%

Houston:

Interstate	1%
Other major urban:	19%
Not within 75 meters of a major roadway:	80%

Although the HAPEM data specify only the fraction of the residential population that live in the vicinity of the roadway, for this application the same probabilities were applied to non-residential indoor microenvironments as well.

6.6.2 Results

Table 6-17 presents the results for the number of children (5-18) exposed to 8-hour average concentrations exceeding 0.07 ppm with moderate or greater exertion. The results show little difference in the predictions compared to the base case.

Table 6-17. Sensitivity to near roadway proximity for indoor sources: 2002 counts of children (ages 5-18) with any or three or more 8-hour ozone exposures above 0.07 ppm concomitant with moderate or greater exertion.

	One or more exposures		**Three or more exposures**	
Urban Area (CSA)	**Base case**	**Near roadway**	**Base case**	**Near roadway**
Boston	444,963	432,772 (-3%)	73,811	64,859 (-12%)
Houston	284,225	281,256 (-1%)	13,561	12,358 (-9%)

Figure 6-1

Days/Person with Exceedances of 8-Hour Average Concentrations (ppm) During Moderate Exertion --All Persons, Boston, 2002--

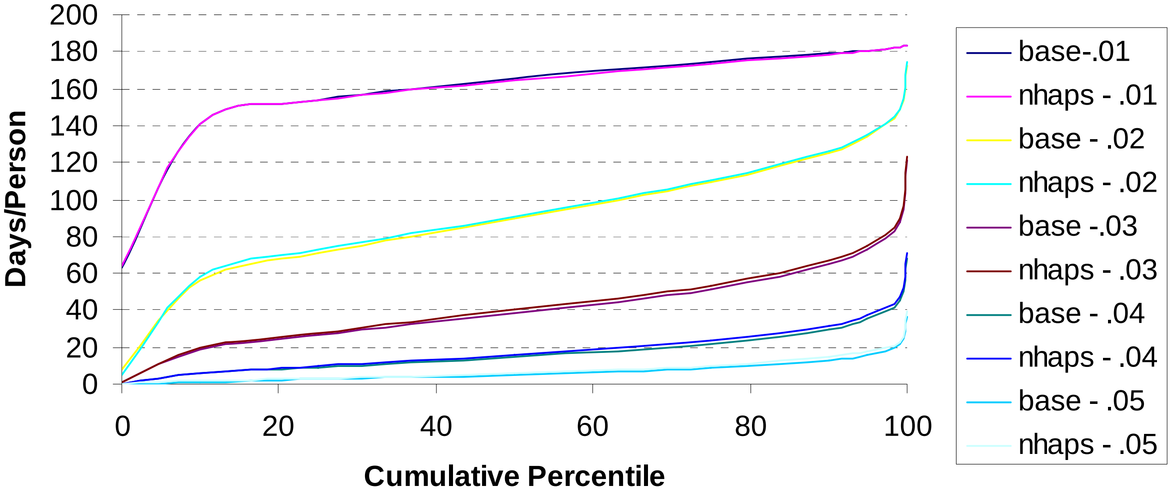


Figure 6-2

Days/Person with Exceedances of 8-Hour Average Concentrations (ppm) During Moderate Exertion --Active Persons, Boston, 2002--

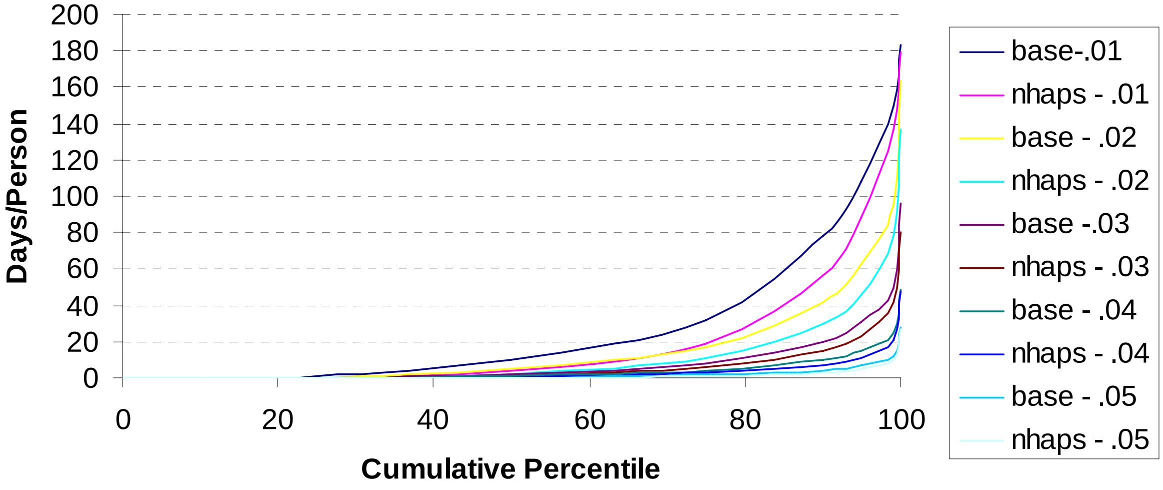


Figure 6-3

**Decay Rate Sensitivity:
Days/Person with Exceedances of 0.05 ppm
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Houston, 2002--**

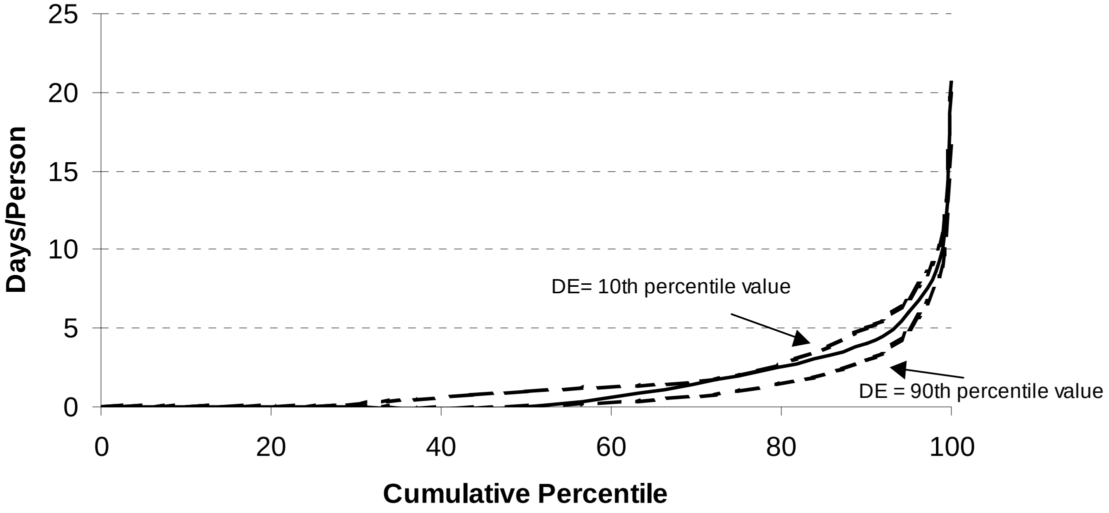


Figure 6-4

**Decay Rate Sensitivity:
Days/Person with Exceedances of 0.05 ppm
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Boston, 2002--**

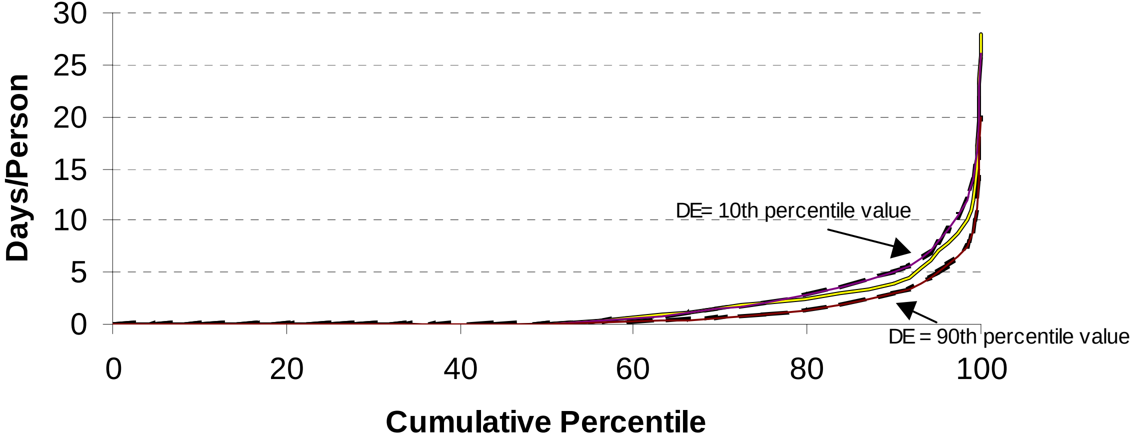


Figure 6-5

**Proximity Factor Sensitivity:
Days/Person with Exceedances of 0.05 ppm
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Houston, 2002--**

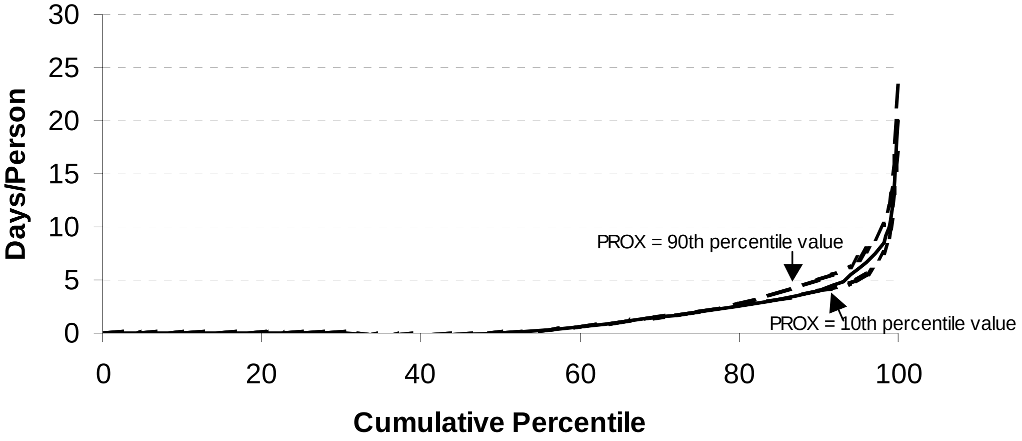


Figure 6-6

**Proximity Factor Sensitivity:
Days/Person with Exceedances of 0.05 ppm
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Boston, 2002--**

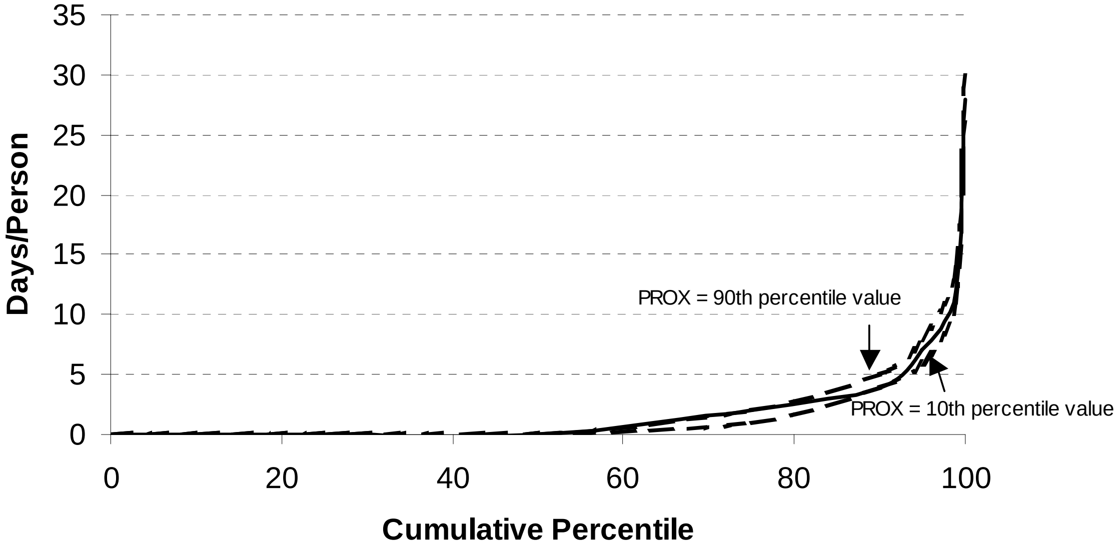


Figure 6-7

Air Exchange Rate Sensitivity:
Days/Person with Exceedances of 0.05 ppm
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Houston, 2002--

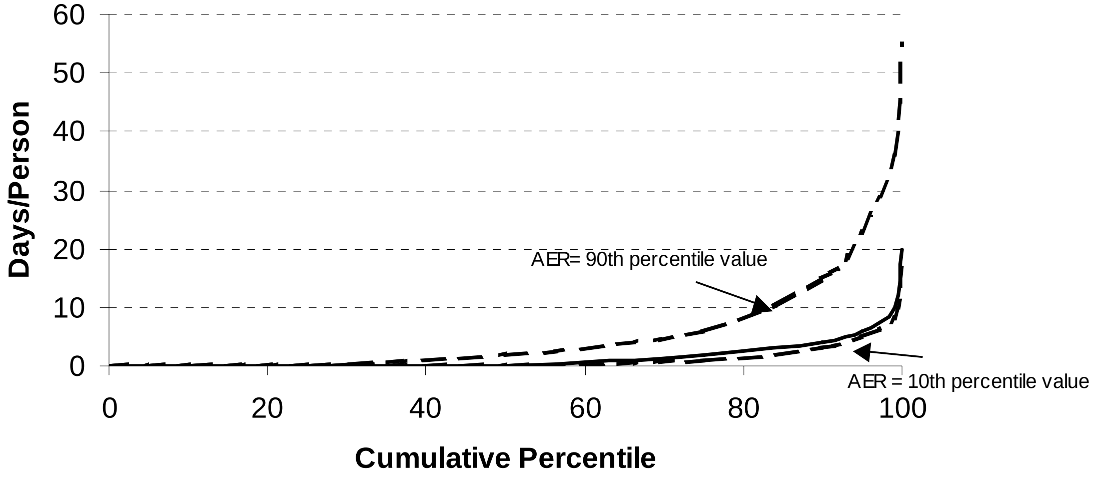


Figure 6-8

Air Exchange Rate Sensitivity:
Days/Person with Exceedances of 0.05 ppm
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Boston, 2002--

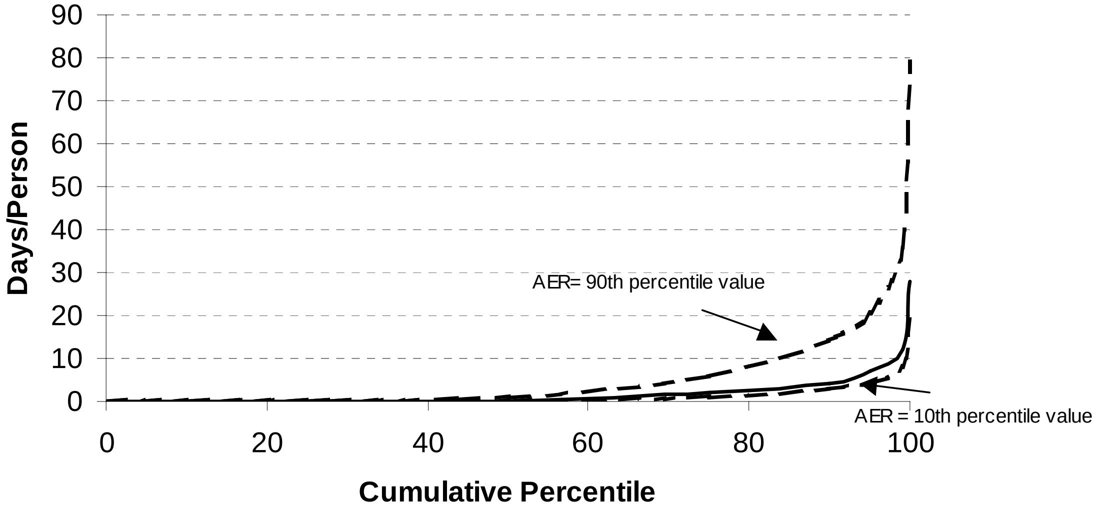


Figure 6-9

Long-term Activity Pattern Sensitivity:
Days/Person with Exceedances of
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Boston, 2002--

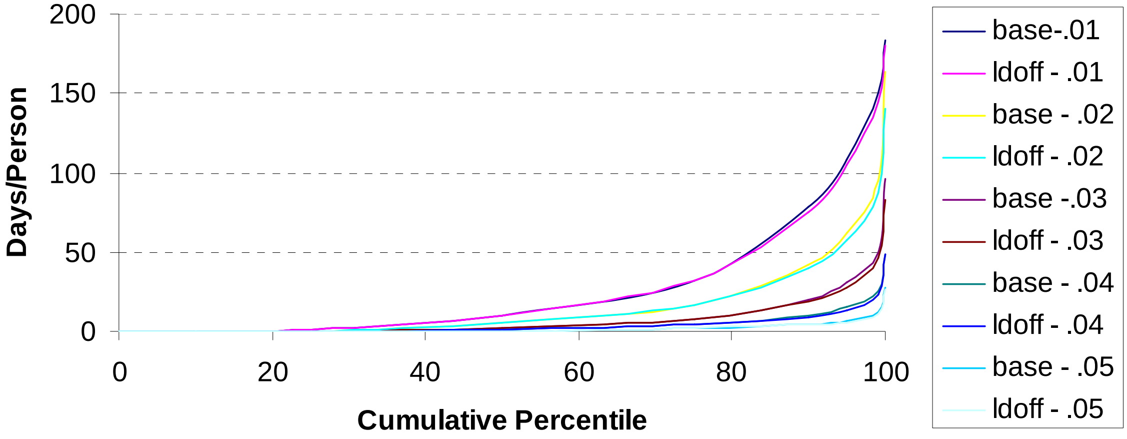
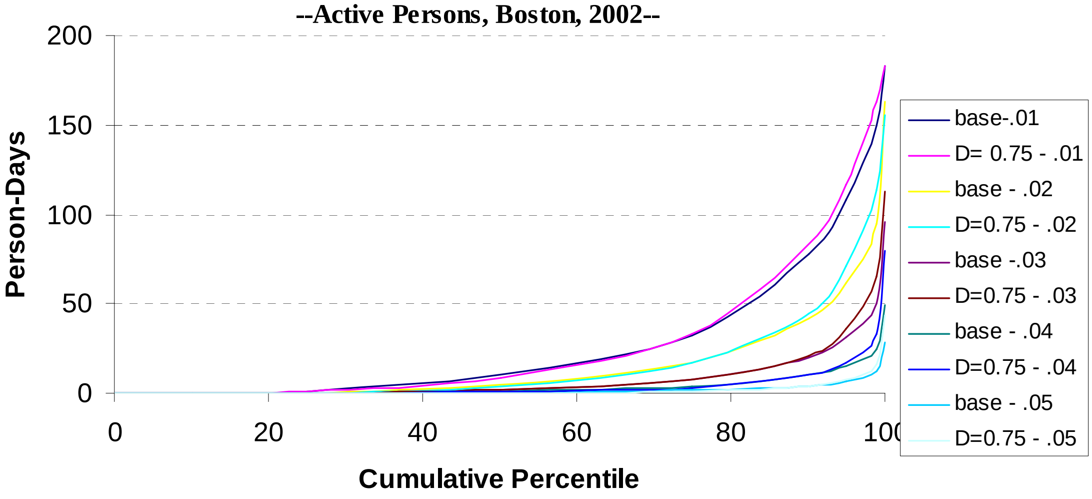


Figure 6-10

Diversity Statistic Sensitivity:
Days/Person with Exceedances of
8-Hour Average Exposure Concentration During Moderate Exertion
--Active Persons, Boston, 2002--



7. MODEL EVALUATION

In order to evaluate the performance of APEX we compared APEX simulation results to personal ozone concentration measurements taken from the Harvard Southern California Chronic Ozone Exposure Study (Xue et al. 2005, Geyh et al. 2000). In this study, 224 children in ages between 7 and 12 yr were followed for 1 year from June 1995 to May 1996. Passive ozone samplers were used to measure ozone personal ozone concentrations, as well as indoor and outdoor concentrations at participants homes for 6 consecutive days each month. The subjects resided in two separate areas of San Bernardino County: urban Upland CA, and the small mountain towns of Lake Arrowhead, Crestline, and Running Springs, CA.

From the original dataset (Geyh et al. 2000), Xue et al. (2005) identified 160 subjects on which longitudinal ozone concentrations have been made at least in 6 of the 12 months of study period. This dataset was used for the APEX model evaluation. The number of 6-day average measured personal exposure concentrations in each data set varies from 7 to 31. In the Upland area, where more than 95% of the subjects had air-conditioning in their homes, the maximum 6-day average exposure concentration was 0.029 ppm, and all other 6-day averages were less than 0.025 ppm. In the Lake Arrowhead area, where none of the subjects' homes were air-conditioned, the maximum exposure concentration was 0.034 ppm with five 6-day periods that had at least one subject with an exposure concentration greater than 0.025 ppm.

For the APEX simulations we used hourly outdoor concentrations from fixed site monitors located in Upland and Crestline as inputs. The air exchange rates used were those developed for Sacramento from measurements taken in the inland portions of the Los Angeles area: Sacramento, Riverside, and San Bernardino Counties. For each 6-day period for which personal measurements were available we simulated 10,000 subjects in the 7 – 12 age range in each of the two study areas. For each case the distribution of simulated 6-day average exposure concentrations was compared to the corresponding distribution of measured values.

Example comparisons for Upland are presented in Figures 7-1[1] and 7-3[2]; both show a systematic underprediction of the measured values by APEX. In Figure 7-1 the underestimate ranges from about 0.005 ppm at the lower end of the distribution to about 0.01 ppm at the upper end of the distribution. The discrepancies are similar in Figure 7-3, ranging from 0.002 to 0.01 ppm.

To provide some insight into the factors driving these discrepancies, comparisons were also made between the continuous measurements made inside the subjects' homes and the APEX indoor residential concentration estimates during the times of exposure. In addition, comparisons were made between the ozone concentrations measured outside the homes of the study subjects and those measured at the nearby fixed site monitors. These comparisons are presented in Figures 7-2 and 7-4, which show that the 6-day average fixed site concentrations input to APEX correspond to the lower end of the ranges measured outside the homes of the study subjects, and are more then 0.005 ppm lower than the mean for the study subjects. The residential indoor concentration estimated by APEX is about 0.002 ppm lower than the mean measured value for the subjects in Figure 7-2, and only slightly lower than the mean measured value in Figure 7-4.

[1] Number of measurements = 25

[2] Number of measurements = 15

This suggests that for the Upland area, the underestimate is primarily a result of the outdoor concentration discrepancy.

Figure 7-1

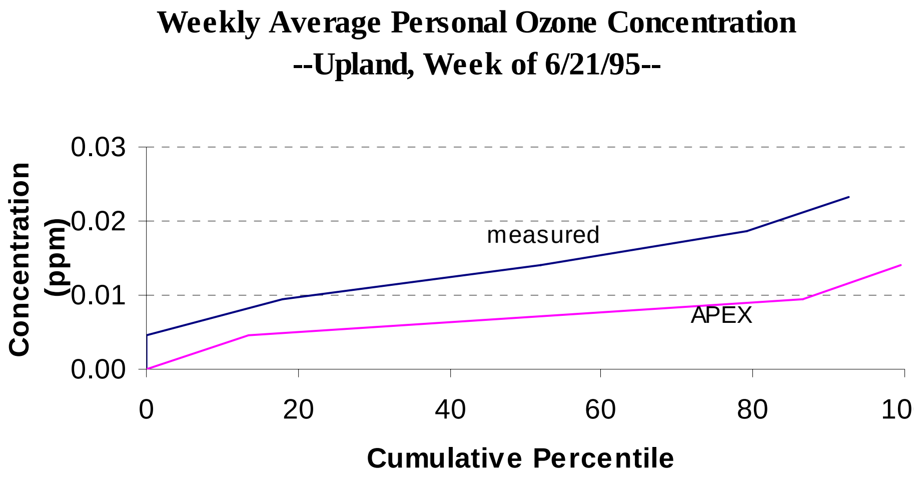


Figure 7-2

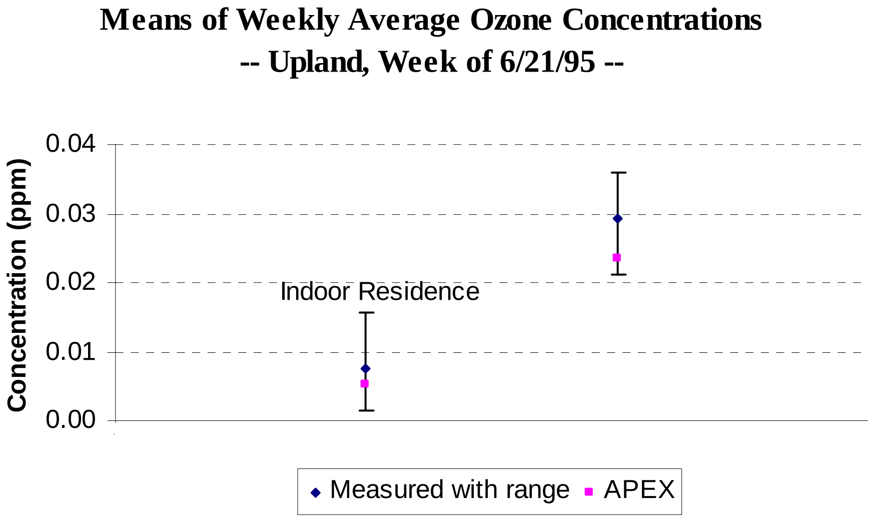


Figure 7-3

Weekly Average Personal Ozone Concentration
--Upland, Week of 5/8/96--

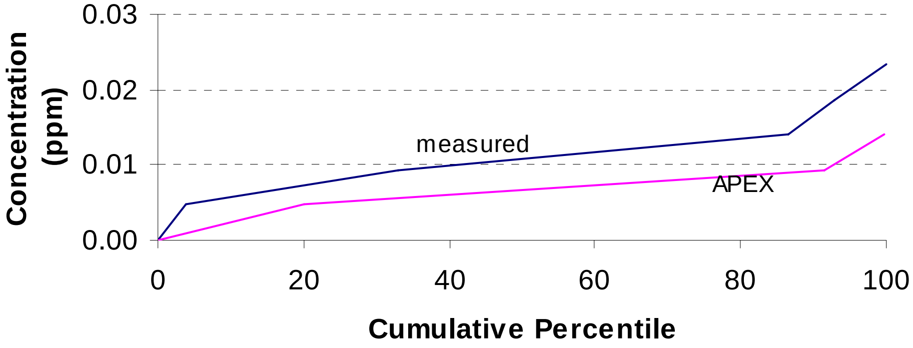
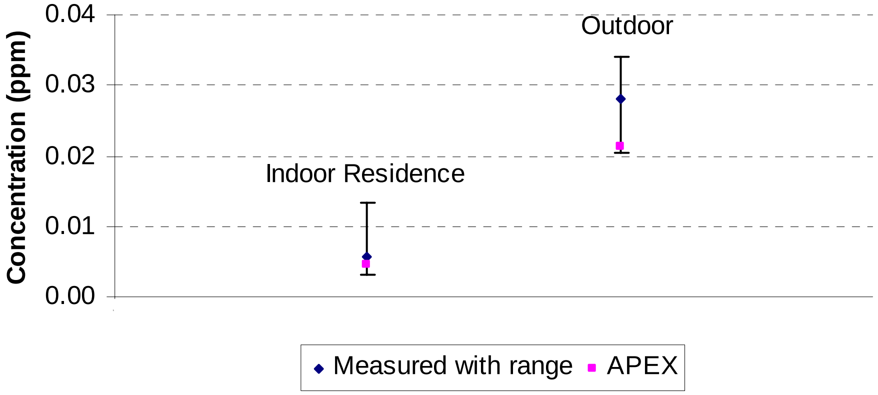


Figure 7-4

Means of Weekly Average Ozone Concentrations
-- Upland, Week of 5/8/96 --



Figures 7-5, 7-7, 7-9, and 7-11 present personal exposure comparisons for the Lake Arrowhead area., with the corresponding supplemental comparisons of indoor and outdoor concentrations presented in Figures 7-6, 7-8, 7-10, and 7-12.

Figure 7-6 shows good agreement between mean indoor and mean outdoor concentrations, Figure 7-5[3] shows an underestimate of personal exposure of about 0.007 ppm in the upper end of the distribution. A possible reason for this discrepancy is that the APEX outdoor concentration inputs for these evaluation simulations do not reflect the substantial variability observed throughout the study area (0.020 to 0.035 ppm), and thus do not account for the elevated concentrations driving the upper end of the distribution.

Figure 7-5

Weekly Average Personal Ozone Concentration
--Lake Arrowhead, Week of 7/12/95--

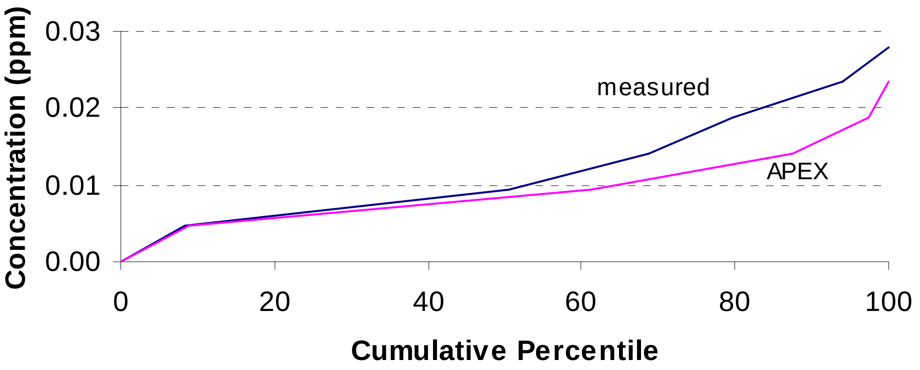
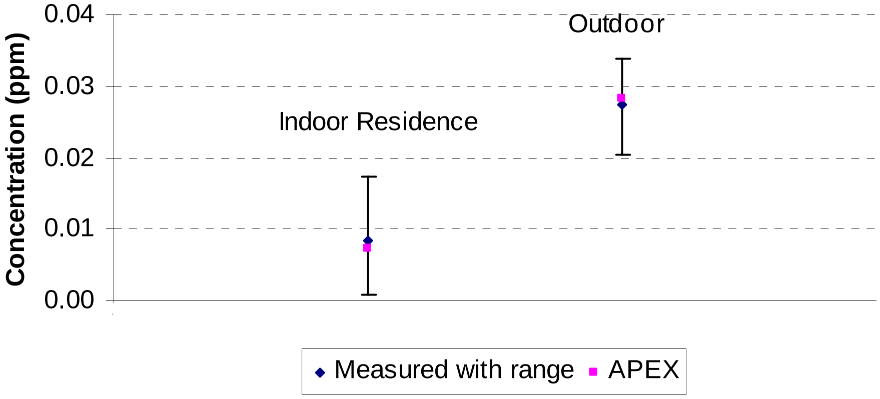


Figure 7-6

Means of Weekly Average Ozone Concentrations
-- Lake Arrowhead, Week of 7/12/95 --



[3] Number of measurements = 18

Figure 7-7[4] shows an underestimate in personal exposure ranging from about 0.005 to 0.01 ppm. Figure 7-8 shows good agreement in the mean outdoor concentrations, but a low APEX estimate for the mean indoor residential concentration. In conjunction with Figures 7-6 and 7-7, this suggests the lower end of the personal exposure distribution is primarily influenced by the indoor concentration estimates.

Figure 7-7

Weekly Average Personal Ozone Concentration
--Lake Arrowhead, Week of 6/21/95--

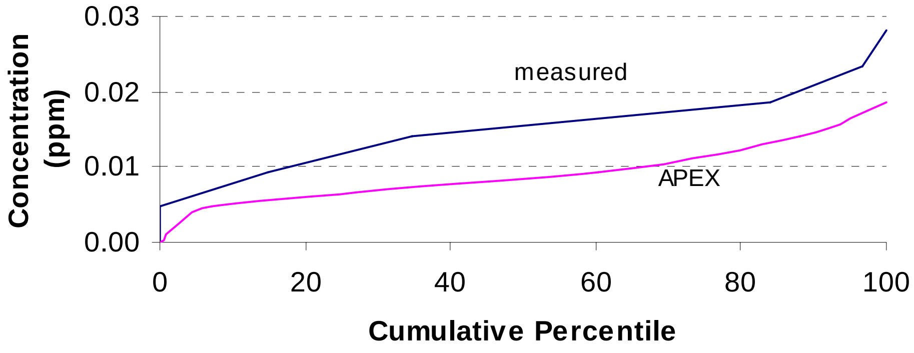
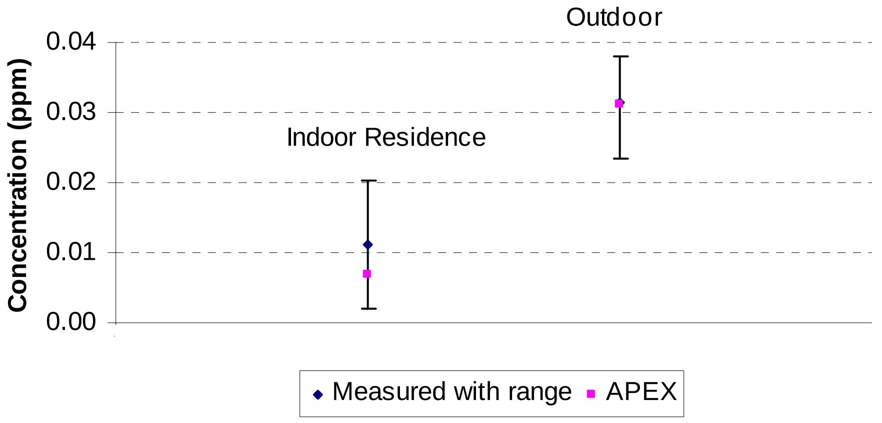


Figure 7-8

Means of Weekly Average Ozone Concentrations
-- Lake Arrowhead, Week of 6/21/95 --



[4] Number of measurements = 31

Figure 7-9[5] shows good agreement between the measured and simulated personal exposure distributions, with an overestimate of about 0.005 ppm at the very upper end. Figure 7-10 shows good agreement between the mean indoor concentrations, but that the fixed site concentration exceeds even the upper end of the range of local outdoor concentrations. These figures are consistent with the suggestions above that the lower end of the personal exposure distribution is primarily influenced by the indoor concentration estimates and the upper end is primarily influenced by the outdoor concentration.

Figure 7-9

Weekly Average Personal Ozone Concentration
--Lake Arrowhead, Week of 6/28/95--

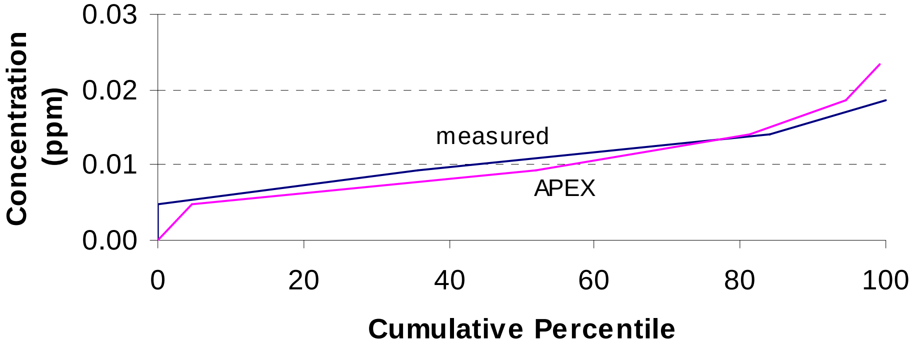
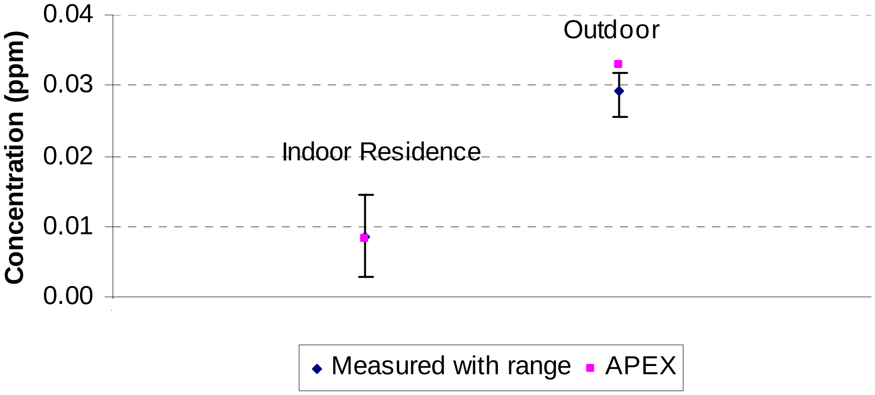


Figure 7-10

Means of Weekly Average Ozone Concentrations
-- Lake Arrowhead, Week of 6/28/95 --



[5] Number of measurements = 15

Figures 7-11[6] and 7-12 illustrate a case where the mean fixed site outdoor concentration corresponds to the upper end of the local outdoor concentration range, and the APEX indoor concentration estimate corresponds to the lower end of the range of measurements. The APEX personal exposure distribution underestimates the measured values by about 0.005 ppm, except the very top end of the distribution. This finding is also consistent with the suggestions that the lower end of the personal exposure distribution is primarily influenced by the indoor residential concentration and the upper end is primarily influenced by the outdoor concentration.

Figure 7-11

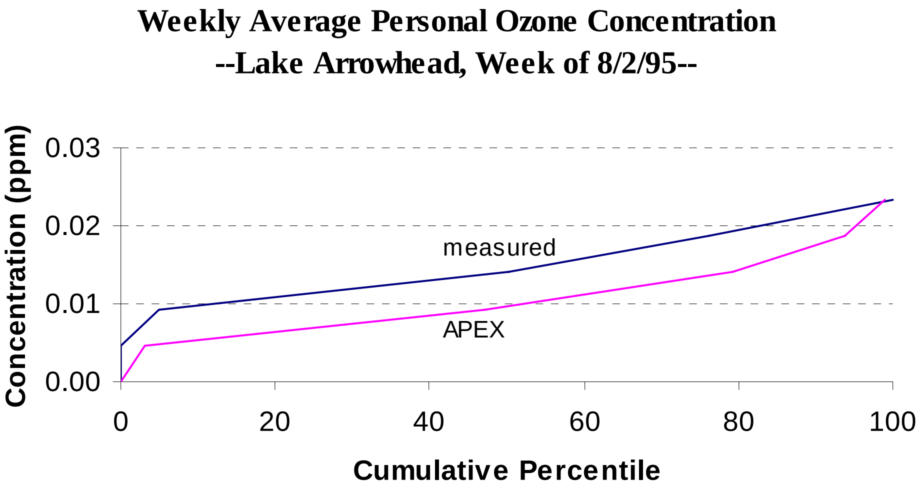
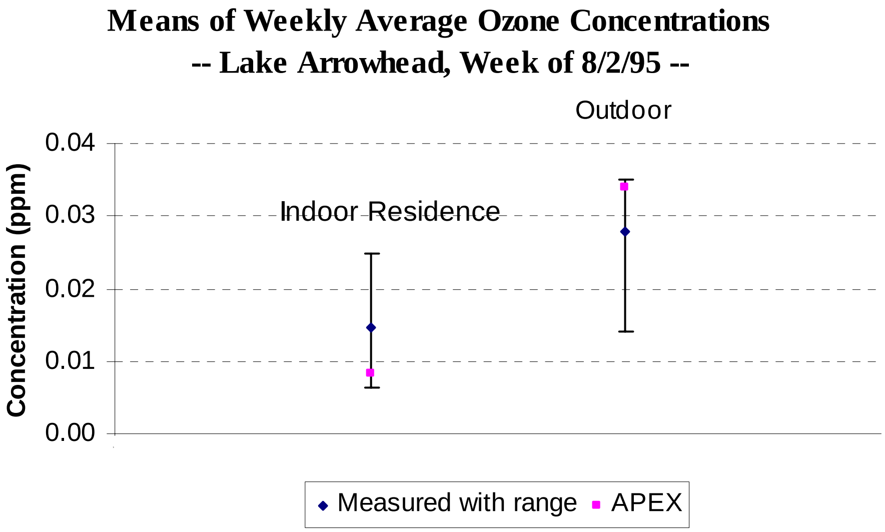


Figure 7-12



[6] Number of measurements = 18

Figures 7-13, 7-14, and 7-15 show the full sets of Upland personal exposure, indoor, and outdoor 6-day average concentration comparisons, respectively. Figure 7-13 shows that the mean personal exposure concentration for each 6-day period generally falls within in the range of measurements. However, there appears to be a tendency for the concentrations at the upper end of the distribution to be underestimated.

Figure 7-15 shows that for time periods with the highest local outdoor concentrations, the fixed site monitor corresponds to the lower end of the range. This suggests that the fixed site Upland monitor is not representative of the location of the ozone peak concentration during times of elevated concentrations, which may explain in part the tendency for underprediction of the higher personal exposure concentrations.

Figure 7-14 shows that the range of mean APEX residential indoor concentrations is smaller than the range of measurement means, with overestimates at the lower end of the temporal distribution and underestimates at the upper end. The underestimate at the upper end of the temporal distribution may be a result of one or more of the following influences.

- The underestimate of the mean APEX outdoor concentration inputs
- The possibility is that the mean indoor residential concentration during times of exposure is lower than the overall mean indoor concentration, i.e., the time spent indoors at home does not tend to occur during the peak diurnal concentrations.
- The possibility that the air exchange rate is underestimated for high temperature ranges for air-conditioned residences in inland California. (For this case the air exchange rate distribution for temperatures > 25° C was estimated from 83 measurements.)

The overestimate at the lower end of the distribution may be a result of overestimating the air exchange rates at low temperature ranges for air-conditioned residences in inland California. (For this case a single air exchange rate distribution was estimated for all temperatures < = 25° C.)

Figure 7-13

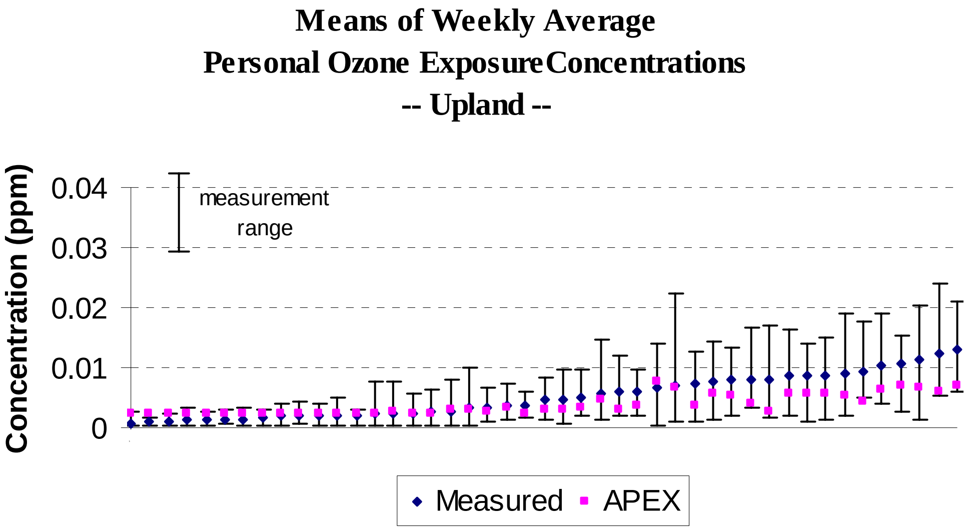


Figure 7-14

Means of Weekly Average Residential Indoor Ozone Concentrations -- Upland --

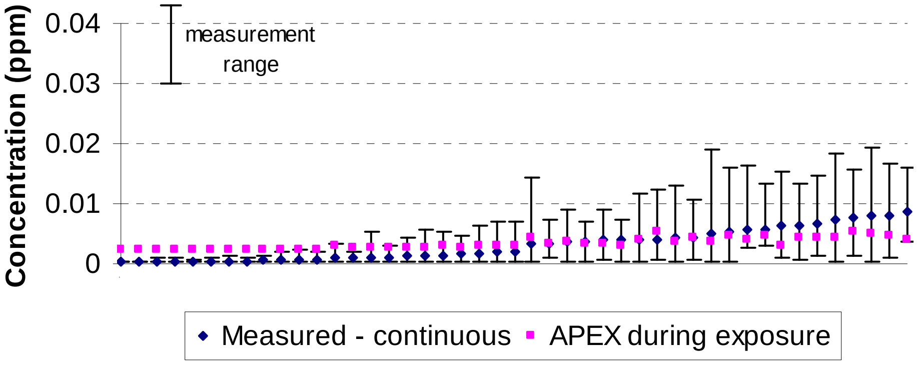
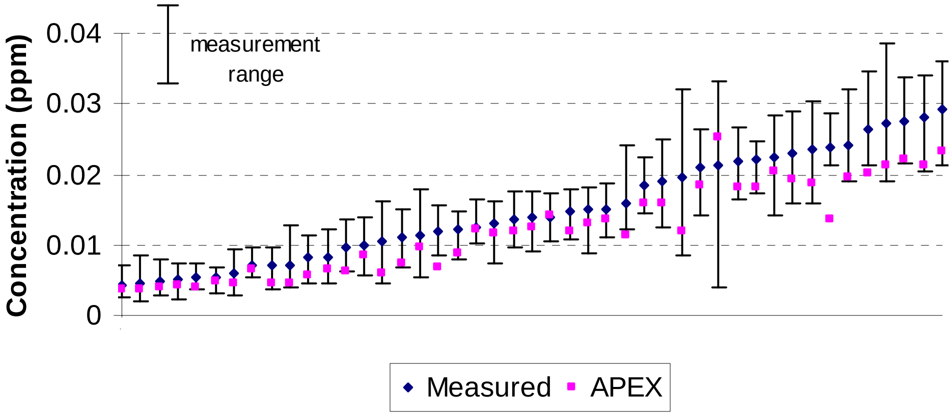


Figure 7-15

Means of Weekly Average Outdoor Ozone Concentrations -- Upland --



Figures 7-16, 7-17, and 7-18 show the Lake Arrowhead personal exposure, indoor, and outdoor 6-day average concentration comparisons, respectively. Figure 7-16 shows that the mean personal exposure concentration for each 6-day period generally falls within in the range of measurements. There appears to be a tendency for the concentrations at the lower end of the distribution to be overestimated and the ones at the upper end of the distribution to be underestimated.

In this case Figure 7-18 does not appear to show a systematic bias in the fixed site monitor concentration compared with the mean local outdoor concentration. But similar to Figure 7-14, Figure 7-17 shows that the range of mean APEX residential indoor concentrations is smaller than the range of measurement means, with overestimates at the lower end of the temporal distribution and underestimates at the upper end. In this case, the underestimate of the mean residential indoor concentration at the upper end of the distribution may result from one or more of the following influences.

- The possibility is that the mean indoor residential concentration during times of exposure is lower than the overall mean indoor concentration, i.e., the time spent indoors at home does not tend to occur during the peak diurnal concentrations.
- The possibility that the residential air exchange rate is underestimated for high temperature ranges for non-air-conditioned residences in inland California. (For this case the air exchange rate for temperatures > 25° C is based on only 14 measurements.)

The overestimate at the lower end of the distribution may be a result of overestimating the air exchange rates at low temperature ranges for non-air-conditioned residences in inland California. (For this case the air exchange rate for temperatures <=10° C is based on only 17 measurements.)

Figure 7-16

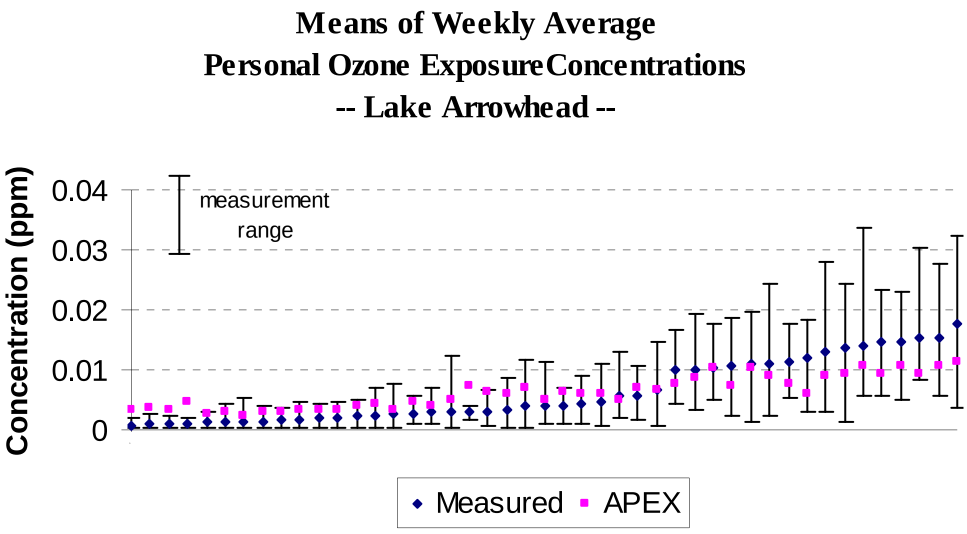
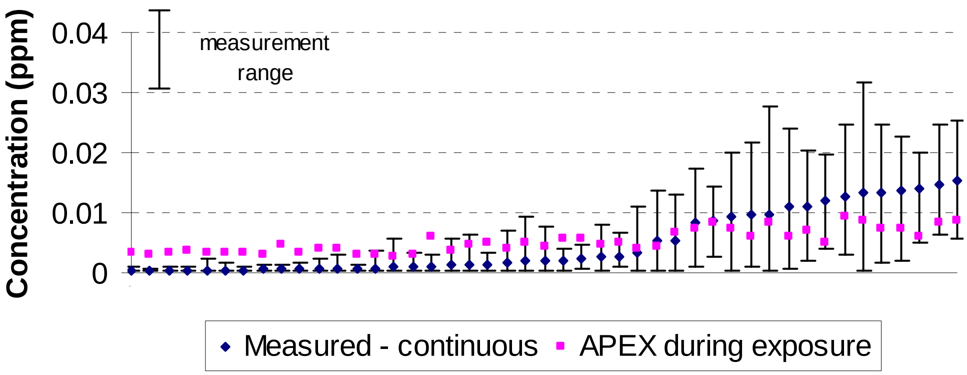
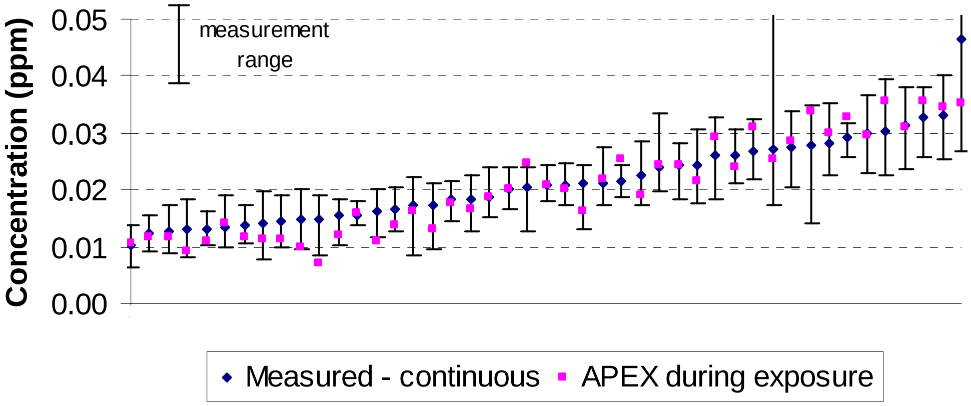


Figure 7-17



Means of Weekly Average Residential Indoor Ozone Concentrations -- Lake Arrowhead --

Figure 7-18



Means of Weekly Average Outdoor Ozone Concentrations -- Lake Arrowhead --

This page intentionally left blank.

8. REFERENCES

Biller, W.F., T.B. Feagans, T.R. Johnson, G.M. Duggan, R.A. Paul, T. McCurdy, and H.C. Thomas. 1981. A general model for estimating exposure associated with alternative NAAQS. Paper No. 81-18.4 in Proceedings of the 74th Annual Meeting of the Air Pollution Control Association, Philadelphia, Pa.

Burmaster, D.E. 1998. Lognormal distributions for skin area as a function of body weight. *Risk Analysis*, 18(1):27-32.

CASTNet 2004. Clean Air Status and Trends Network (CASTNet) 2003 Annual Report, Prepared by: MACTEC Engineering and Consulting, Inc. Prepared for:
U.S. Environmental Protection Agency, Office of Air and Radiation, Clean Air Markets Division, Washington, DC. Available at http://www.epa.gov/CASTNet .

Federal Highway Administration, U.S. Department of Transportation. 2004 (Publication Date). Highway Statistics 2003, Urbanized Areas, Miles and Daily Vehicle Miles of Travel. Table HM-71. Website: http://www.fhwa.dot.gov/policy/ohim/hs03/htm/hm71.htm.

Geyh, AS, Xue, J, Ozkaynak, H, and Spengler, JD. 2000. The Harvard Southern California chronic ozone exposure study: Assessing ozone exposure of grade-school-age children in two Southern California communities. *Environ Health Persp.* 108:265-270.

Graham S.E. and T. McCurdy. 2005. Revised ventilation rate (V_E) equations for use in inhalation-oriented exposure models. EPA/600/X-05/008.

Johnson, T.R., and R.A. Paul. 1983. The NAAQS Exposure Model (NEM) Applied to Carbon Monoxide. EPA-450/5-83-003. Prepared for the U.S. Environmental Agency by PEDCo Environmental Inc., Durham, N.C. under Contract No. 68-02-3390. U.S. Environmental Protection Agency, Research Triangle Park, N.C.

Johnson, T., J. Capel, E. Olaguer, and L. Wijnberg. 1992. Estimation of Ozone Exposures Experienced by Residents of ROMNET Domain Using a Probabilistic Version of NEM. Report prepared by IT Air Quality Services for the Office of Air Quality Planning and Standards, U. S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Johnson, T., A. Pakrasi, A. Wisbeth, G. Meiners, W. M. Ollison. 1995. Ozone exposures within motor vehicles – results of a field study in Cincinnati, Ohio. *Proceedings 88th annual meeting and exposition of the Air & Waste Management Association, June 18-23, 1995.* San Antonio, TX. Preprint paper 95-WA84A.02.

Johnson, T., J. Capel, and M. McCoy. 1996a. Estimation of Ozone Exposures Experienced by Urban Residents Using a Probabilistic Version of NEM and 1990 Population Data. Report prepared by IT Air Quality Services for the Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, September.

Johnson, T., J. Capel, J. Mozier, and M. McCoy. 1996b. Estimation of Ozone Exposures Experienced by Outdoor Children in Nine Urban Areas Using a Probabilistic Version of NEM. Report prepared for the Air Quality Management Division under Contract No. 68-DO-30094, April.

Johnson, T., J. Capel, M. McCoy, and J. Mozier. 1996c. Estimation of Ozone Exposures Experienced by Outdoor Workers in Nine Urban Areas Using a Probabilistic Version of NEM. Report prepared for the Air Quality Management Division under Contract No. 68-DO-30094, April.

Johnson, T. 2002. A Guide to Selected Algorithms, Distributions, and Databases Used in Exposure Models Developed By the Office of Air Quality Planning and Standards. Revised Draft. Prepared for U.S. Environmental Protection Agency under EPA Grant No. CR827033.

Langstaff, J. E. 2007. OAQPS Staff Memorandum to Ozone NAAQS Review Docket (OAR-2005-0172). Subject: Analysis Of Uncertainty In Ozone Population Exposure Modeling. [January 31, 2007].

Lee, Vallarino, Dumyahn, Ozkaynak, and Spengler. 1999. Ozone decay rates in residences. JAWMA. 49:1238-1244.

McCurdy, T. 2000. Conceptual basis for multi-route intake dose modeling using an energy expenditure approach. *Journal of Exposure Analysis and Environmental Epidemiology*. 10:1-12.

McCurdy, T., G. Glen, L. Smith, and Y. Lakkadi. 2000. The National Exposure Research Laboratory's Consolidated Human Activity Database, *Journal of Exposure Analysis and Environmental Epidemiology* 10: 566-578.

NCDC Surface Weather Observations. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite Data and Information Service, National Climatic Data Center, Asheville, North Carolina. http://lwf.ncdc.noaa.gov/oa/ncdc.html

Occupational Health and Safety Administration (OSHA). 2005. http://www.osha.gov/dts/chemicalsampling/data/CH_259300.html. Last accessed: July 19, 2005.

Persily, A., J. Gorfain, G. Brunner. 2005. Ventilation design and performance in U.S. office buildings. *ASHRAE Journal.* April 2005, 30-35.

Roddin, M.F., H.T. Ellis, and W.M. Siddiqee. 1979. Background Data for Human Activity Patterns, Vols. 1, 2. Draft Final Report prepared for Strategies and Air Standards Division, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, N.C.

U.S. Census Bureau, Employment Status: 2000- Supplemental Tables. Housing and Household Economic Statistics Division. http://www.census.gov/population/www/cen2000/phc-t28.html.

U.S. Environmental Protection Agency. 1999. Total Risk Integrated Methodology. Website: http://www.epa.gov/ttnatw01/urban/trim/trimpg.html.

U.S. Environmental Protection Agency. 2002. Consolidated Human Activities Database (CHAD) Users Guide. The database and documentation are available electronically on the internet at: http://www.epa.gov/chadnet1/.

U.S. Environmental Protection Agency. 2006a. Air Quality Criteria for Ozone and Related Photochemical Oxidants (Final). National Center for Environmental Assessment, U.S.

Environmental Protection Agency, Research Triangle Park, NC, EPA/600/R-05/004aF-cF. Available electronically on the internet at: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=149923

U.S. Environmental Protection Agency. 2006c. Total Risk Integrated Methodology (TRIM) - Air Pollutants Exposure Model Documentation (TRIM.Expo / APEX, Version 4) Volume I: User's Guide. Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC. June 2006. Available at: http://www.epa.gov/ttn/fera/human_apex.html.

U.S. Environmental Protection Agency. 2006d. Total Risk Integrated Methodology (TRIM) - Air Pollutants Exposure Model Documentation (TRIM.Expo / APEX, Version 4) Volume II: Technical Support Document. Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC. June 2006. Available at: http://www.epa.gov/ttn/fera/human_apex.html.

U.S. Environmental Protection Agency. 2007. Review of National Ambient Air Quality Standards for Ozone: Assessment of Scientific and Technical Information - OAQPS Staff Paper. Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC. Available electronically on the internet at: http://www.epa.gov/ttn/naaqs/standards/ozone/s_o3_cr_sp.html.

Xue J, Liu SV, Ozkaynak H, Spengler J. 2005. Parameter evaluation and model validation of ozone exposure assessment using Harvard Southern California Chronic Ozone Exposure Study Data. *J. Air & Waste Manage. Assoc.* **55**:1508–1515.

Xue J, McCurdy T, Spengler J, Ozkaynak H. 2004. Understanding variability in time spent in selected locations for 7-12-year old children. *J Expo Anal Environ Epidemiol* 14(3):222-33.

This page intentionally left blank.

APPENDIX A. ANALYSIS OF AIR EXCHANGE RATE DATA

This page intentionally left blank.

ICF INTERNATIONAL

DRAFT MEMORANDUM

To: **John Langstaff**

From: **Jonathan Cohen, Hemant Mallya, Arlene Rosenbaum**

Date: **September 30, 2005**

Re: **EPA 68D01052, Work Assignment 3-08. Analysis of Air Exchange Rate Data**

EPA is planning to use the APEX exposure model to estimate ozone exposure in 12 cities / metropolitan areas: Atlanta, GA; Boston, MA; Chicago, IL; Cleveland, OH; Detroit, MI; Houston, TX; Los Angeles, CA; New York, NY; Philadelphia, PA; Sacramento, CA; St. Louis, MO-IL; Washington, DC. As part of this effort, ICF Consulting has developed distributions of residential and non-residential air exchange rates (AER) for use as APEX inputs for the cities to be modeled. This memorandum describes the analysis of the AER data and the proposed APEX input distributions. Also included in this memorandum are proposed APEX inputs for penetration and proximity factors for selected microenvironments.

Residential Air Exchange Rates

Studies. Residential air exchange rate (AER) data were obtained from the following seven studies:

> **Avol:** Avol et al, 1998. In this study, ozone concentrations and AERs were measured at 126 residences in the greater Los Angeles metropolitan area between February and December, 1994. Measurements were taken in four communities: Lancaster, Lake Gregory, Riverside, and San Dimas. Data included the daily average outdoor temperature, the presence or absence of an air conditioner (either central or room), and the presence or absence of a swamp (evaporative) cooler. Air exchange rates were computed based on the total house volume and based on the total house volume corrected for the furniture. These data analyses used the corrected AERs.
>
> **RTP Panel:** Williams et al, 2003a, 2003b. In this study particulate matter concentrations and daily average AERs were measured at 37 residences in central North Carolina during 2000 and 2001 (averaging about 23 AER measurements per residence). The residences belong to two specific cohorts: a mostly Caucasian, non-smoking group aged at least 50 years having cardiac defibrillators living in Chapel Hill; a group of non-smoking, African Americans aged at least 50 years with controlled hypertension living in a low-to-moderate SES neighborhood in Raleigh. Data included the daily average outdoor temperature, and the number of air conditioner units (either central or room). Every residence had at least one air conditioner unit.
>
> **RIOPA:** Meng et al, 2004, Weisel et al, 2004. The Relationship of Indoor, Outdoor, and Personal Air (RIOPA) study was undertaken to estimate the impact of outdoor sources of air toxics to indoor concentrations and personal exposures. Volatile organic compounds,

carbonyls, fine particles and AERs were measured once or twice at 310 non-smoking residences from summer 1999 to spring 2001. Measurements were made at residences in Elizabeth, NJ, Houston TX, and Los Angeles CA. Residences in California were randomly selected. Residences in New Jersey and Texas were preferentially selected to be close (< 0.5 km) to sources of air toxics. The AER measurements (generally over 48 hours) used a PMCH tracer. Data included the daily average outdoor temperature, and the presence or absence of central air conditioning, room air conditioning, or a swamp (evaporative) cooler.

TEACH: Chillrud at al, 2004, Kinney et al, 2002, Sax et al, 2004. The Toxic Exposure Assessment, a Columbia/Harvard (TEACH) study was designed to characterize levels of and factors influencing exposures to air toxics among high school students living in inner-city neighborhoods of New York City and Los Angeles, CA. Volatile organic compounds, aldehydes, fine particles, selected trace elements, and AER were measured at 87 high school student's residences in New York City and Los Angeles in 1999 and 2000. Data included the presence or absence of an air conditioner (central or room) and hourly outdoor temperatures (which were converted to daily averages for these analyses).

Wilson 1984: Wilson et al, 1986, 1996. In this 1984 study, AER and other data were collected at about 600 southern California homes with three seven-day tests (in March and July 1984, and January, 1985) for each home. We obtained the data directly from Mr. Wilson. The available data consisted of the three seven-day averages, the month, the residence zip code, the presence or absence of a central air conditioner, and the presence or absence of a window air conditioner. We matched these data by month and zip code to the corresponding monthly average temperatures obtained from EPA's SCRAM website as well as from the archives in www.wunderground.com (personal and airport meteorological stations). Residences more than 25 miles away from the nearest available meteorological station were excluded from the analysis. For our analyses, the city/location was defined by the meteorological station, since grouping the data by zip code would not have produced sufficient data for most of the zip codes.

Wilson 1991: Wilson et al, 1996. Colome et al, 1993, 1994. In this 1991 study, AER and other data were collected at about 300 California homes with one two-day test in the winter for each home. We obtained the data directly from Mr. Wilson. The available data consisted of the two-day averages, the date, city name, the residence zip code, the presence or absence of a central air conditioner, the presence or absence of a swamp (evaporative) cooler, and the presence or absence of a window air conditioner . We matched these data by date, city, and zip code to the corresponding daily average temperatures obtained from EPA's SCRAM website as well as from the archives in www.wunderground.com (personal and airport meteorological stations). Residences more than 25 miles away from the nearest available meteorological station were excluded from the analysis. For our analyses, the city/location was defined by the meteorological station, since grouping the data by zip code would not have produced sufficient data for most of the zip codes.

Murray and Burmaster: Murray and Burmaster (1995). For this article, Murray and Burmaster corrected and compiled nationwide residential AER data from several studies conducted between 1982 and 1987. These data were originally compiled by the Lawrence Berkeley National Laboratory. We acknowledge Mr. Murray's assistance in obtaining these data for us. The available data consisted of AER measurements, dates, cities, and degree-days. Information on air conditioner presence or absence was not available.

Table A-1 summarizes these studies.

For each of the studies, air conditioner usage, window status (open or closed), and fan status (on or off) was not part of the experimental design, although some of these studies included information on whether air conditioners or fans were used (and for how long) and whether windows were closed during the AER measurements (and for how long).

As described above, in the following studies the homes were deliberately sampled from specific subsets of the population at a given location rather than the entire population: The RTP Panel study selected two specific cohorts of older subjects with specific diseases. The RIOPA study was biased towards residences near air toxics sources. The TEACH study focused on inner-city neighborhoods. Nevertheless, we included all these studies because we determined that any potential bias would be likely to be small and we preferred to keep as much data as possible.

Table A-1. Summary of Studies of Residential Air Exchange Rates

	Avol	RTP Panel	RIOPA	TEACH	Wilson 1984	Wilson 1991	Murray and Burmaster
Locations	Lancaster, Lake Gregory, Riverside, San Dimas. All in Southern CA	Research Triangle Park, NC	CA; NJ; TX	Los Angeles, CA; New York City, NY	Southern CA	Southern CA	AZ, CA, CO, CT, FL, ID, MD, MN, MT, NJ
Years	1994	2000; 2001	1999; 2000; 2001	1999; 2000	1984, 1985	1984	1982 – 1987
Months/Seasons	Feb; Mar; Apr; May; Jun; Jul; Aug; Sep; Oct; Nov	2000 (Jun; Jul; Aug; Sep; Oct; Nov), 2001 (Jan; Feb; Apr; May)	1999 (July to Dec); 2000 (all months); 2001 (Jan and Feb)	1999 (Feb; Mar; Apr; Jul; Aug); 2000 (Jan; Feb; Mar; Sep; Oct)	Mar 1984, Jul 1984, Jan 1985	Jan, Mar, Jul	Various
Number of Homes	86	37	284	85	581	288	1,884
Total AER Measurements	161	854	524	151	1,362	316	2,844
Average Number of Measurements per Home	1.87	23.08	1.85	1.78	2.34	1.10	1.51
Measurement Duration	Not Available	24 hour	24 to 96 hours	Sample time (hours) reported. Ranges from about 1 to 7 days.	7 days	7 days	Not available
Measurement Technique	Not Available	Perflourocarbon tracer.	PMCH tracer	Perflourocarbon tracer.	Perflourocarbon tracer.	Perflourocarbon tracer.	Not available
Min AER Value	0.01	0.02	0.08	0.12	0.03	0.01	0.01
Max AER Value	2.70	21.44	87.50	8.87	11.77	2.91	11.77
Mean AER Value	0.80	0.72	1.41	1.71	1.05	0.57	0.76
Min Temperature (C)	-0.04	-2.18	-6.82	-1.36	11.00	3.00	Not available

	Avol	RTP Panel	RIOPA	TEACH	Wilson 1984	Wilson 1991	Murray and Burmaster
Max Temperature (C)	36.25	30.81	32.50	32.00	28.00	25.00	Not available
Air Conditioner Categories	No A/C; Central or Room A/C; Swamp Cooler only; Swamp + [Central or Room]	Central or Room A/C (Y/N)	Window A/C (Y/N); Evap Coolers (Y/N)	Central or Room A/C (Y/N)	Central A/C (Y/N); Room A/C (Y/N);	Central A/C (Y/N); Room A/C (Y/N); Swamp Cooler(Y/N)	Not available
Air Conditioner Measurements	A/C use in minutes	Not Available	Duration measurements in Hrs and Mins	Not Available	Not Available	Not Available	Not available
Fan Categories	Not available	Fan (Y/N)	Fan (Y/N)	Not Available	Not Available	Not Available	Not available
Fan Measurements	Time on or off for various fan types during sampling was recorded, but not included in database provided.	Not Available	Duration measurements in Hrs and Mins	Not Available	Not Available	Not Available	Not available
Window Open/ Closed Data	Duration open between times 6am-12 pm; 12pm - 6 pm; and 6pm - 6am	Windows (open / closed along with duration open in inch-hours units	Windows (Open / Closed) along with window open duration measurements	Not Available	Not Available	Not Available	Not available
Comments			CA sample was a random sample of homes. NJ and TX homes were deliberately chosen to be near to ambient sources.	Restricted to inner-city homes with high school students.	Contemporaneous temperature data obtained for these analyses from SCRAM and www.wunderground.com meteorological data.	Contemporaneous temperature data obtained for these analyses from SCRAM and www.wunderground.com meteorological data.	

We compiled the data from these seven studies to create the following variables, of which some had missing values:

- Study
- Date
- Time – Time of the day that the AER measurement was made
- House_ID – Residence identifier
- Measurement_ID – Uniquely identifies each AER measurement for a given study
- AER – Air Exchange Rate (per hour)
- AER_Duration – Length of AER measurement period
- Have_AC – Indicates if the residence has any type of air conditioner (A/C), either a room A/C or central A/C or swamp cooler or any of them in combination. "Y" = "Yes." "N" = "No."
- Type_of_AC1 – Indicates the types of A/C or swamp cooler available in each house measured. Possible values: "Central A/C" "Central and Room A/C" "Central or Room A/C" "No A/C" "Swamp + (Central or Room)" "Swamp Cooler only" "Window A/C" "Window and Evap"
- Type_of_AC2 – Indicates if a house measured has either no A/C or some A/C. Possible values are "No A/C" and "Central or Room A/C."
- Have_Fan – Indicates if the house studied has any fans
- Mean_Temp – Daily average outside temperature
- Min_Temp – Minimum hourly outside temperature
- Max_Temp – Maximum hourly outside temperature
- State
- City
- Location – Two character abbreviation
- Flag – Data status. Murray and Burmaster study: "Used" or "Not Used." Other studies: "Used"; "Missing" (missing values for AER, Type_of_AC2, and/or Mean_Temp); "Outlier".

The main data analysis was based on the first six studies. The Murray and Burmaster data were excluded because of the absence of information on air conditioner presence. (However, a subset of these data was used for a supplementary analysis described below.) .

Based on our review of the AER data we excluded seven outlying high AER values – above 10 per hour. The main data analysis used all the remaining data that had non-missing values for AER, Type_of_AC2, and Mean_Temp. We decided to base the A/C type variable on the broad characterization "No A/C" versus "Central or Room A/C" since this variable could be calculated from all of the studies (excluding Murray and Burmaster). Information on the presence or absence of swamp coolers was not available from all the studies, and, also importantly, the corresponding information on swamp cooler prevalence for the subsequent ozone modeling cities was not available from the American Housing Survey. It is plausible that AER distributions depend upon the presence or absence of a swamp cooler. It is also plausible that AER distributions also depend upon whether the residence specifically has a central A/C, room or window A/C, or both. However we determined to use the broader A/C type definition, which in effect assumes that the exact A/C type and the presence of a swamp cooler are approximately proportionately represented in the surveyed residences.

Most of the studies had more than one AER measurement for the same house. It is reasonable to assume that the AER varies with the house as well as other factors such as the temperature. (The A/C type can be assumed to be the same for each measurement of the same house). We expected the temperature to be an important factor since the AER will be affected by the use of the available ventilation (air conditioners, windows, fans), which in turn will depend upon the outside meteorology. Therefore it is not appropriate to average data for the same house under different conditions, which might have been one way to account for dependence between multiple measurements on the same house. To simplify the data analysis, we chose to ignore possible dependence between measurements on the same house on different days and treat all the AER values as if they were statistically independent.

Summary Statistics. We computed summary statistics for AER and its natural logarithm LOG_AER on selected strata defined from the study, city, A/C type, and mean temperature. Cities were defined as in the original databases, except that for Los Angeles we combined all the data in the Los Angeles ozone modeling region, i.e. the counties of Los Angeles, Orange, Ventura, Riverside, and San Bernardino. A/C type was defined from the Type_of_AC2 variable, which we abbreviated as “NA” = “No A/C” and “AC” = “Central or Room A/C.” The mean temperature was grouped into the following temperature bins: -10 to 0 ºC, 0 to 10 ºC, 10 to 20 ºC, 20 to 25 ºC, 25 to 30 ºC, 30 to 40 ºC.(Values equal to the lower bounds are excluded from each interval.) Also included were strata defined by study = “All” and/or city = “All,” and/or A/C type = “All” and/or temperature bin = “All.” The following summary statistics for AER and LOG_AER were computed:

- Number of values
- Arithmetic Mean
- Arithmetic Standard Deviation
- Arithmetic Variance
- Deciles (Min, 10th, 20th … 90th percentiles, Max)

These calculations exclude all seven outliers and results are not used for strata with 10 or fewer values, since those summary statistics are extremely unreliable.

Examination of these summary tables clearly demonstrates that the AER distributions vary greatly across cities and A/C types and temperatures, so that the selected AER distributions for the modeled cities should also depend upon the city, A/C type and temperature. For example, the mean AER for residences with A/C ranges from 0.39 for Los Angeles between 30 and 40 ºC to 1.73 for New York between 20 and 25 ºC. The mean AER for residences without A/C ranges from 0.46 for San Francisco between 10 and 20 ºC to 2.29 for New York between 20 and 25 ºC. The need to account for the city as well as the A/C type and temperature is illustrated by the result that for residences with A/C and between 20 and 25 ºC, the mean AER ranges from 0.52 for Research Triangle Park to 1.73 for New York. Statistical comparisons are described below.

Statistical Comparisons. Various statistical comparisons were carried out between the different strata, for the AER and its logarithm. The various strata are defined as in the Summary Statistics section, excluding the "All" cases. For each analysis, we fixed one or two of the variables Study, City, A/C type, temperature, and tested for statistically significant differences among other variables. The comparisons are listed in Table A-2.

Table A-2. Summary of Comparisons of Means

Comparison Analysis Number.	Comparison Variable(s) "Groups Compared"	Stratification Variable(s) (not missing in worksheet)	Total Comparisons	Cases with significantly different means (5 % level)	
				AER	Log AER
1.	City	Type of A/C AND Temp. Range	12	8	8
2.	Temp. Range	Study AND City	12	5	5
3.	Type of A/C	Study AND City	15	5	5
4.	City	Type of A/C	2	2	2
5.	City	Temp. Range	6	5	6
6.	Type of A/C AND Temp. Range	Study AND City	17	6	6

For example, the first set of comparisons fix the Type of A/C and the temperature range; there are twelve such combinations. For each of these twelve combinations, we compare the AER distributions across different cities. This analysis determines whether the AER distribution is appropriately defined by the A/C type and temperature range, without specifying the city. Similarly, for the sixth set of comparisons, the study and city are held fixed (17 combinations) and in each case we compare AER distributions across groups defined by the combination of the A/C type and the temperature range.

The F Statistic comparisons compare the mean values between groups using a one way analysis of variance (ANOVA). This test assumes that the AER or log(AER) values are normally distributed with a mean that may vary with the comparison variable(s) and a constant variance. We calculated the F Statistic and its P-value. P-values above 0.05 indicate cases where all the group means are not statistically significantly different at the 5 percent level. Those results are summarized in the last two columns of the above table "Summary of Comparisons of Means" which gives the number of cases where the means are significantly different. Comparison analyses 2, 3, and 6 show that for a given study and city, slightly less than half of the comparisons show significant differences in the means across temperature ranges, A/C types, or both. Comparison analyses 1, 4, and 5 show that for the majority of cases, means vary significantly across cities, whether you first stratify by temperature range, A/C type, or both.

The Kruskal-Wallis Statistic comparisons are non-parametric tests that are extensions of the more familiar Wilcoxon tests to two or more groups. The analysis is valid if the AER minus the group median has the same distribution for each group, and tests whether the group medians are equal. (The test is also consistent under weaker assumptions against more general alternatives) The P-values show similar patterns to the parametric F test comparisons of the means. Since the logarithm is a strictly increasing function and the test is non-parametric, the Kruskal-Wallis tests give identical results for AER and Log (AER).

The Mood Statistic comparisons are non-parametric tests that compare the scale statistics for two or more groups. The scale statistic measures variation about the central value, which is a non-parametric generalization of the standard deviation. Specifically, suppose there is a total of N AER or log(AER) values, summing across all the groups. These N values are ranked from 1 to N, and the j'th highest value is given a score of $\{j - (N+1)/2\}^2$. The Mood statistic uses a one way ANOVA statistic to compare the total scores for each group. Generally, the Mood statistics show that in most cases the scale statistics are not statistically significantly different. Since the logarithm is a strictly increasing function and the test is non-parametric, the Mood tests give identical results for AER and Log (AER).

Fitting Distributions. Based on the summary statistics and the statistical comparisons, the need to fit different AER distributions to each combination of A/C type, city, and temperature is apparent. For each combination with a minimum of 11 AER values, we fitted and compared exponential, log-normal, normal, and Weibull distributions to the AER values.

The first analysis used the same stratifications as in the above "Summary Statistics" and "Statistical Comparisons" sections. Results are not reported for all strata because of the minimum data requirement of 11 values. Results for each combination of A/C type, city, and temperature (i.e., A, C, and T) were analyzed. Each combination has four rows, one for each fitted distribution. For each distribution we report the fitted parameters (mean, standard deviation, scale, shape) and the p-value for three standard goodness-of-fit tests: Kolmogorov-Smirnov (K-S), Cramer-Von-Mises (C-M), Anderson-Darling (A-D). Each goodness-of-fit test compares the empirical distribution of the AER values to the fitted distribution. The K-S and C-M tests are different tests examining the overall fit, while the Anderson-Darling test gives more weight to the fit in the tails of the distribution. For each combination, the best-fitting of the four distributions has the highest p-value and is marked by an x in the final three columns. The mean and standard deviation (Std_Dev) are the values for the fitted distribution. The scale and shape parameters are defined by:

- Exponential: density = $\sigma^{-1} \exp(-x/\sigma)$, where shape = mean = σ
- Log-normal: density = $\{\sigma x\sqrt{(2\pi)}\}^{-1} \exp\{ -(\log x - \zeta)^2 / (2\sigma^2)\}$, where shape = σ and scale = ζ. Thus the geometric mean and geometric standard deviation are given by $\exp(\zeta)$ and $\exp(\sigma)$, respectively.
- Normal: density = $\{\sigma\sqrt{(2\pi)}\}^{-1} \exp\{ -(x - \mu)^2 / (2\sigma^2)\}$, where mean = μ and standard deviation = σ
- Weibull: density = $(c/\sigma)\ (x/\sigma)^{c-1} \exp\{-(x/\sigma)^c\}$, where shape = c and scale = σ

Generally, the log-normal distribution was the best-fitting of the four distributions, and so, for consistency, we recommend using the fitted log-normal distributions for all the cases.

One limitation of the initial analysis was that distributions were available only for selected cities, and yet the summary statistics and comparisons demonstrate that the AER distributions depend upon the city as well as the temperature range and A/C type. As one option to address this issue, we considered modeling cities for which distributions were not available by using the AER distributions across all cities and dates for a given temperature range and A/C type.

Another important limitation of the initial analysis was that distributions were not fitted to all of the temperature ranges due to inadequate data. There are missing values between temperature ranges, and the temperature ranges are all bounded. To address this issue, the temperature ranges were regrouped to cover the entire range of temperatures from minus to plus infinity, although obviously the available data to fit these ranges have finite temperatures. Stratifying by A/C type, city, and the new temperature ranges produces results for four cities: Houston (AC and NA); Los Angeles (AC and NA); New York (AC and NA); Research Triangle Park (AC). For each of the fitted distributions we created histograms to compare the fitted distributions with the empirical distributions.

AER Distributions for The First Nine Cities. Based upon the results for the above four cities and the corresponding graphs, we propose using those fitted distributions for the three cities Houston, Los Angeles, and New York. For another 6 of the cities to be modeled, we propose using the distribution for one of the four cities thought to have similar characteristics to the city to be modeled with respect to factors that might influence AERs. These factors include the age composition of housing stock, construction methods, and other meteorological variables not explicitly treated in the analysis, such as humidity and wind speed patterns. The distributions proposed for these cities are as follows:

- Atlanta, GA, A/C: Use log-normal distributions for Research Triangle Park. Residences with A/C only.
- Boston, MA: Use log-normal distributions for New York
- Chicago, IL: Use log-normal distributions for New York
- Cleveland, OH: Use log-normal distributions for New York
- Detroit, MI: Use log-normal distributions for New York
- Houston, TX: Use log-normal distributions for Houston
- Los Angeles, CA: Use log-normal distributions for Los Angeles
- New York, NY: Use log-normal distributions for New York
- Philadelphia, PA: Use log-normal distributions for New York

Since the AER data for Research Triangle Park was only available for residences with air conditioning, AER distributions for Atlanta residences without air conditioning are discussed below.

To avoid unusually extreme simulated AER values, we propose to set a minimum AER value of 0.01 and a maximum AER value of 10.

Obviously, we would be prefer to model each city using data from the same city, but this approach was chosen as a reasonable alternative, given the available AER data.

AER Distributions for Sacramento and St. Louis. For these two cities, a direct mapping to one of the four cities Houston, Los Angeles, New York, and Research Triangle Park is not recommended because the cities are likely to be too dissimilar. Instead, we decided to use the distribution for the inland parts of Los Angeles to represent Sacramento and to use the aggregate distributions for all cities outside of California to represent St. Louis. The results for the city Sacramento were obtained by combining all the available AER data for Sacramento, Riverside, and San Bernardino counties. The results for the city St. Louis were obtained by combining all non-California AER data.

AER Distributions for Washington DC. Washington DC was judged likely to have similar characteristics both to Research Triangle Park and to New York City. To choose between these two cities, we compared the Murray and Burmaster AER data for Maryland with AER data from each of those cities. The Murray and Burmaster study included AER data for Baltimore and for Gaithersburg and Rockville, primarily collected in March. April, and May 1987, although there is no information on mean daily temperatures or A/C type. We collected all the March, April, and May AER data for Research Triangle Park and for New York City, and compared those distributions with the Murray and Burmaster Maryland data for the same three months.

The results for the means and central values show significant differences at the 5 percent level between the New York and Maryland distributions. Between Research Triangle Park and Maryland, the central values and the mean AER values are not statistically significantly different, and the differences in the mean log (AER) values are much less statistically significant than between New York and Maryland. The scale statistic comparisons are not statistically significantly different between New York and Maryland, but were statistically significantly different between Research Triangle Park and Maryland. Since matching central and mean values is generally more important than matching the scales, we propose to model Washington DC residences with air conditioning using the Research Triangle Park distributions, stratified by temperature:

- Washington DC, A/C: Use log-normal distributions for Research Triangle Park. Residences with A/C only.

Since the AER data for Research Triangle Park was only available for residences with air conditioning, the estimated AER distributions for Washington DC residences without air conditioning are discussed below.

AER Distributions for Washington DC and Atlanta GA Residences With No A/C. For Atlanta and Washington DC we have proposed to use the AER distributions for Research Triangle Park. However, all the Research Triangle Park data (from the RTP Panel study) were from houses with air conditioning, so there are no available distributions for the "No A/C" cases. For these two cities, one option is to use AER distributions fitted to all the study data for residences without A/C, stratified by temperature. We propose applying the "No A/C" distributions for modeling these two cities for residences without A/C. However, since Atlanta and Washington DC residences are expected to be better represented by residences outside of California, we instead propose to use the "No A/C" AER distributions aggregated across cities outside of California, which is the same as the recommended choice for the St. Louis "No A/C" AER distributions.

A/C Type and Temperature Distributions. Since the proposed AER distribution is conditional on the A/C type and temperature range, these values also need to be simulated using APEX in order to select the appropriate AER distribution. Mean daily temperatures are one of the available APEX inputs for each modeled city, so that the temperature range can be determined for each modeled day according to the mean daily temperature. To simulate the A/C type, we obtained estimates of A/C prevalence from the American Housing Survey. Thus for each city/metropolitan area, we obtained the estimated fraction of residences with Central or Room A/C (see Table A-3), which gives the probability p for selecting the A/C type "Central or Room A/C." Obviously, 1-p is the probability for "No A/C." For comparison with Washington DC and Atlanta, we have included the A/C type percentage for Charlotte, NC (representing Research Triangle Park, NC). As discussed above, we propose modeling the 96-97 % of Washington DC and Atlanta residences with A/C using the Research Triangle Park AER distributions, and modeling the 3-4 % of Washington DC and Atlanta residences without A/C using the combined study No A/C AER distributions.

Table A-3. Fraction of residences with central or room A/C (from American Housing Survey)

CITY	SURVEY AREA & YEAR	PERCENTAGE
Atlanta	Atlanta, 2003	97.01
Boston	Boston, 2003	85.23
Chicago	Chicago, 2003	87.09
Cleveland	Cleveland, 2003	74.64
Detroit	Detroit, 2003	81.41
Houston	Houston, 2003	98.70
Los Angeles	Los Angeles, 2003	55.05
New York	New York, 2003	81.57
Philadelphia	Philadelphia, 2003	90.61
Sacramento	Sacramento, 2003	94.63
St. Louis	St. Louis, 2003	95.53
Washington DC	Washington DC, 2003	96.47
Research Triangle Park	Charlotte, 2002	96.56

Other AER Studies

We recently became aware of some additional residential and non-residential AER studies that might provide additional information or data. Indoor / outdoor ozone and PAN distributions were studied by Jakobi and Fabian (1997). Liu et al (1995) studied residential ozone and AER distributions in Toronto, Canada. Weschler and Shields (2000) describes a modeling study of ventilation and air exchange rates. Weschler (2000) includes a useful overview of residential and non-residential AER studies.

AER Distributions for Other Indoor Environments

To estimate AER distributions for non-residential, indoor environments (e.g., offices and schools), we obtained and analyzed two AER data sets: "Turk" (Turk et al, 1989); and "Persily" (Persily and Gorfain 2004; Persily et al. 2005).

The earlier "Turk" data set (Turk et al, 1989) includes 40 AER measurements from offices (25 values), schools (7 values), libraries (3 values), and multi-purpose (5 values), each measured using an SF6 tracer over two- or four-hours in different seasons of the year.

The more recent "Persily" data (Persily and Gorfain 2004; Persily et al. 2005) were derived from the U.S. EPA Building Assessment Survey and Evaluation (BASE) study, which was conducted to assess indoor air quality, including ventilation, in a large number of randomly selected office buildings throughout the U.S. The data base consists of a total of 390 AER measurements in 96 large, mechanically ventilated offices; each office was measured up to four times over two days, Wednesday and Thursday AM and PM. The office spaces were relatively large, with at least 25 occupants, and preferably 50 to 60 occupants. AERs were measured both by a volumetric method and by a CO2 ratio method, and included their uncertainty estimates. For these analyses, we used the recommended "Best Estimates" defined by the values with the lower estimated uncertainty; in the vast majority of cases the best estimate was from the volumetric method.

Another study of non-residential AERs was performed by Lagus Applied Technology (1995) using a tracer gas method. That study was a survey of AERs in 16 small office buildings, 6 large office buildings, 13 retail establishments, and 14 schools. We plan to obtain and analyze these data and compare those results with the Turk and Persily studies.

Due to the small sample size of the Turk data, the data were analyzed without stratification by building type and/or season. For the Persily data, the AER values for each office space were averaged, rather using the individual measurements, to account for the strong dependence of the AER measurements for the same office space over a relatively short period.

Summary statistics of AER and log (AER) for the two studies are presented in Table A-4.

Table A-4. AER summary statistics for offices and other non-residential buildings

Study	Variable	N	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
Persily	AER	96	1.9616	2.3252	0.0712	0.5009	1.0795	2.7557	13.8237
Turk	AER	40	1.5400	0.8808	0.3000	0.8500	1.5000	2.0500	4.1000
Persily	Log(AER)	96	0.1038	1.1036	-2.6417	-0.6936	0.0765	1.0121	2.6264
Turk	Log(AER)	40	0.2544	0.6390	-1.2040	-0.1643	0.4055	0.7152	1.4110

The mean values are similar for the two studies, but the standard deviations are about twice as high for the Persily data. The proposed AER distributions were derived from the more recent Persily data only.

Similarly to the analyses of the residential AER distributions, we fitted exponential, log-normal, normal, and Weibull distributions to the 96 office space average AER values. The results are shown in Table A-5.

Table A-5. Best fitting office AER distributions from the Persily et al. (2004, 2005)

Scale	Shape	Mean	Std_Dev	Distribution	P-Value Kolmogorov-Smirnov	P-Value Cramer-von Mises	P-Value Anderson-Darling
1.9616		1.9616	1.9616	Exponential	0.13	0.04	0.05
0.1038	1.1036	2.0397	3.1469	Lognormal	0.15	0.46	0.47
		1.9616	2.3252	Normal	0.01	0.01	0.01
1.9197	0.9579	1.9568	2.0433	Weibull		0.01	0.01

(For an explanation of the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling P-values see the discussion residential AER distributions above.) According to all three goodness-of-fit measures the best-fitting distribution is the log-normal. Reasonable choices for the lower and upper bounds are the observed minimum and maximum AER values.

We therefore propose the following indoor, non-residential AER distributions.

- AER distribution for indoor, non-residential microenvironments: Lognormal, with scale and shape parameters 0.1038 and 1.1036, i.e., geometric mean = 1.1094, geometric standard deviation = 3.0150. Lower Bound = 0.07. Upper bound = 13.8.

Proximity and Penetration Factors For Outdoors, In-vehicle, and Mass Transit

For the APEX modeling of the outdoor, in-vehicle, and mass transit micro-environments, an approach using proximity and penetration factors is proposed, as follows.

Outdoors Near Road

Penetration factor = 1.

For the Proximity factor, we propose using ratio distributions developed from the Cincinnati Ozone Study (American Petroleum Institute, 1997, Appendix B; Johnson et al. 1995). The field study was conducted in the greater Cincinnati metropolitan area in August and September, 1994. Vehicle tests were conducted according to an experimental design specifying the vehicle type, road type, vehicle speed, and ventilation mode. Vehicle types were defined by the three study vehicles: a minivan, a full-size car, and a compact car. Road types were interstate highways (interstate), principal urban arterial roads (urban), and local roads (local). Nominal vehicle speeds (typically met over one minute intervals within 5 mph) were at 35 mph, 45 mph, or 55 mph. Ventilation modes were as follows:

- Vent Open: Air conditioner off. Ventilation fan at medium. Driver's window half open. Other windows closed.
- Normal A/C. Air conditioner at normal. All windows closed.
- Max A/C: Air conditioner at maximum. All windows closed.

Ozone concentrations were measured inside the vehicle, outside the vehicle, and at six fixed site monitors in the Cincinnati area.

The proximity factor can be estimated from the distributions of the ratios of the outside-vehicle ozone concentrations to the fixed-site ozone concentrations, reported in Table 8 of Johnson et al. (1995). Ratio distributions were computed by road type (local, urban, interstate, all) and by the fixed-site monitor (each of the six sites, as well as the nearest monitor to the test location). For this analysis we propose to use the ratios of outside-vehicle concentrations to the concentrations at the nearest fixed site monitor, as shown in Table A-6.

Table A-6. Ratio of outside-vehicle ozone to ozone at nearest fixed site[1]

Road Type[1]	Number of cases[1]	Mean[1]	Standard Deviation[1]	25th Percentile[1]	50th Percentile[1]	75th Percentile[1]	Estimated 5th Percentile[2]
Local	191	0.755	0.203	0.645	0.742	0.911	0.422
Urban	299	0.754	0.243	0.585	0.722	0.896	0.355
Interstate	241	0.364	0.165	0.232	0.369	0.484	0.093
All	731	0.626	0.278	0.417	0.623	0.808	0.170

1. From Table 8 of Johnson et al. (1995). Data excluded if fixed-site concentration < 40 ppb.
2. Estimated using a normal approximation as Mean – 1.64 × Standard Deviation

For the outdoors-near- road microenvironment, we recommend using the distribution for local roads, since most of the outdoors-near-road ozone exposure will occur on local roads. The summary data from the Cincinnati Ozone Study are too limited to allow fitting of distributions, but the 25th and 75th percentiles appear to be approximately equidistant from the median (50th percentile). Therefore we propose using a normal distribution with the observed mean and standard deviation. A plausible upper bound for the proximity factor equals 1. Although the normal distribution allows small positive values and can even produce impossible, negative values (with a very low probability), the titration of ozone concentrations near a road is limited. Therefore, as an empirical approach, we recommend a lower bound of the estimated 5th percentile, as shown in the final column of the above table. Therefore in summary we propose:

- Penetration factor for outdoors, near road: 1.

- Proximity factor for outdoors, near road: Normal distribution. Mean = 0.755. Standard Deviation = 0.203. Lower Bound = 0.422. Upper Bound = 1.

Outdoors, Public Garage / Parking Lot

This micro-environment is similar to the outdoors-near-road microenvironment. We therefore recommend the same distributions as for outdoors-near-road:

- Penetration factor for outdoors, public garage / parking lot: 1.
- Proximity factor for outdoors, public garage / parking lot: Normal distribution. Mean = 0.755. Standard Deviation = 0.203. Lower Bound = 0.422. Upper Bound = 1.

Outdoors, Other

The outdoors, other ozone concentrations should be well represented by the ambient monitors. Therefore we propose:

- Penetration factor for outdoors, other: 1.
- Proximity factor for outdoors, other: 1.

In-Vehicle

For the proximity factor for in-vehicle, we also recommend using the results of the Cincinnati Ozone Study presented in Table A-6. For this microenvironment, the ratios depend upon the road type, and the relative prevalences of the road types can be estimated by the proportions of vehicle miles traveled in each city. The proximity factors are assumed, as before, to be normally distributed, the upper bound to be 1, and the lower bound to be the estimated 5th percentile.

- Proximity factor for in-vehicle, local roads: Normal distribution. Mean = 0.755. Standard Deviation = 0.203. Lower Bound = 0.422. Upper Bound = 1.
- Proximity factor for in-vehicle, urban roads: Normal distribution. Mean = 0.754. Standard Deviation = 0.243. Lower Bound = 0.355. Upper Bound = 1.
- Proximity factor for in-vehicle, interstates: Normal distribution. Mean = 0.364. Standard Deviation = 0.165. Lower Bound = 0.093. Upper Bound = 1.

To complete the specification, the distribution of road type needs to be estimated for each city to be modeled. Vehicle miles traveled (VMT) in 2003 by city (defined by the Federal-Aid urbanized area) and road type were obtained from the Federal Highway Administration. (http://www.fhwa.dot.gov/policy/ohim/hs03/htm/hm71.htm). For local and interstate road types, the VMT for the same DOT categories were used. For urban roads, the VMT for all other road types was summed (Other freeways/expressways, Other principal arterial, Minor arterial, Collector). The computed VMT ratios for each city are shown in Table A-7.

Table A-7. Vehicle Miles Traveled by City and Road Type in 2003 (FHWA, October 2004)

FEDERAL-AID URBANIZED AREA	FRACTION VMT BY ROAD TYPE		
	INTERSTATE	URBAN	LOCAL
Atlanta	0.38	0.45	0.18
Boston	0.31	0.55	0.14
Chicago	0.30	0.59	0.12
Cleveland	0.39	0.45	0.16
Detroit	0.26	0.63	0.11
Houston	0.24	0.72	0.04
Los Angeles	0.29	0.65	0.06
New York	0.18	0.67	0.15
Philadelphia	0.23	0.65	0.11
Sacramento	0.21	0.69	0.09
St. Louis	0.36	0.45	0.19
Washington	0.31	0.61	0.08

Note that a “Federal-Aid Urbanized Area" is an area with 50,000 or more persons that at a minimum encompasses the land area delineated as the urbanized area by the Bureau of the Census. Urbanized areas that have been combined with others for reporting purposes are not shown separately. The Illinois portion of Round Lake Beach-McHenry-Grayslake has been reported with Chicago.

Thus to simulate the proximity factor in APEX, we propose to first select the road type according to the above probability table of road types, then select the AER distribution (normal) for that road type as defined in the last set of bullets.

For the penetration factor for in-vehicle, we recommend using the inside-vehicle to outside-vehicle ratios from the Cincinnati Ozone Study. The ratio distributions were summarized for all the data and for stratifications by vehicle type, vehicle speed, road type, traffic (light, moderate, or heavy), and ventilation. The overall results and results by ventilation type are shown in Table A-8.

Table A-8. Ratio of inside-vehicle ozone to outside-vehicle ozone[1]

Ventilation[1]	Number of cases[1]	Mean[1]	Standard Deviation[1]	25th Percentile[1]	50th Percentile[1]	75th Percentile[1]	Estimated 5th Percentile[2]
Vent Open	226	0.361	0.217	0.199	0.307	0.519	0.005
Normal A/C	332	0.417	0.211	0.236	0.408	0.585	0.071
Maximum A/C	254	0.093	0.088	0.016	0.071	0.149	0.000[3]
All	812	0.300	0.232	0.117	0.251	0.463	0.000[3]

1. From Table 7 of Johnson et al.(1995). Data excluded if outside-vehicle concentration < 20 ppb.
2. Estimated using a normal approximation as Mean – 1.64 × Standard Deviation
3. Negative estimate (impossible value) replaced by zero.

Although the data in Table A-8 indicate that the inside-to-outside ozone ratios strongly depend upon the ventilation type, it would be very difficult to find suitable data to estimate the ventilation type distributions for each modeled city. Furthermore, since the Cincinnati Ozone Study was scripted, the ventilation conditions may not represent real-world vehicle ventilation scenarios. Therefore, we propose to use the overall average distributions.

- Penetration factor for in-vehicle: Normal distribution. Mean = 0.300. Standard Deviation = 0.232. Lower Bound = 0.000. Upper Bound = 1.

Mass Transit

The mass transit microenvironment is expected to be similar to the in-vehicle microenvironment. Therefore we recommend using the same APEX modeling approach:

- Proximity factor for mass transit, local roads: Normal distribution. Mean = 0.755. Standard Deviation = 0.203. Lower Bound = 0.422. Upper Bound = 1.
- Proximity factor for mass transit, urban roads: Normal distribution. Mean = 0.754. Standard Deviation = 0.243. Lower Bound = 0.355. Upper Bound = 1.
- Proximity factor for mass transit, interstates: Normal distribution. Mean = 0.364. Standard Deviation = 0.165. Lower Bound = 0.093. Upper Bound = 1.
- Road type distributions for mass transit: See Table A-6
- Penetration factor for mass transit: Normal distribution. Mean = 0.300. Standard Deviation = 0.232. Lower Bound = 0.000. Upper Bound = 1.

References

American Petroleum Institute (1997). *Sensitivity testing of pNEM/O3 exposure to changes in the model algorithms.* Health and Environmental Sciences Department.

Avol, E. L., W. C. Navidi, and S. D. Colome (1998) Modeling ozone levels in and around southern California homes. *Environ. Sci. Technol.* 32, 463-468.

Chilrud, S. N., D. Epstein, J. M. Ross, S. N. Sax, D. Pederson, J. D. Spengler, P. L. Kinney (2004). Elevated airborne exposures of teenagers to manganese, chromium, and iron from steel dust and New York City's subway system. *Environ. Sci. Technol.* 38, 732-737.

Colome, S.D., A. L. Wilson, Y. Tian (1993). *California Residential Indoor Air Quality Study, Volume 1, Methodology and Descriptive Statistics.* Report prepared for the Gas Research Institute, Pacific Gas & Electric Co., San Diego Gas & Electric Co., Southern California Gas Co.

Colome, S.D., A. L. Wilson, Y. Tian (1994). *California Residential Indoor Air Quality Study, Volume 2, Carbon Monoxide and Air Exchange Rate: An Univariate and Multivariate Analysis. Chicago, IL.* Report prepared for the Gas Research Institute, Pacific Gas & Electric Co., San Diego Gas & Electric Co., Southern California Gas Co. GRI-93/0224.3

Jakobi, G and Fabian, P. (1997). Indoor/outdoor concentrations of ozone and peroxyacetyl nitrate (PAN). *Int. J. Biometeorol.* 40: 162-165..

Johnson, T., A. Pakrasi, A. Wisbeth, G. Meiners, W. M. Ollison (1995). Ozone exposures within motor vehicles – results of a field study in Cincinnati, Ohio. *Proceedings 88th annual meeting and exposition of the Air & Waste Management Association, June 18-23, 1995.* San Antonio, TX. Preprint paper 95-WA84A.02.

Kinney, P. L., S. N. Chillrud, S. Ramstrom, J. Ross, J. D. Spengler (2002). Exposures to multiple air toxics in New York City. *Environ Health Perspect* 110, 539-546.

Lagus Applied Technology, Inc. (1995) *Air change rates in non-residential buildings in California.* Sacramento CA, California Energy Commission, contract 400-91-034.

Liu, L.-J. S, P. Koutrakis, J. Leech, I. Broder, (1995) Assessment of ozone exposures in the greater metropolitan Toronto area. *J. Air Waste Manage. Assoc*. 45: 223-234.

Meng, Q. Y., B. J. Turpin, L. Korn, C. P. Weisel, M. Morandi, S. Colome, J. J. Zhang, T. Stock, D. Spektor, A. Winer, L. Zhang, J. H. Lee, R. Giovanetti, W. Cui, J. Kwon, S. Alimokhtari, D. Shendell, J. Jones, C. Farrar, S. Maberti (2004). Influence of ambient (outdoor) sources on residential indoor and personal $PM_{2.5}$ concentrations: Analyses of RIOPA data. *Journal of Exposure Analysis and Environ Epidemiology*. Preprint.

Murray, D. M. and D. E. Burmaster (1995). Residential Air Exchange Rates in the United States: Empirical and Estimated Parametric Distributions by Season and Climatic Region. *Risk Analysis*, Vol. 15, No. 4, 459-465.

Persily, A. and J. Gorfain.(2004). *Analysis of ventilation data from the U.S. Environmental Protection Agency Building Assessment Survey and Evaluation (BASE) Study.* National Institute of Standards and Technology, NISTIR 7145, December 2004.

Persily, A., J. Gorfain, G. Brunner.(2005). Ventilation design and performance in U.S. office buildings. *ASHRAE Journal.* April 2005, 30-35.

Sax, S. N., D. H. Bennett, S. N. Chillrud, P. L. Kinney, J. D. Spengler (2004) Differences in source emission rates of volatile organic compounds in inner-city residences of New York City and Los Angeles. *Journal of Exposure Analysis and Environ Epidemiology*. Preprint.

Turk, B. H., D. T. Grimsrud, J. T. Brown, K. L. Geisling-Sobotka, J. Harrison, R. J. Prill (1989). *Commercial building ventilation rates and particle concentrations*. ASHRAE, No. 3248.

Weschler, C. J. (2000) Ozone in indoor environments: concentration and chemistry. Indoor Air 10: 269-288.

Weschler, C. J. and Shields, H. C. (2000) The influence of ventilation on reactions among indoor pollutants: modeling and experimental observations. *Indoor Air*. 10: 92-100.

Weisel, C. P., J. J. Zhang, B. J. Turpin, M. T. Morandi, S. Colome, T. H. Stock, D. M. Spektor, L. Korn, A. Winer, S. Alimokhtari, J. Kwon, K. Mohan, R. Harrington, R. Giovanetti, W. Cui, M. Afshar, S. Maberti, D. Shendell (2004). Relationship of Indoor, Outdoor and Personal Air (RIOPA) study; study design, methods and quality assurance / control results. *Journal of Exposure Analysis and Environ Epidemiology*. Preprint.

Williams, R., J. Suggs, A. Rea, K. Leovic, A. Vette, C. Croghan, L. Sheldon, C. Rodes, J. Thornburg, A. Ejire, M. Herbst, W. Sanders Jr. (2003a). The Research Triangle Park particulate matter panel study: PM mass concentration relationships. *Atmos Env* 37, 5349-5363.

Williams, R., J. Suggs, A. Rea, L. Sheldon, C. Rodes, J. Thornburg (2003b). The Research Triangle Park particulate patter panel study: modeling ambient source contribution to personal and residential PM mass concentrations. *Atmos Env* 37, 5365-5378.

Wilson, A. L., S. D. Colome, P. E. Baker, E. W. Becker (1986). *Residential Indoor Air Quality Characterization Study of Nitrogen Dioxide, Phase I, Final Report.* Prepared for Southern California Gas Company, Los Angeles.

Wilson, A. L., S. D. Colome, Y. Tian, P. E. Baker, E. W. Becker, D. W. Behrens, I. H. Billick, C. A. Garrison (1996). California residential air exchange rates and residence volumes. *Journal of Exposure Analysis and Environ Epidemiology*. Vol. 6, No. 3.

APPENDIX B. THEORETICAL DEVELOPMENT OF A UNIFIED ALGORITHM FOR ADJUSTING MET VALUES IN HUMAN EXPOSURE MODELING FOR FATIGUE AND EXCESS POST-EXERCISE OXYGEN CONSUMPTION (EPOC)

This page intentionally left blank.

TECHNICAL MEMORANDUM

TO: Tom McCurdy, U.S. EPA, WA Manager, NERL WA 131

FROM: Kristin Isaacs, Graham Glen, and Luther Smith, Alion Science and Technology Inc.

DATE: June 16, 2005

SUBJECT: **Theoretical Development of a Unified Algorithm for Adjusting MET Values in Human Exposure Modeling for Fatigue and EPOC**

I. INTRODUCTION

The CHAD activity database assigns distributions for energy expenditure to each diary event, based on the reported event activity. This is done using the MET paradigm, which uses ratios of activity-specific to basal energy expenditure. However, the basic or "raw" MET distributions do not consider sequences of events. It is well known that a person's capacity for work will diminish as they get tired, and in practice, this means that the upper bound on MET is lowered if events in the recent past have been at unusually high MET levels. Furthermore, once high activity levels have ended, people tend to breathe heavily even while resting, as they recover their accumulated oxygen deficit. This effect is called excess post-exercise oxygen consumption (EPOC), and results in raising the MET levels above the 'raw' values pulled from the activity-based distributions.

Historically, the logic for the downward adjustments (downward limitations on the maximum MET with increasing fatigue) was developed before the EPOC adjustments. The pNEM model included downward adjustments, both for single events and averages over many diary events. The rules for these adjustments are given in a report[1] by Ted Johnson describing the pNEM algorithms. These rules were incorporated into CHAD and APEX without alteration. The rules for the EPOC adjustments were developed later by G. Glen and added to CHAD. They were not included in APEX or any of the SHEDS models.

Rather than separately accounting for these effects, it is more logical to make both adjustments simultaneously. This would prevent the possibility of making a downward adjustment so that the MET average conforms to a given limit, but then have the EPOC adjustment boost the average back above that limit. Also, the current method of making the adjustments is computationally burdensome. For these reasons, we have developed a new approach.

The proposed adjustment algorithm imposes limits on MET via the value of an oxygen deficit an individual has incurred. The method is more computationally efficient than previous MET-adjustment algorithms, and eliminates some of the problematic features of the current methods.

II. THEORETICAL DEVELOPMENT OF THE METHOD

Background: Oxygen Deficit, Physiological Limits on MET, and EPOC

At the beginning of exercise, there is a lag between work expended and oxygen consumption.[2] During this work/ventilation mismatch, an individual's energy needs are met by anaerobic processes. The magnitude of the mismatch between expenditure and consumption is termed the oxygen deficit. During heavy exercise, further oxygen deficit (in addition to that associated with the start of exercise) may be accumulated. At some point, oxygen deficit reaches a maximum value, and performance and energy expenditure deteriorate.

After exercise ceases, ventilation and oxygen consumption will remain elevated above baseline levels. This increased oxygen consumption was historically labeled the "oxygen debt" or "recovery oxygen consumption." However, recently the term "excess post-exercise oxygen consumption" (EPOC) has been adopted for the phenomenon.

The new method for adjusting the MET values is based on keeping a running total of the oxygen deficit as one proceeds chronologically through an activity diary. The oxygen deficit calculations were derived from numerous published studies. Oxygen deficit is measured as a percentage of the maximum oxygen deficit an individual can attain prior to deterioration of performance. Limitations on MET levels corresponding to post-exercise diary events were based on maintaining an oxygen deficit below this maximum value. In addition, adjustments to MET were simultaneously made for EPOC. The EPOC adjustments are based in part on the modeled oxygen deficit and in part on data from published studies on EPOC, oxygen deficit, and oxygen consumption.

As instructed by the EPA WAM, the methods were constructed in terms of reserve MET rather than total MET. The reserve is the amount over the basal rate (MET=1). Furthermore, we defined M as the normalized reserve, so that M=0 at MET=1, and M=1 at maximum MET:

$$M = \frac{\text{METS} - 1}{\text{METS}_{\max} - 1} \qquad (1)$$

Using a normalized reserve assures that the method can be applied identically to a population of individuals having widely different $\text{MET}_{\max}$ values.

Nomenclature

MET	Metabolic equivalent (unitless)
$\text{MET}_{\max}$	Maximum achievable metabolic equivalent for an individual (unitless)
M	Normalized MET reserve (unitless, M, bounded between 0 and 1)
ΔM	Change in M from one diary event to the next (M)
$D_{\max}$	Absolute maximum oxygen deficit that can be obtained (M-hr)
F	Fractional oxygen deficit (percent of individual maximum, unitless)
t_e	Duration of activity diary event (hours)
t_r	Time required to recover from an F of 1 to an F of 0 at rest (recovery time, hours)

dF_{inc}	Rate of change of F due to deficit increase (F/hr, will have a positive value)
dF_{rec}	Rate of change of F due to deficit recovery (F/hr, will have a negative value)
dF_{tot}	Total rate of change of F, dF_{inc+} dF_{rec} (F/hr)
ΔF_{inc}	Increase in F due to anaerobic energy expenditure (F)
ΔF_{rec}	Decrease in F due to recovery of oxygen deficit (F)
ΔF_{tot}	Change in F due to simultaneous anaerobic work and oxygen recovery, $\Delta F_{inc+}\Delta F_{rec}$ (F)
ΔF_{fast}	Total change in F during the fast recovery phase (F)
S_{fast}	Magnitude of the rate of change in M during fast component (M/hr)
$EPOC_{fast}$	Change in M due to fast-component EPOC (M)
$EPOC_{slow}$	Change in M due to slow-component EPOC (M)
PAI	Physical activity index (median of daily average MET, dimensionless)

Simulation of Oxygen Deficit

This section presents the theoretical development of the equations describing the accumulation of oxygen deficit. We developed the method using a large number of studies on oxygen consumption, oxygen deficit, and EPOC. Individual studies will be referenced below. The first two sections below describe the equations themselves, while the last section describes the determination of appropriate values for the model parameters.

Fast Processes. There exists a component of the accumulated oxygen deficit that is due to transition from one M level to another.[2] This component derives from the anaerobic work that is required by sudden muscular motion. There is also a corresponding fast component of oxygen recovery which occurs very quickly after a change from a high M level to a lower one. In the absence of any data to the contrary, it is assumed that these fast deficit accumulation and fast recovery processes occur at the same rate. These processes are illustrated in the Figure 1. The adjustment to F is equal to the area of the triangle associated with either a positive or negative change in M, normalized by the maximum obtainable accumulated oxygen deficit (D_{max}). The normalized area can thus be calculated as:

$$\Delta F_{fast} = 0.5\frac{\Delta M|\Delta M|}{S_{fast} D_{max}} \tag{2}$$

where $\Delta M = M_i - M_{i-1}$ and S_{fast} is the slope of the change in M (in M/hr). Note that this change in F will be positive if ΔM is positive, and negative otherwise.

Slow Processes. The slow component of the increase in oxygen deficit corresponds to the accumulation of deficit over a period of heavier exercise (rather than that associated with an increase in activity level). The starting point for the analyses is the table of data[3-15] assembled by T. McCurdy for the 1998 and 1999 EPOC work. Data from a selection of these studies in which persons exercised to exhaustion are given in Table 1. The table includes the time it took for subjects to reach exhaustion, their accumulated oxygen deficit, their MET_{max}, the MET value at which they exercised, and the corresponding normalized reserve MET (M). (Note that the MET and MET_{max} quantities in this table were derived from VO_2 and VO_{2max} measurements.) A plot of M versus duration is shown in Figure 2. There is one data point having M > 1, for one subject who exercised briefly at a level above his/her $\mathrm{MET}_{\max}$. The data indicate that oxygen deficit accumulates at a much faster rate when M is high. For example, an M value near 0.5 requires about 5 times longer to reach exhaustion than an M value near 0.75 (on average), indicating that F is nonlinear in M.

Let the rate of increase in F be given by dF_{inc}. Based upon the relationship depicted in Fig. 2, we postulate a simple nonlinear relationship between dF_{inc} and M as a power law:

$$dF_{inc} = aM^{b} \tag{3}$$

However, before estimating a and b, one must account for slow recovery of oxygen debt, as it occurs simultaneously with debt accumulation. We assume a slow, but continual, process for recovering oxygen deficit that is independent of the MET level. For modeling purposes, time-varying processes are very difficult to handle, especially when using finite time-step models. In our exposure models, the time step may be as large as one hour. To avoid problems, we model the slow EPOC recovery as constant over time, until the oxygen deficit is erased. Assuming this takes t_r hours, the slow recovery of oxygen deficit occurs at a rate

$$dF_{rec} = -\frac{1}{t_r} \tag{4}$$

The total net rate of change in F from slow processes during an event i with duration t_e is given by

$$dF_{slow} = dF_{inc} + dF_{rec} \tag{5}$$

and the associated change in F is

$$\Delta F_{slow} = \left(aM_i^b - \frac{1}{t_r} \right) t_e \tag{6}$$

For an individual starting with an F of 0 and exercising to exhaustion (neglecting the transitory effects), the change in ΔF is 1.0. In this case, rearranging and taking the logarithm gives

$$\log\left(\frac{1}{t} + \frac{1}{t_r} \right) = \log(a) + b\log(M) \tag{7}$$

This equation can be used to fit data to estimate the parameters a and b (this will be discussed in the next section).

The starting normalized oxygen deficit for the next event (i +1), taking into account both the fast and slow changes in F, is then

$$F_{i+1} = F_i + \Delta F_{slow} + \Delta F_{Fast} \tag{8}$$

Appropriate values for t_r, a, and b. These parameters were derived from summaries of published data that were supplied by EPA (i.e., the data in Table 1). It should be noted that these data were collected and analyzed some years ago and should be updated to include any recent additions to the literature. As additional data become available, the parameter values estimated here may be adjusted without changing the structure of the algorithm.

Several of the studies in Table 1 reported t_r values. However, due to variability in measurement and protocol differences, these recovery times varied from 0.5 hours to 24 hours. From a modeling viewpoint, it would be unacceptable to allow recovery to significantly carry over from one day to the next. To do so could lead to a perpetual delay in recovering an oxygen deficit, for example, by repeatedly encountering new exercise events before recovery is complete. For the results section, we chose t_r from a uniform distribution having a minimum of 8 and a maximum of 16 hours. (In practice, the values selected for t_r do not affect the result significantly.) The user could replace this distribution, if desired.

Eq. 7 was fit to the data (Table 1) using different values of t_r to obtain estimates of a and b. The results are shown in Table 2. The results were summarized to obtain the following expressions for a and b:

$$a = 5.20 - \left(\frac{1.54}{t_r}\right) + \left(\frac{3.92}{t_r^2}\right), \tag{9}$$

$$b = 3.93 - \left(\frac{3.57}{t_r}\right) + \left(\frac{3.66}{t_r^2}\right). \tag{10}$$

Values for D_{max}. Appropriate distributions for maximum oxygen debt (MOD) in ml/kg were derived from data from a number of studies in adults,[16-29] adolescents,[30] and children.[31-32] The studies covered multiple types of exercise protocols, some having more than one protocol per study. We chose to define normal distributions for MOD in all three age groups, based on average mean and standard deviation values from the studies:

adults:	54.95±14.46 (ml/kg)
adolescents:	63.95±21.12 (ml/kg)
children:	34.74±13.10 (ml/kg)

Values were selected from normal distributions with these characteristics. The bounds of these distributions were selected as two standard deviations from the mean; these ranges were found to be reasonable when compared to reported ranges.[29] The means for each exercise protocol from the studies for all three age groups are shown in the plots in Fig. 3, and the data for all the studies are given in Table 5. For use in Eq. 2, we transformed these values to $D_{max,}$ via a units conversion factor and the normalization needed for use with reserve MET:

$$D_{max}\ \text{(M-hr)} = \left(\frac{\text{MOD}}{60\ \text{METStoO}_2} \right) (\text{METS}_{max} - 1)^{-1} \tag{12}$$

where METtoO$_2$ is the conversion factor[2] for mlO$_2$ to MET-min, 3.5 [(mlO2/min)/kg]/MET. Note that the variability in this factor is not addressed here.

Values for $S_{fast.}$ A number of studies on EPOC[33-42] were used to derive S_{fast}. These were all studies in which oxygen consumption was measured relatively soon (within a few minutes) after the end of exercise and at a frequency high enough to capture the kinetics of the change in oxygen consumption. The data were found to be relatively uniform from the minimum (0.6 MET/min) to the maximum (3.7 MET/min) slope values, and so values were selected from a uniform distribution having these bounds. Converting units and normalizing to M, one obtains:

$$S_{fast}\ \text{(M/hr)} = \frac{60\ \text{Uniform}\ (0.6, 3.7)}{(\text{METS}_{max} - 1)} \tag{13}$$

The data for all studies are given in Table 6.

Adjustments to M for Fatigue

The equations provided in the previous section describe a method for keeping a running total of the fractional oxygen deficit (F) for each diary event for an individual. We used these event F values to limit M for each event to appropriate values. Basically, the maximum M value that can be maintained for an entire event is the value that would result in an F_{i+1} (eq. 8) equal to 1 (i.e., the maximum value) at the end of the diary event. Ideally, one would wish to solve Eqs. 2, 6, and 8 explicitly for M_i for a value of F_{i+1} =1. However, the equations are non-linear in M_i. The approach used here is to set M for each event equal to the raw MET value, and test if F_{i+1}>1. If it is, then the M_i value is reduced by a predetermined amount (currently 0.01) and F_{i+1} is recalculated. The process continues until an appropriate value of M_i, called $M_{max,i,}$ is found. As the exposure model marches through the events of the activity diary, the M values associated with each event are adjusted if necessary:

$$M_i = \min(M_i, M_{max,\ i}) \tag{14}$$

Adjustments to M for EPOC

As noted above, it has been observed in many studies that EPOC is characterized by both slow and fast components. The fast component occurs within minutes of exercise, while the slow component may persist for many hours. Both fast and slow EPOC components were modeled.

Fast Processes. The fast EPOC component, which takes place in the first few minutes after exercise, is also characterized by the slope $S_{fast.}$ The energy recovered during those first few minutes corresponds to the recovery triangle in Fig. 1, and this increase in the rate of energy expenditure for a post-exercise event is modeled as the area of the triangle divided by the event duration:

$$\mathrm{EPOC}_{\mathrm{fast}} = 0.5\frac{(\Delta \mathrm{M})^2}{\mathrm{S}_{\mathrm{fast}}\mathrm{t}_{\mathrm{e}}} \tag{15}$$

$\mathrm{EPOC}_{\mathrm{fast}}$ will thus have units of M (normalized reserve MET). The M level for the post-exercise events will be incremented by $\mathrm{EPOC}_{\mathrm{fast.}}$

Slow Processes. We estimate the increase in M associated with the slow EPOC component as the amount required to maintain the slow recovery of F. Since a deficit D_{max} is recovered in full in the recovery time t_r, the time-averaged adjustment to MET for the slow recovery process must be

$$\mathrm{EPOC}_{\mathrm{slow}}=\frac{\mathrm{D}_{\mathrm{max}}}{\mathrm{t}_{\mathrm{r}}} \tag{16}$$

Every diary event with the full rate of slow recovery will have its M value adjusted upward by $\mathrm{EPOC}_{\mathrm{slow}}$. An appropriate fraction of EPOC slow is used if only partial recovery is needed to eliminate the deficit (i.e., return F to 0). The final adjusted M value for the diary event is thus

$$\mathrm{M}_{\mathrm{adj}} = \mathrm{M} + \mathrm{EPOC}_{\mathrm{fast}} + \mathrm{EPOC}_{\mathrm{slow}} \tag{17}$$

and the new MET value for the event is

$$\mathrm{METS}_{\mathrm{adj}} = \mathrm{M}_{\mathrm{adj}}(\mathrm{METS}_{\mathrm{max}} - 1) + 1 \tag{18}$$

II. DISCUSSION

Note: As the main focus of this document was the presentation of the method, only a general summary of the modeling results are presented here. An in-depth analysis of PAI, dose, and ventilation modeling results for children within 36 age and gender cohorts were presented in the report *Analysis of Data Relating to EPOC and Duration-Dependent Limits on MET* (Kristin Isaacs and Graham Glen, February 5, 2005). Though that report utilized an earlier version of MET-adjustment, it is likely that the results presented there would not vary greatly from those obtained using the method discussed here. The PAI results presented herein demonstrate that the new method decreases MET (and thus PAI) a bit more than the earlier method. However, it is expected that this decrease will be fairly uniform across cohorts.

MET Limits for Fatigue

For periods of constant exercise, Eq. 7 results in a function having a horizontal asymptote. This asymptote is the M level that the individual can sustain indefinitely, and above which oxygen deficit accumulates. At this M, the net change in F is zero because recovery exactly balances the increase in F. A plot of M_{max} assuming constant exercise at M_{max} and a t_r of 12 hours is given in

Fig. 4 These results are very close to those predicted by the Bink[43] equation (as modified by Erb[44]), which was used in the previous method to limit MET. This demonstrates that for continuous exercise, the limits on intensity predicted by the unified method decline appropriately with time.

Existing methods have a tendency to overcorrect MET values for relatively low-activity events to fulfill limits on subsequent high-MET events. By imposing limits on MET via the current value of the oxygen deficit, the unified method avoids overcorrection and implements more localized adjustments in MET. For example, consider the case of the 3 year-old whose one, four, and nine-hour running MET averages are shown in Fig. 5. The flat dotted lines represent the MET limits predicted by Bink[43] for constant exercise at these time intervals. In Fig. 5 note that the new method allows the MET curve to approach the Bink limit without exceeding it, whereas overcorrection in the earlier methods prevents MET from even approaching this limit.

The actual adjustments made to MET time series varied greatly. Much of this variation was dependent on individual differences in $MET_{max.}$ Three examples for children of different age, gender, and MET_{max} are given in Figure 6.

EPOC

The adjustments to single-event MET for EPOC when the algorithm was applied to CHAD were very small compared to the adjustments made for fatigue. In general, the increases in MET for EPOC were small, usually less than one MET. In a few cases the adjustments were bigger (on the order of 4-6 MET), due to the fact that the adjustments were applied to a very short event. An example of a MET time series with EPOC adjustments is shown in Fig. 7.

As more conclusive data become available, the slow EPOC processes could be modeled in a similar manner to the fast component, (i.e., with a slope term). Currently, data on the duration and magnitude of the slow EPOC component are inconclusive and vary greatly from study to study. However, the change in M for slow EPOC is extremely small, and thus it would be expected that a different modeling method for this component would have a negligible effect on M (and thus ventilation and dose).

EFFECT ON PAI IN CHILDREN

Mean values of PAI for age and gender cohorts are given in Tables 3 and 4. In general, the unified algorithm resulted in a decreased PAI. A frequency distribution for PAI is given in Fig. 8. The algorithm shifted the distribution of PAI to the left, with the higher end of the distribution being most affected. That is, the unified algorithm mainly adjusted the highest values of PAI.

III. SUMMARY AND CONCLUSIONS

We have developed a new method for simultaneously correcting MET values for fatigue and excess post-exercise oxygen consumption. The method is based on the calculation of an accumulated oxygen deficit. The method's equations were derived from data from a large number of studies on oxygen deficit and EPOC. Furthermore, the model variables can be easily updated to incorporate data from future studies as they become available. However, the method as presented here returns qualitatively appropriate results for time-dependent averages of MET levels for children, though fine tuning of the results might be obtained by updating the model parameter estimates using new data.

The new method is more computationally efficient and theoretically straightforward than the previous ones. It requires no maintenance of multiple running averages of MET values (as was required by the previous algorithm) or recursive nonlinear adjustment of oxygen deficit (as was required by the other methods).

References

1. Johnson, T. and Capel, J. "Software for estimating ventilation (respiration) rates for use in dosimetry models. May, 2002.

2. McArdle, WD, Katch, FI, and Katch, VL. Exercise Physiology: Energy, Nutrition, and Human Performance, Fifth Edition. Lippincott, Williams, and Wilkins, Philadephia, 2001.

3. Bahr R. Excess postexercise oxygen consumption--magnitude, mechanisms and practical implications. Acta Physiol Scand Suppl. 605:1-70, 1992.

4. Bahr R, Ingnes I, Vaage O, Sejersted OM, Newsholme EA. Effect of duration of exercise on excess postexercise O2 consumption. J Appl Physiol. 62(2):485-90, 1987.

5. Sedlock DA. Postexercise energy expenditure following upper body exercise. Res Q Exerc Sport. 62(2):213-6, 1991.

6. Sedlock DA. Effect of exercise intensity on postexercise energy expenditure in women. Br J Sports Med. 25(1):38-40, 1991.

7. Bielinski R, Schutz Y, Jequier E. Energy metabolism during the postexercise recovery in man. Am J Clin Nutr. 42(1):69-82, 1985.

8. Brockman L, Berg K, Latin R. Oxygen uptake during recovery from intense intermittent running and prolonged walking. J Sports Med Phys Fitness. 33(4):330-6, 1993.

9. Gillette CA, Bullough RC, Melby CL. Postexercise energy expenditure in response to acute aerobic or resistive exercise. Int J Sport Nutr. Dec;4(4):347-60, 1994.

10. Gore CJ, Withers RT. Effect of exercise intensity and duration on postexercise metabolism. J Appl Physiol. 68(6):2362-8, 1990.

11. Hagberg JM, Hickson RC, Ehsani AA, Holloszy JO. Faster adjustment to and recovery from submaximum exercise in the trained state.J Appl Physiol. 48(2):218-24, 1980.

12. Maehlum S, Grandmontagne M, Newsholme EA, Sejersted OM. Magnitude and duration of excess postexercise oxygen consumption in healthy young subjects. Metabolism. 35(5):425-9, 1986.

13. Harris JM, Hobson EA, and Hollingsworth DF. Individual variations in energy expenditure and intake. Proc Nutr Soc. 21: 157-169, 1962.

14. Kaminsky LA, and Whaley MH. Effect of interval-type exercise on excess post-exercise oxygen consumption in obese and normal-weight women. Med Exer Nutr Health. 2:106-111, 1993.

15. Katch FI, Girandola RN, and Henry FM. The influence of the estimated oxygen cost of ventilation on oxygen deficit and recovery oxygen intake for moderately heavy bicycle ergometer exercise. Med Sci Sports 4: 71-76, 1972.

16. Gastin PB, Costill DL, Lawson DL, Krzeminski K, McConell GK. Accumulated oxygen deficit during supramaximum all-out and constant intensity exercise. Med Sci Sports Exerc. 27(2):255-63, 1995.

17. Gastin PB, Lawson DL. Variable resistance all-out test to generate accumulated oxygen deficit and predict anaerobic capacity. Eur J Appl Physiol Occup Physiol. 69(4):331-6, 1994.

18. Weber CL, Schneider DA. Maximum accumulated oxygen deficit expressed relative to the active muscle mass for cycling in untrained male and female subjects. Eur J Appl Physiol. 82(4):255-61, 2000.

19. Doherty M, Smith PM, Schroder K. Reproducibility of the maximum accumulated oxygen deficit and run time to exhaustion during short-distance running. J Sports Sci. 18(5):331-8, 2000.

20. Renoux JC, Petit B, Billat V, Koralsztein JP. Oxygen deficit is related to the exercise time to exhaustion at maximum aerobic speed in middle distance runners. 1: Arch Physiol Biochem. 107(4):280-5, 1999.

21. Roberts AD, Clark SA, Townsend NE, Anderson ME, Gore CJ, Hahn AG. Changes in performance, maximum oxygen uptake and maximum accumulated oxygen deficit after 5, 10 and 15 days of live high:train low altitude exposure. Eur J Appl Physiol. 88(4-5):390-5, 2003.

22. Maxwell NS, Nimmo MA. Anaerobic capacity: a maximum anaerobic running test versus the maximum accumulated oxygen deficit. Can J Appl Physiol. 21(1):35-47, 1996.

23. Buck D, McNaughton L. Maximum accumulated oxygen debt must be calculated using 10 min time periods. Med Sci Sports Exerc. 31(9):1346-1349, 1999.

24. Hill DW, Ferguson CS, Ehler KL.An alternative method to determine maximum accumulated O2 deficit in runners. Eur J Appl Physiol Occup Physiol. 79(1):114-7, 1998.

25. Demarle AP, Slawinski JJ, Laffite LP, Bocquet VG, Koralsztein JP, Billat VL. Decrease of O(2) deficit is a potential factor in increased time to exhaustion after specific endurance training. J Appl Physiol. 90(3):947-53, 2001.

26. Bickham D, Le Rossignol P, Gibbons C, Russell AP. Re-assessing accumulated oxygen deficit in middle-distance runners.J Sci Med Sport. 5(4):372-82, 2002.

27. Faina M, Billat V, Squadrone R, De Angelis M, Koralsztein JP, Dal Monte A. Anaerobic contribution to the time to exhaustion at the minimal exercise intensity at which maximum

oxygen uptake occurs in elite cyclists, kayakists and swimmers. Eur J Appl Physiol Occup Physiol. 76(1):13-20, 1997.

28. Billat V, Beillot J, Jan J, Rochcongar P, Carre F. Gender effect on the relationship of time limit at 100% VO2max with other bioenergetic characteristics. Med Sci Sports Exerc. 28(8):1049-55, 1996.

29. Olesen HL. Accumulated oxygen deficit increases with inclination of uphill running. J Appl Physiol. 73(3):1130-4, 1992.

30. Naughton GA, Carlson JS, Buttifant DC, Selig SE, Meldrum K, McKenna MJ, Snow RJ. Accumulated oxygen deficit measurements during and after high-intensity exercise in trained male and female adolescents. Eur J Appl Physiol Occup Physiol. 76(6):525-31, 1998.

31. Carlson JS, Naughton GA. An examination of the anaerobic capacity of children using maximum accumulated oxygen debt. Pediatr Exerc Sci. 5:60-71, 1993.

32. Berthoin S, Baquet G, Dupont G, Blondel N, Mucci P. Critical velocity and anaerobic distance capacity in prepubertal children. Can J Appl Physiol. 28(4):561-75, 1996.

33. Knuttgen HG. Oxygen debt after submaximum physical exercise. J Appl Physiol 29(5):651-657, 1970.

34. Harms CA, Cordain L, Stager JM, Sockler JM, and Harris M. Body fat mass affects postexercise oxygen metabolism in males of similar lean body mass Med Exer Nutr Health 4:33-39, 1995.

35. Trost S, Wilcox A, Gillis D. The effect of substrate utilization, manipulated by nicotinic acid, on excess postexercise oxygen consumption Int J Sports Med 18(2):83-88, 1997.

36. Dawson B, Straton S, and Randall N. Oxygen consumption during recovery from prolonged submaximum cycling below the anaerobic threshold. J Sports Med Phys Fitness, 36:77-84, 1996.

37. Short KR, Sedlock DA. Excess postexercise oxygen consumption and recovery rate in trained and untrained subjects. J Appl Physiol, 83(1):153-159, 1997.

38. Pivarnik JM, Wilkerson JE. Recovery metabolism and thermoregulation of endurance trained and heat acclimatized men. Sports Med Phys Fitness 28(4):375-80, 1988.

39. Almuzaini KS, Potteiger JA, Green SB. Effects of split exercise sessions on excess postexercise oxygen consumption and resting metabolic rate. Can J Appl Physiol 23(5):433-43, 1998.

40. Frey GC, Byrnes WC, and Mazzeo RS. Factors influencing excess postexercise oxygen consumption in trained and untrained women. Metabolism 42(7):822-828, 1993.

41. Kaminsky LA, Padjen S, LaHam-Saeger. J Effect of split exercise sessions on excess post-exercise oxygen consumption. Br J Sports Med 1990 Jun;24(2):95-8.

42. Maresh CM, Abraham A, De Souza MJ, Deschenes MR, Kraemer WJ, Armstrong LE, Maguire MS, Gabaree CL, Hoffman JR. Oxygen consumption following exercise of moderate intensity and duration. Eur J Appl Physiol Occup Physiol 1992;65(5):421-6.

43. Bink B. The physical working capacity in relation to working time and age Ergonomics, 5(1):29-31, 1962.

44. Erb BD. Applying work physiology to occupational medicine. Occup Health and Safety 50(6):20-24, 1981.

Table 1. Data used for estimation of oxygen deficit. Table gives data for 22 subjects exercising to exhaustion. Oxygen deficit was assumed to be 0 at the start of exercise.

Observation	**Oxygen Deficit**	**Time to Reach Exhaustion**	**METSMax**	**METS**	**M**
	(ml/kg)	(Hours)			
1	153.52	1.00	13.3	9.31	0.67561
2	109.95	1.17	13.8	10.35	0.730469
3	106.43	1.33	13.46	9.57	0.687801
4	93.92	1.27	16.11	11.43	0.690271
5	68.23	0.75	13.3	9.31	0.67561
6	57.86	1.33	19.18	13.426	0.683498
7	57.14	1.33	17.8	12.46	0.682143
8	55.13	0.10	15.51	16.7508	1.085513
9	44.42	0.67	15.86	10.9434	0.669139
10	39.76	0.83	17.8	12.46	0.682143
11	37.08	3.00	19.86	10.30734	0.493496
12	32.88	3.00	17.8	8.9	0.470238
13	31.09	0.58	11.3	9.2773	0.803621
14	29.16	0.40	15.4	12.32	0.786111
15	29	0.50	13.3	9.31	0.67561
16	27.88	0.50	20.95	14.665	0.684962
17	24.87	0.33	17.94	13.1859	0.719357
18	24.27	0.33	15.8	11.85	0.733108
19	21.08	0.75	11.3	7.458	0.62699
20	19.57	0.33	13.8	11.04	0.784375
21	18.32	0.17	18.87	15.30357	0.800424
22	13.72	0.83	7.79	5.53869	0.668437

Table 2. Values of a and b for different recovery times.

d_r *(hours)*	a	b	R^2
8	5.08538	3.54442	0.79008
9	5.09260	3.58195	0.79121
10	5.09932	3.61289	0.79216
11	5.10550	3.63885	0.73296
12	5.11114	3.66094	0.79364
13	5.11627	3.67997	0.79423
14	5.12095	3.69653	0.79475
15	5.12522	3.71109	0.79520
16	5.12912	3.72398	0.79561

Table 3. Mean PAI Values (Males). Unified algorithm values are for a 50000-person simulation (~1400 persons per cohort).

Age	CHAD (UNCORRECTED)	UNIFIED ALGORITHM	LITERATURE VALUE*
0	1.84	1.59	-
0.33	-	-	1.15
0.50	-	-	1.23
0.75	-	-	1.34
1	1.83	1.58	1.32
2	1.89	1.59	1.38
3	1.88	1.62	-
4	1.86	1.65	-
5	1.91	1.71	1.36
6	1.91	1.75	1.39
7	1.93	1.77	1.33
8	1.91	1.79	1.39
9	1.90	1.79	1.41
10	1.90	1.81	1.59
11	1.86	1.74	1.65
12	1.85	1.78	1.74
13	1.84	1.78	1.46
14	1.86	1.81	1.73
15	1.85	1.81	1.89
16	1.94	1.90	-
17	1.93	1.89	-

* Provided by EPA

Table 4. Mean PAI Values (Females). Values are for a 50000-person simulation (~1400 persons per cohort).

Age	CHAD (UNCORRECTED)	UNIFIED ALGORITHM	LITERATURE VALUE*
0	1.85	1.57	-
0.33	-	-	1.2
0.50	-	-	1.31
0.75	-	-	1.29
1	1.86	1.56	1.3
2	1.88	1.57	1.36
3	1.86	1.59	-
4	1.87	1.66	-
5	1.85	1.69	1.33
6	1.86	1.73	1.35
7	1.84	1.75	1.41
8	1.85	1.76	1.47
9	1.85	1.77	1.6
10	1.83	1.76	1.55
11	1.83	1.78	1.59
12	1.80	1.77	-
13	1.79	1.76	1.60
14	1.79	1.76	1.66
15	1.77	1.75	1.74
16	1.94	1.90	-
17	1.83	1.81	-

*Provided by EPA

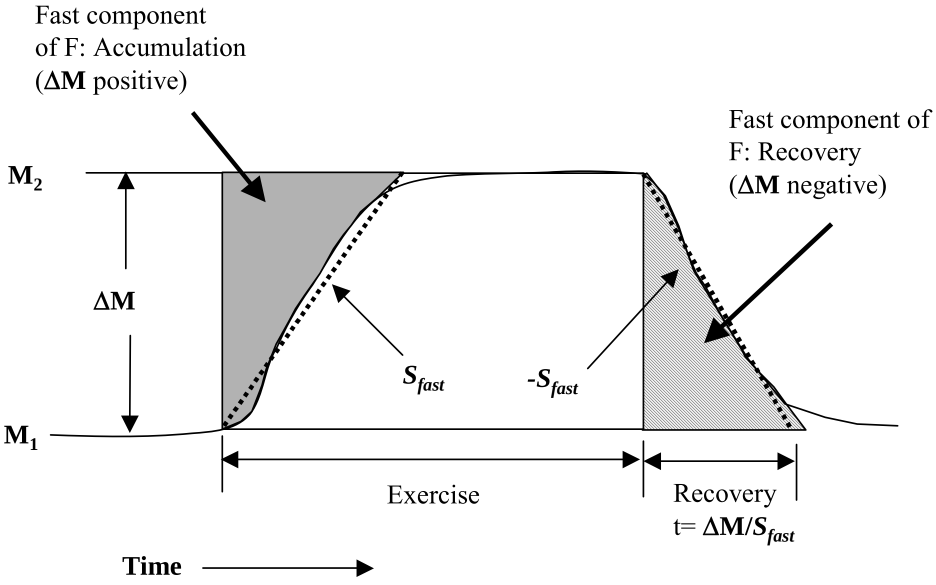


Figure 1. Fast components of oxygen deficit and recovery.

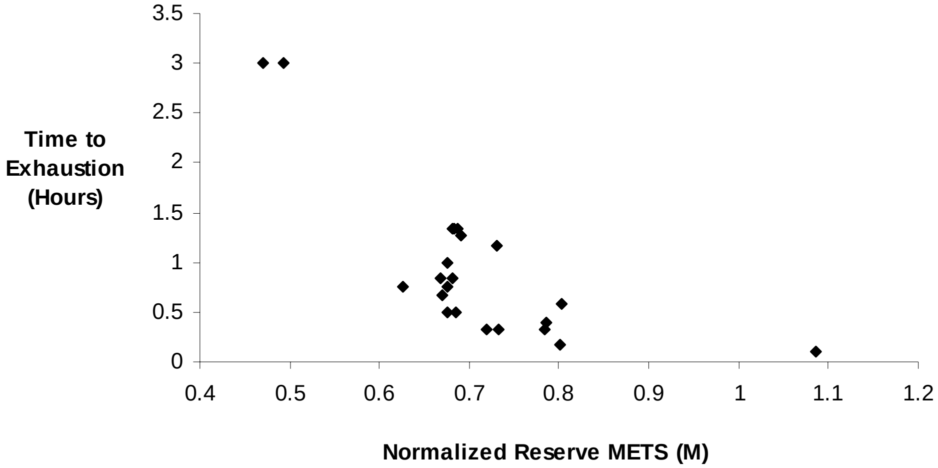


Figure 2. Exercise level in normalized reserve METS (M) versus time to reach exhaustion. Note nonlinear relationship.

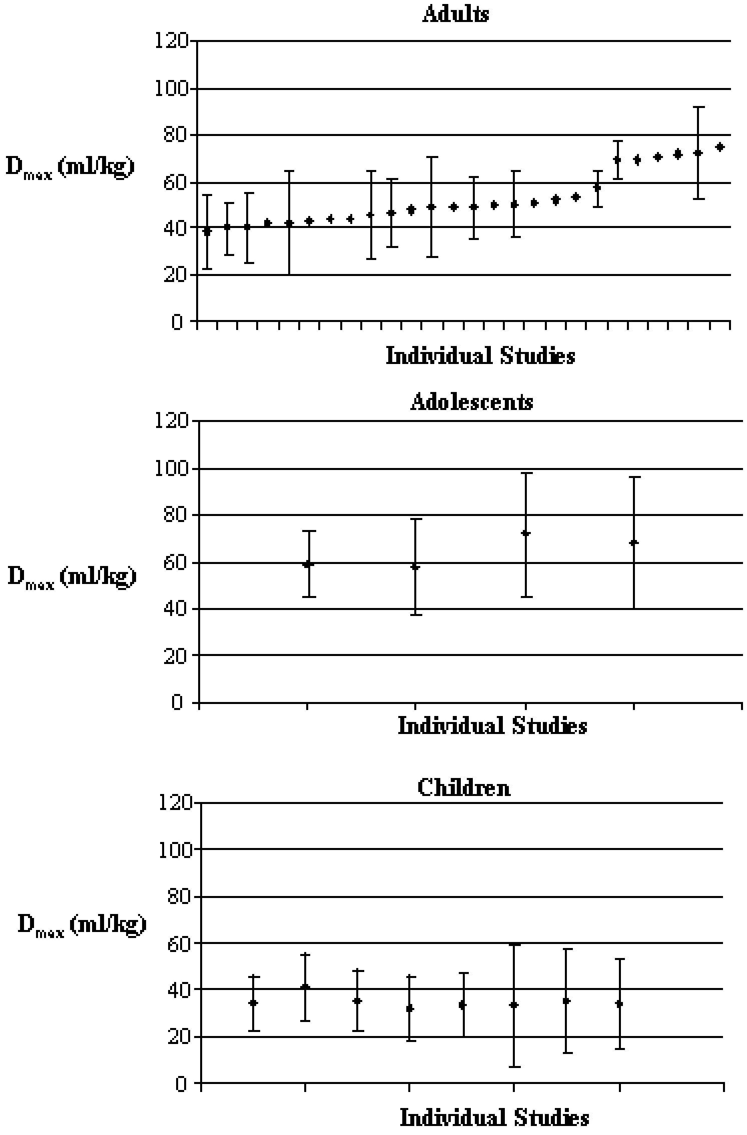


Figure 3. Values of the maximum accumulated oxygen deficit at exhaustion in adults[16-29], adolescents[30], and children[31-32]. The diamonds represent means from different exercise protocols. Bars are ± one standard deviation.

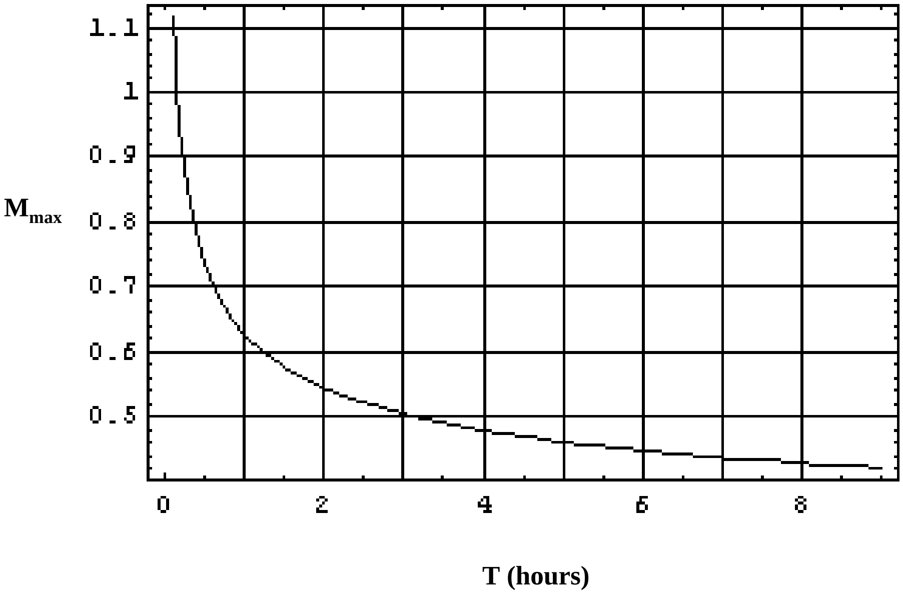


Figure 4. Maximum M (METS reserve) value that can be sustained during constant–intensity exercise of duration T.

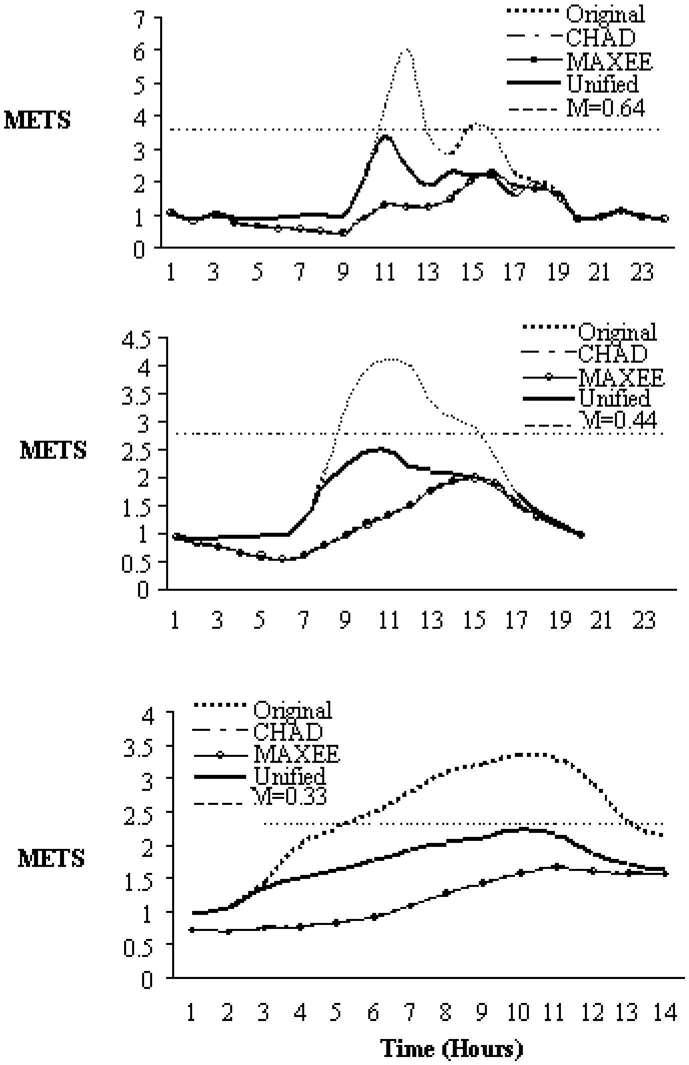


Figure 5. From the top, one, four, and nine hour running averages of METS for different METS-adjustment algorithms (for a male, age= 3 years, METSmax=5). Straight lines are the limits imposed by the Bink[43] equation.

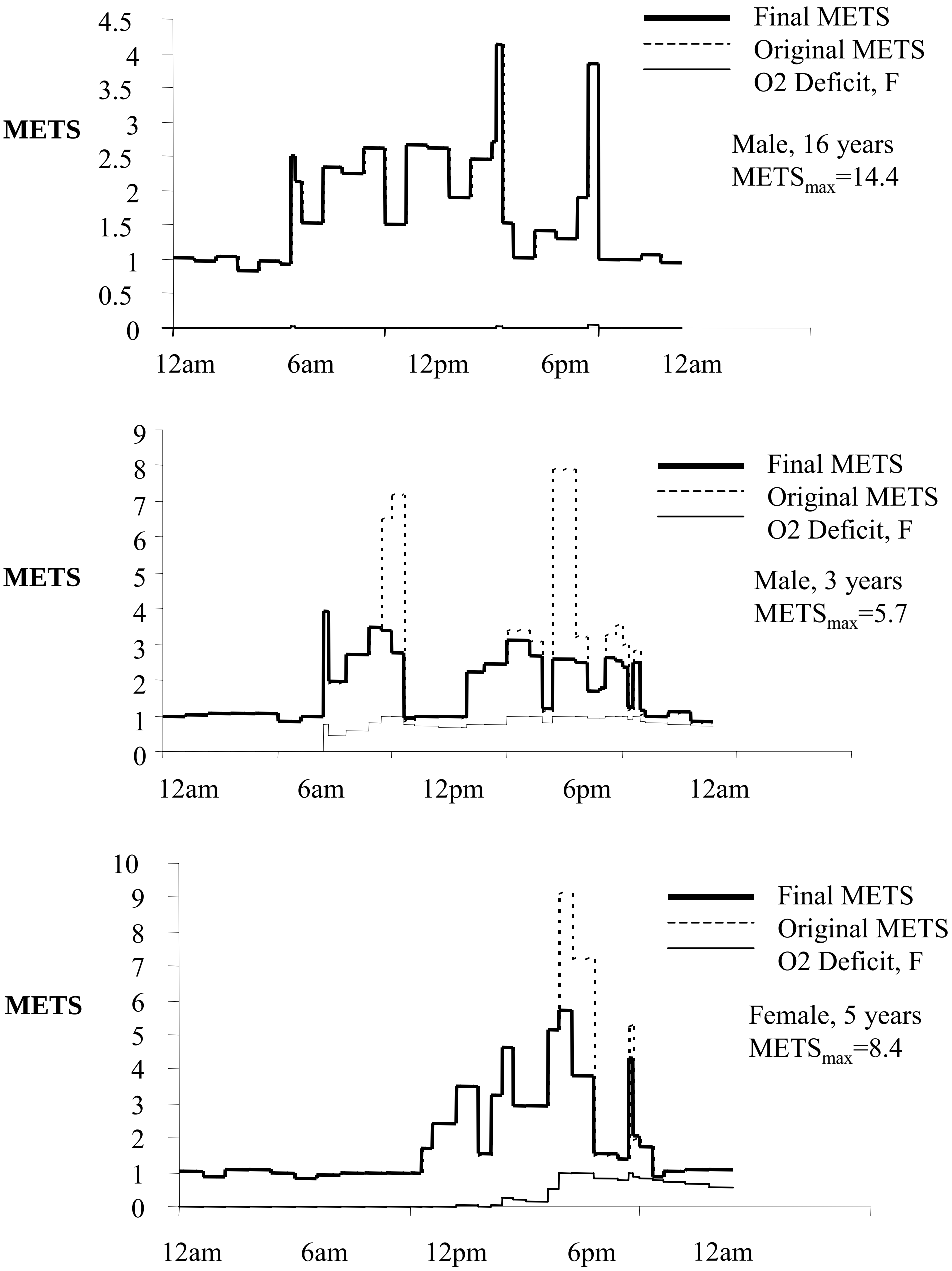


Figure 6. Three examples of original and corrected METS event time series. In the example in the top panel, the METS values pulled from the CHAD diary required no adjustment. The METS series for the individuals represented in the bottom two panels were adjusted for fatigue.

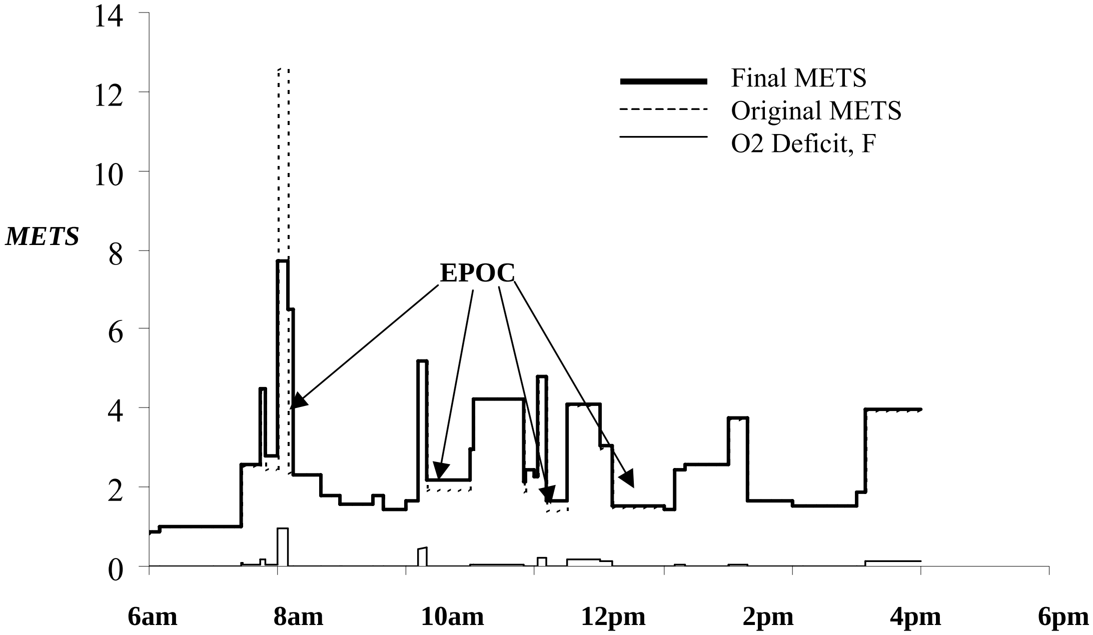


Figure 7. Event series for a 13-year-old female (METSmax=14), showing small upwards adjustments in METS for EPOC.

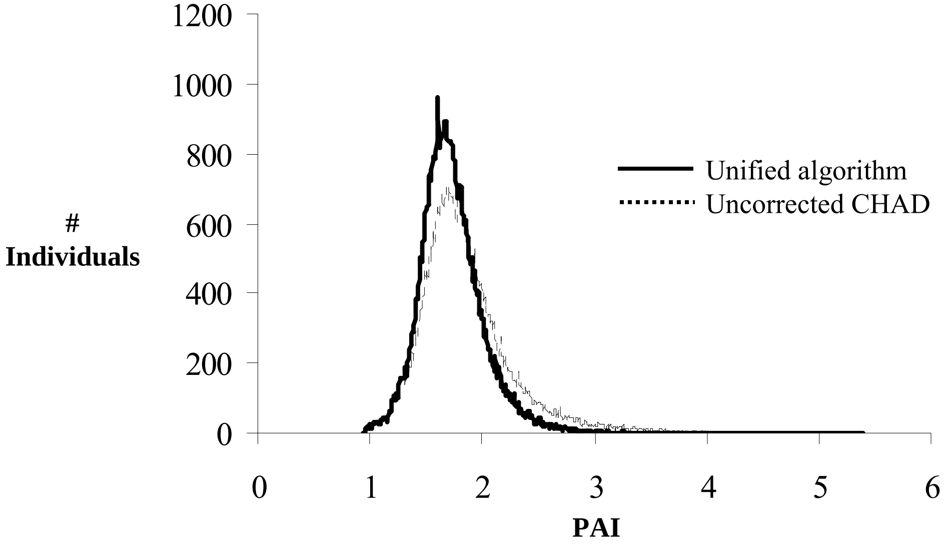


Figure 8. PAI distributions calculated using the uncorrected CHAD METS values compared with the values resulting from the unified algorithm.

Table 5. Data for Calculation of Maximum Accumulated Oxygen Deficit, D_{max}

Study	VO2Max (ml/kg-min)	SD	SE	Dmax (ml/kg)	SD	SE	age	gender
Berthoin et al. 2003	43.3	5.3		34.3	11.8		c	f
Berthoin et al. 2003	48.7	8.1		33.6	13.6		c	m
Bickham et al. 2002	64.4	6.1		43.3			a	b
Billat et al. 1996	63.2	4.2		40.1	14.9		a	f
Billat et al. 1996	77	6.4		48.9	21.3		a	m
Buck and McNaughton 1999	57.5	2.4		53.4			a	m
Carlson and Naughton 1993	43.3		1	41	14.4	2.4	c	f
Carlson and Naughton 1993	43.3		1	35	13.2	2.2	c	f
Carlson and Naughton 1993	43.3		1	32	13.8	2.3	c	f
Carlson and Naughton 1993	53.9		2.3	33	25.8	4.3	c	m
Carlson and Naughton 1993	53.9		2.3	35	22.2	3.7	c	m
Carlson and Naughton 1993	53.9		2.3	34	19.2	3.2	c	m
Doherty et al. 2000	58	4.6		69			a	m
Doherty et al. 2000	58	4.6		70.4			a	m
Doherty et al. 2000	58	4.6		71.4			a	m
Faina et al. 1997	72	4		45.9	19		a	m
Gastin et al. 1995	57	3		42			a	b
Gastin et al. 1995	57	3		43.9			a	b
Gastin et al. 1995	57	3		44.1			a	b
Gastin et al. 1995	55	3		51.2			a	b
Gastin et al. 1995	55	3		52.1			a	b
Gastin and Lawson 1994	53.1	2.1		47.6			a	m
Gastin and Lawson 1994	53.1	2.1		49			a	m
Gastin and Lawson 1994	53.1	2.1		49.6			a	m
Hill et al.1998	48.2	9.1		42	22		a	b
Maxwell and Nimmo 1996	112.2	5.2		74.6			a	m
Naughton et al. 1997	49.6		3.5	58.6	22.2	3.7	ad	f
Naughton et al. 1997	49.6		3.5	58.1	28.2	4.7	ad	f
Naughton et al. 1997	61.7		2.2	71.5	35.4	5.9	ad	m
Naughton et al. 1997	61.7		2.2	67.6	38.4	6.4	ad	m
Olesen 1992	53.5			40	11		a	b
Olesen 1992	62.5			57	8		a	b
Olesen 1992	53.5			69	8		a	b
Olesen 1992	53.5			72	20		a	b
Roberts et al. 2003	62.3	9		49.1	13		a	b
Roberts et al. 2003	62.3	9		50.5	14.1		a	b
Weber and Schneider 2000	38.5		1.8	38.2	15.6	2.6	a	f
Weber and Schneider 2000	43.4		1.5	46.3	14.4	2.4	a	m
Woolford et al. 1999	74.2	2.3		38.7	5.4		ad	b
Woolford et al. 1999	74.4	3.5		54.4	9.7		ad	b
Woolford et al. 1999	76.2	2.9		56.8	9.1		ad	b

Abbreviations

SD	=	Standard deviation
SE	=	Standard error
c	=	Children
ad	=	Adolescents
a	=	Adults
m	=	Males
f	=	Females
b	=	Both

Table 6. Data for Calculation of the Slope of the Fast EPOC Component

Study	Peak VO2	Baseline	VO2 post-EPOCfast	Duration of EPOCfast	Slope	Slope
	(ml/min)	(ml/min)	ml/min	min	ml/min/min	METS/min
Dawson et al. 1996	1900	250	450	2.5	580.00	2.320
Almuzaini et al. 1998	2500	250	425	2.75	754.55	3.018
Knuttgen 1970	2500	250	400	2.5	840.00	3.360
Short and Sedlock 1997	1800	250	575	2	612.50	2.450
Short and Sedlock 1997	1500	250	400	2	550.00	2.200
Harms et al. 1995	2976	300	399	7	368.14	1.227
Harms et al. 1995	2688	300	420	7	324.00	1.080
Trost et al. 1997	1900	250	550	4	337.50	1.350
Pivarnik and Wilkerson 1988	3300	250	900	5	480.00	1.920
Pivarnik and Wilkerson 1988	2600	250	650	5	390.00	1.560
Pivarnik and Wilkerson 1988	1650	250	520	5	226.00	0.904
Frey et al. 1993	2610	350	725	5	377.00	1.077
Frey et al. 1993	2003	350	580	5	284.60	0.813
Frey et al. 1993	1688	300	609	5	215.80	0.719
Frey et al. 1993	1373	300	493	5	176.00	0.587
Kaminsky et al. 1990	2100	220	475	2	812.50	3.693
Maresh et al. 1992	2262	312	624	5	327.60	1.050
Maresh et al. 1992	2340	312	702	5	327.60	1.050
MEAN				4.263888889	443.54	1.688
Std dev				1.605304219	204.55	0.949

This page intentionally left blank.

APPENDIX C. A NEW METHOD OF LONGITUDINAL DIARY ASSEMBLY

This page intentionally left blank.

A New Method of Longitudinal Diary Assembly

Graham Glen and Luther Smith

June, 2005 (revised October, 2005)

Alion Science and Technology, Inc.
Durham, NC 27713

prepared for WA 131

EPA Contract 68-D-00-206

Tom McCurdy
Work Assignment Contract Officer's Representative
National Exposure Research Laboratory
U. S. Environmental Protection Agency

Introduction

Exposure models like APEX and SHEDS are microenvironmental personal simulation models. The determination of exposure requires time series for both (a) microenvironmental pollutant concentrations and (b) personal time-activity patterns. To estimate longitudinal exposure patterns, it is necessary to produce a longitudinal time-activity diary for each simulated person which covers the entire simulation period. The human time-activity databases used by exposure models contain no longitudinal diaries of sufficient length. (Models are typically run for a year or more.) Various methods of assembling single-day diaries into a longitudinal pattern are currently implemented in EPA exposure models. This report describes a new method that correctly meets user-defined targets for both variance and autocorrelation.

The output from an exposure model like APEX or SHEDS consists of a set of exposure time series, one for each simulated individual. Of course, the mean exposure is important, both within an individual (the mean over time) and across individuals (the population mean). The existing diary assembly methods are good at determining these means. However, there is a growing recognition that variation in exposure is also important. One such aspect is within-person variation, which is useful for determining the frequency and intensity of high-exposure events, even for persons whose mean exposure is low. Another aspect is the between-person variance, especially in some long-term measure of exposure. For example, to assess the carcinogenic risk from pollutants that slowly accumulate in the body, the average daily dose (ADD) over a period of several years may be a useful measure of exposure. Then the distribution of risk across the population depends on the distribution of ADD. A large part of the variance in this distribution may be due to persistent differences in activities among individuals. To characterize this distribution correctly, it is necessary to have longitudinal activity diaries with persistent differences in activities between individuals, even for persons in the same age-gender cohort.

Another aspect of longitudinal diary assembly is similarity in diaries from day to day, reflecting the degree of repetitiveness in human behavior. Statistically, this can be measured by autocorrelation. The proposed method uses a one day lag. Longer lag times could be considered, but the strength of the correlation decreases rapidly with elapsed time (Xue et al., 2004; MacIntosh, 2001).

Cohorts and Diary Pools

Nearly all diary assembly methods depend on some method of cohort specification. Diaries are drawn from *cohorts,* which are population subgroups whose members have certain common characteristics. It is reasonable to expect that at least on average, people who are closely matched in age and gender (and possibly other properties such as employment status) would have activity patterns that are more similar than people of widely differing demographic status. Hence, if one were attempting to construct a longitudinal activity diary for a 30 year old working female, it is reasonable to use a set of single-day diaries belonging to (say) the cohort of working females ages 25-44. Note that the cohort cannot be defined too narrowly, or there might not be enough single-day diaries in the database to allow the proper variation in activities. This is the main reason why cohorts often consist of a range of ages, rather than a single year of age.

The creation of cohorts involves a trade-off between two factors. A narrower or smaller age range for each cohort increases the similarity between the people supplying the diaries and the target individual for whom the diary is assembled (Graham and McCurdy, 2004). However, for statistical stability it is necessary that the pool of available diaries from which the selections are made does not get too small.

Within cohorts, additional criteria for diary selection may be imposed. For example, it is often the case that diaries are matched by day of week and season, and sometimes by temperature and/or rainfall as well. The set of diaries available for possible selection on a given simulation day is called a *diary pool* or *subgroup*. In short, the term 'cohort' refers to restrictions on the universe of available diaries that apply to a given person throughout the entire simulation, whereas 'pool' refers to restrictions that apply on a particular simulation day, but may change on subsequent days. Each simulated person belongs to only one cohort, but may move through several diary pools as the simulation progresses. It is permissible for the diaries in the diary pool to have unequal selection probabilities. For example, perhaps a diary that is an exact match in age to the simulated individual is given a higher *a priori* selection probability than a diary from a person of slightly different age.

This appendix does not address the questions of cohort or pool definition. Once these definitions are given, the next step is to specify the method of selecting one-day diaries from each pool for assembly into a single longitudinal diary. This selection process should result in a 'realistic' distribution of the dominant exposure-related variable on the diaries. One of the strengths of the proposed diary assembly method is that it does not directly depend on the cohort or pool definitions; the same method (and computer code) is applicable in all cases.

Indexing the diary database by scores for a key variable

For this discussion, it is assumed that there is some measurable property of the diaries that has a dominant influence on exposure. To obtain credible exposure estimates, it is necessary to assemble longitudinal diaries that have a realistic distribution for this key property. A specific example of this key property could be the total time spent outdoors, which is currently used by the SHEDS-Wood model for assembling longitudinal activity diaries. For other pollutants the key variable might be travel time or time performing a particular activity, for example. The key or index variable could also be a composite formed from several different variables, for example, a sum or perhaps a weighted average of other variables. The necessary condition for implementing the method is that every single-day diary be assigned a numeric value for this key variable. This allows the set of available diaries in every diary pool to be ranked in terms of this key variable, from lowest to highest. While the diary assembly method does not depend on how this key variable is defined, in examples given below it is assumed (for specificity) that the key variable is outdoor time.

An important aspect of this approach is that all references to the key variable are in terms of *scores*. This means that within every pool of diaries, the individual diaries are ranked from lowest to highest in terms of the key variable and assigned a score which indicates their place in the list. This score is bounded between zero and one. If there are K diaries in a pool, and each diary has equal statistical weight, then the score for the diary at rank R is

$$\text{score} = (R - 1/2) / K \qquad (1)$$

Similarly, when individuals are being ranked within a group of P persons, then the score for the person at rank R in the group is

$$\text{score} = (R-1/2) / P. \qquad (2)$$

The scores are useful for several reasons. First, the distributional properties are known, whereas the distributional properties of the key variable itself would depend on its definition and, furthermore, might well vary from cohort to cohort and from pool to pool. Knowing the distributional properties allows the specification of methods that target certain statistics. Second, the score reflects the behavior of an individual relative to their peer group (for example, a score of 0.75 means that the person ranks above 75% of the people in the same cohort and pool, in terms of the key variable). Third, scores can be moved across diary pools, whereas absolute values for the key variable might not. For example, there might be a diary with six hours of outdoor time in the Sunday pool, but no such diary in the Monday pool. But a score of 0.75 has meaning on all days and can be mapped to a specific diary. The ability to move scores across day types is important in the autocorrelation matching, as described below. Fourth, the use of scores helps in ensuring that all the available diaries are collectively sampled with the correct frequency.

Note that the use of ranks or scores does not preclude the ability to return to the original key variable. In terms of diary assembly, it is necessary to specify which diary should be used on a given simulation day. For this purpose, requesting the available diary nearest to score 0.38 is no different than requesting the available diary nearest to (say) 73 minutes of outdoor time. Once the diary is chosen, the exact value of the key variable on that diary can be recovered.

The statistics D and A

For this assessment, two statistics are used. The first is called ‘D’, which measures long-term differences between persons in the same cohort. The second is called ‘A’; it is the mean across persons of the daily autocorrelation coefficient of the scores. Detailed mathematical properties of D and A are given in the appendix. Both D and A are collective properties of a group of persons. To calculate them, a time series for the key variable is needed for each person. There may be some gaps or missing values in the time series, but to calculate D it is necessary that there is substantial temporal overlap between persons, as each person is ranked relative to the others on each day.

The following discussion of how D and A are calculated assumes that a longitudinal diary is available for each individual. The discussion of how one constructs longitudinal diaries that collectively have desired values of D and A for model simulation runs comes later.

D is calculated as follows. For each day, rank each person relative to their cohort and use equation (2) to generate a score. Here P may possibly vary from day to day; it is the number of persons with non-missing values on each day. The underlying assumption is that the sample on any given day is representative, so that a score of 0.38 would mean that the person ranks above about 38% of all the other persons in the cohort, even if only part of the cohort was sampled on that particular day. Days with a very small diary pool should therefore be excluded from the analysis.

This yields a time series of daily scores for each person. Find the mean and variance of the scores for each person over their time series. The overall within-person variance σ_w^2 in the group is the mean of these individual time variances. The between-person variance σ_b^2 is the variance across persons in the mean scores for each time series. The statistic D is then given by

$$D = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2). \qquad (3)$$

Since both variances must be non-negative, it is clear that D is a proper fraction, bounded between zero and one. D=0 means that σ_b^2 is zero, or that each person has the same mean score. A small D means that σ_b^2 is substantially smaller than σ_w^2. A D near one means that σ_b^2 is much larger than σ_w^2, or that each person shows little variation over time relative to the variability between persons.

The criteria for defining cohorts and diary pools are determined by the user, and the proposed method places no restrictions on these criteria. However, the calculation of D can provide a useful indicator of whether cohorts have been reasonably defined. A large value for D indicates large variability in long-term behavior between the individuals, and this is contradictory to the concept of cohorts.

The autocorrelation A is even simpler to calculate than D, because each time series can be examined independently. The first step is to determine the score for each day, relative to the entire time series. If there are J days in the time series, and a given day is at rank R in terms of the rank for the key variable among the J days, then the score for that day is (R-1/2) / J. The overall mean and variance in these scores for the time series is then calculated. However, due to the properties of the discrete uniform distribution of the scores (neglecting tied scores), the mean must be 1/2 and the variance is $(1/12)\ (1-1/J^2)$, which is very close to 1/12 for J large. The lag-one covariance is also determined; it is (1/J) times the sum of the paired products (score(j)-1/2)*(score(j+1)-1/2), where score(j) is the score on day 'j' (see, for example, Box et al., 1994). The lag-one autocorrelation for the time series is given by the ratio of the covariance to the variance. This calculation is repeated for each time series, and the statistic A is the mean of these individual autocorrelations. The statistic A has a range from -1 to +1, with positive values indicating that each day has a tendency to resemble the day before. Random selection of diaries from day to day produces A values near zero. Negative A values imply dissimilarity between consecutive days.

A study of children conducted in Southern California (see Xue et al. 2004) provides about 60 days of data on each of 163 children. The time series are not continuous, as the monitoring consisted of twelve six-day periods, one per month over a year. Furthermore, only about 40 children were measured simultaneously, as the other children were sampled in different weeks. However, a sample size of 40 is sufficient to calculate reliable rankings across persons. The number of consecutive day pairs was substantially less than the number of days, due to the gaps in the time series. However, D and A statistics were calculated for three variables directly recorded on the activity diaries (outdoor time, travel time, and indoor time), and also for a fourth variable, the physical activity index or PAI (McCurdy et al., 2000). The analyses were performed for all children together and for two gender cohorts. The separation into two cohorts reduces the number of children measured simultaneously to fewer than 20. Further division into more cohorts is therefore not practical, as the reliability of the scores would decrease. The results for these analyses are given in Table 1.

Table 1: D and A statistics derived from the Southern California study data

Variable	Group	D	A
outdoor time	all	0.19	0.22
outdoor time	boys	0.21	0.21
outdoor time	girls	0.15	0.24
travel time	all	0.18	0.07
travel time	boys	0.18	0.05
travel time	girls	0.18	0.08
indoor time	all	0.17	0.22
indoor time	boys	0.21	0.20
indoor time	girls	0.17	0.24
physical activity	all	0.16	0.23
physical activity	boys	0.16	0.20
physical activity	girls	0.16	0.25

Here 'physical activity' is measured by PAI, which is the ratio of total energy expenditure per day to the basal metabolic energy expenditure per day, estimated from the diary times. For all variables and each group, the standard deviation between persons for autocorrelation was about 0.20, and the standard error in the mean A was about 0.02. Table 1 indicates that gender differences for both D and A are small, if present at all.

It should be noted that the variables in Table 1 are not really independent. The sum of the three time variables equals 24 hours in all cases. Furthermore, PAI is derived from the same three times, so part of the similarity across variables is due to these relationships.

Generating longitudinal diaries

Exposure models like APEX and SHEDS construct a number of 'simulated individuals', whose demographic characteristics are intended to be representative of the target population. A longitudinal activity diary is constructed for each such person; it is to be hoped that the collective properties of these diaries are also representative of the target population, or at least the distribution of the key variable affecting exposure. As mentioned earlier, the new diary assembly method does not impose any constraints on the methods of constructing cohorts and diary pools, so it is up to the modeler to ensure that these are defined appropriately. The new method just ensures that the selections from these pools match the requested targets for D (variance ratio) and A (autocorrelation). The target values for D and A are supplied by the modeler.

First, construct a beta distribution (with parameters as specified in the Supplement) for the distribution of personal mean scores. For each simulated person, first select a mean target score T at random from this beta distribution. Then, for each individual, construct another beta distribution with mean equal to T. From this second beta distribution, pick a set of independent random values containing approximately 3% more numbers than there are days in the simulation period. Call this the set of X-scores and let K be the number of scores selected. At this point, one has P sets of X-scores, each containing K values.

The second part of the process is to generate the requested autocorrelation by reordering the collection of selected values. First, choose a target autocorrelation for each individual. This is selected from a beta distribution with a mean of A. For each individual, the set of X-scores are ranked from lowest to highest. For the first simulation day, choose any X-score at random. For each subsequent day, construct a new beta distribution (the parameters of the beta depend on A and the selected value for the prior day, as detailed in the Supplement), and pick one value Y from it. Find the nearest X-score (in rank) to K*Y that has not already been assigned to a prior day in the time series. Continue this process until all simulation days are assigned values. The reason for the extra values is that without them, the last few days of the simulation would have very few choices left, and this lack of freedom would inhibit meeting the requested autocorrelation.

The result of these steps is a vector of X-scores, one value per simulation day, for each person. It remains to now associate a diary with each X-score. Recall that the user has specified the appropriate diary pool for each simulation day. The diaries in the pool are assigned a cumulative probability distribution as follows. First, they are sorted by the key variable. Then a selection probability is assigned to each diary as determined by the diary pool structure (for many models, equal probabilities are used).

The following example illustrates how a diary is assigned to an X-score. Suppose the pool for a particular day had only four diaries, with probabilities in sorted order of 12%, 33%, 41%, and 14% of being used. The cumulative probability vector is then (0.12, 0.45,0.86,1.00). The X-score assigned to this day is then used to determine which diary is selected. If the X-score is lower than 0.12 then the diary ranked lowest on the key variable is chosen. If X is between 0.12 and 0.45 the second lowest diary is picked. For X between 0.45 and 0.86 the next highest diary is used. Finally, if X is greater than 0.86 then the diary ranked highest on the key variable is selected. This process is repeated for each day of the simulation period.

Results

The following tables present some results obtained using the new method. Tables 2 and 3 present comparisons of D and A statistics, respectively, calculated both from ranks and from key variable values, for both the Southern California data and simulations using the new method. Table 4 displays the performance of the new method over the full range of both D and A. Table 5 shows the performance of the proposed method for different simulation lengths, for a variety of D and A values.

Table 2. Computation of the D statistic calculated from ranks and key variable values, both directly from the southern California study and from simulations using the proposed method. Simulations constructed 20,000 longitudinal diaries for periods of forty-eight days.

Key variable	Group	Ranks		Values	
		Study	Simulation	Study	Simulation
outdoor time	all	.19	.19	.12	.14
outdoor time	girls	.15	.15	.11	.11
outdoor time	boys	.21	.21	.17	.18
travel time	all	.18	.18	.10	.13
travel time	girls	.18	.18	.10	.13
travel time	boys	.18	.18	.12	.14
indoor time	all	.17	.17	.12	.14
indoor time	girls	.17	.17	.11	.13
indoor time	boys	.21	.21	.16	.17
PAI	all	.16	.16	.12	.13
PAI	girls	.16	.16	.13	.12
PAI	boys	.16	.16	.13	.14

Table 3. Computation of the A statistic calculated from ranks and key variable values, both directly from the southern California study and from simulations using the proposed method. Simulations constructed 20,000 longitudinal diaries for periods of forty-eight days.

Key variable	Group	Ranks		Values	
		Study	Simulation	Study	Simulation
outdoor time	all	.22	.21	.24	.19
outdoor time	girls	.24	.23	.26	.20
outdoor time	boys	.21	.20	.21	.19
travel time	all	.07	.07	.06	.06
travel time	girls	.08	.08	.07	.06
travel time	boys	.05	.06	.04	.05
indoor time	all	.22	.21	.23	.19
indoor time	girls	.24	.23	.26	.20
indoor time	boys	.20	.19	.19	.18
PAI	all	.23	.22	.26	.19
PAI	girls	.25	.24	.29	.21
PAI	boys	.20	.20	.23	.17

Table 4. Performance of proposed method in hitting targeted values at selected points across the ranges of the D and A statistics.

Requested		Obtained	
D	A	D	A
0	0	.00	.00
0	.50	.01	.50
0	.99	.03	.99
0	-.50	.00	-.49
0	-.99	.00	-.99
.50	0	.51	.01
.50	.50	.51	.50
.50	.99	.53	.99
.50	-.50	.50	-.49
.50	-.99	.51	-.99
.99	0	.99	.01
.99	.50	.99	.50
.99	.99	.99	.99
.99	-.50	.99	-.49
.99	-.99	.99	-.99

Table 5. Performance of proposed method in hitting targeted values of the D and A statistics over different lengths of the simulation period. The values of D=.19 and A=.22 are the values for outdoor time obtained from the southern California study.

Simulation period length	Requested		Obtained	
	D	A	D	A
30 days	.19	.22	.20	.24
90 days	.19	.22	.20	.24
1 year	.19	.22	.20	.22
30 days	.10	.40	.11	.40
90 days	.10	.40	.10	.41
1 year	.10	.40	.10	.40
30 days	.40	.10	.41	.13
90 days	.40	.10	.41	.12
1 year	.40	.10	.41	.10
30 days	.81	-.22	.81	-.17
90 days	.81	-.22	.81	-.20
1 year	.81	-.22	.82	-.21

Discussion

1) Use of ranks rather than the original key variable
2) Use of beta distributions rather than other forms
3) Ensuring no sampling bias within diary pools
4) Performance over full range of D and A values
5) Performance of simulations of various lengths
6) Varying targets for D and/or A within a simulation
7) Movement of X-scores across day-types
8) The frequency distribution for relatively rare diary events
9) Ease of use

1) *Use of ranks rather than the original key variable*
The new method makes use of rankings of the key variable in computing D and A statistics and in the generation of X-scores, rather than using the original values of the key variable. This provides both a modeling advantage and a mathematical advantage. The modeling advantage is that it permits the maintenance of persistent differences while allowing a natural transition across diary pools. A person with a mean or target X-score of T has a tendency for a higher value for the key variable than a fraction T of his/her peer group. In the absence of information to the contrary, it is reasonable to suppose that this tendency would persist. If the key variable is outdoor time, on cold and rainy days the entire group may spend less time outdoors, but this does not suggest that the relative position of individuals within the group would change. Once the diaries are assembled, most persons will show drops in outdoor time on such days due to the change in the diary pool, even though the X-scores themselves do not drop on such days. This combination of maintaining persistent differences between individuals while allowing the diary pools to define the distribution of the key variable would be very difficult to attain using the original (non-ranked) variable.

The mathematical argument for using ranks is that the method becomes much more general, since the distribution of ranks does not depend on the choice of the key variable, or on the definition of cohorts, diary pools, or day-types. By contrast, the development of a parametric method that tried to match statistics on the original values of the key variable would have to depend on characterizing the distribution of that variable for the specific application of the model. For some variables like outdoor time, the distribution has a relatively low mean and positive skewness (a long tail to the right), but for indoor time the mean is high and the distribution is negatively skewed. Furthermore, the distribution would depend on the specific definition of the cohort, and would change as well with day-type and season. It would also be likely to change when going from one geographic region to another. Every time the distribution changes, the mathematical algorithms would have to change to reproduce the given distribution while simultaneously meeting targets for both variability (here represented by D) and episodic behavioral tendencies (here represented by A). The complexity of such approaches would add both a computational burden and a quality assurance burden to the exposure model.

The performance of the proposed method was numerically evaluated against measured key variable values using data from the southern California study (see Tables 2 and 3). Note that the protocol for this study did not match the assumptions used in developing this method; in particular, different children reported diaries on different days, and each child had breaks in their time series. The new method was applied to three different key variables (outdoor time, travel time, and indoor time), each with two cohort groupings (all children together, and separated by gender). Synthetic longitudinal diaries were constructed from the single-day diaries reported during this study. Both D and A statistics were calculated for the study and for the synthetic diaries, using both the ranks and the key variable values.

The D statistics on rankings were essentially the same for the original diaries and the synthetic diaries. The D statistics on ranks were consistently higher than those on key variable values (average D on ranks ~ 0.18, average D on key variable ~ 0.12). This is consistent with the observation in the physical activity literature that people have more fixed tendencies in terms of rankings than in the original variable (Anderssen et al., 1996; Kelder et al.,1994; Schwab et al., 1992). However, this may not apply universally to all variables (DeBourdeauhuij et al.,2002).

More within-person consistency translates to less within-person variance for the rankings than for the original variable. By the form of the definition of D, this implies higher values for D for the rankings. This effect is evident in Table 2 for the four variables considered there. For D calculated on key variable values, the synthetic diaries (average D ~ 0.14) tended to exceed the study (average D ~ 0.12) by only a small amount.

Autocorrelations in the key variable values proved to be close to the autocorrelations in ranks, for both the study and for the simulated diaries. The simulated diaries were consistently close in A to the study when measured using ranks. Using the key variable values to calculate A, the synthetic diaries tended to be lower than the study (differences ranging from 0.01 to 0.07), except when the key variable was time spent in travel.

2) *Use of beta distributions*

All of the random number generation in the new method involves drawing numbers from beta distributions. This is convenient though not strictly necessary. All of the random number distributions are bounded both above and below, which is a natural property of the beta distribution. For instance, it would be quite feasible to select personal targets for autocorrelation that were normally distributed about the overall population mean A, but since autocorrelation is bounded between -1 and +1, it would then be necessary to truncate the normal on both ends. Most programming languages have built-in beta distribution functions, and for the ones that do not (like Fortran), there are a number of well-tested algorithms developed for this purpose. Alternate distributions for generating the X-scores have been tested, for example a two-level uniform (one probability inside a given sub-interval and a different probability elsewhere) has been successfully used for this purpose.

Given fixed end points, the beta distribution has two shape parameters which allow a great variety of forms. Both shape parameters must be positive. If both parameters exceed one, then the distribution has the ‘usual’ form of a central peak, monotonically decreasing on either side until reaching the bounds. The location and width of this peak can be targeted separately, which is convenient for targeting both a mean and a variance. If the parameters fall on different sides of unity, then the distribution is monotonic over its entire range (either increasing or decreasing), often called a J-shaped beta. If both parameters are less than one, a U-shaped distribution results, with peak probability at each end. Such U-shaped distributions are never used for X-scores or diary reordering, but may be used to assign individual targets T. A beta distribution with both shape parameters equal to one is a uniform distribution. In fact, if D=0 is requested, then all the X-scores are chosen from such uniform beta distributions, and all persons have a common target mean of T=0.5. If D is set to one, then the targets T have a uniform distribution, but the X-scores all become equal to T since the beta for them narrows to zero variance. In practice this would lead to numerical difficulties, so in implementation the code would usually contain a restriction that D<0.99 (or some similar bound). If a simulated person is given a target autocorrelation of zero, then the beta distributions used to order the X-scores all reduce to uniform distributions.

3) *Ensuring no sampling bias within dairy pools*
If a given pool of one-day diaries is believed to be representative for a given cohort on a given day, then to avoid any bias it is necessary that over a large population of simulated individuals, all the diaries in the pool be used about equally often. That is, the mean and variance of any variable on that day for the group of simulated individuals should match the mean and variance seen in the diary pool itself, since the pool is supposed to be representative of the real population. This is most easily achieved by the simple expedient of uniformly sampling from the diary pool.

In the new method, the selection probabilities from the diary pool are not uniform for one individual; they tend to be higher for diaries near to the target score T than for ones further away. To avoid overall biases, it is necessary that the mixture of all the personal betas over a large group of persons be very close to uniform. So that, for example, a person who preferentially samples diaries at the low end of the rankings should be balanced by a person who preferentially samples the upper end. An important constraint on the beta distributions used for the T scores and the X-scores, is that the overall distribution of X-scores over a large simulated population should be close to uniform. In general, exact uniformity cannot be achieved by mixing betas; some particular X-scores may remain oversampled or undersampled by about 2% relative to others. However, it is possible to arrange these effects so that both the mean and variance of the beta mixture match the mean and variance of a uniform distribution, which ensures that the mean and variance of the key variable on the diaries is the same for the group of simulated individuals as for the diary pool itself (in the limit of a very large number of individuals), on each simulation day. See the Supplement for details.

4) *Performance over the full range of D and A*
The D statistic is bounded between zero and one, and A is bounded between minus one and one. There are no restrictions on D and A together; any A may be used with any D. The limiting values on both parameters imply total order, which is incompatible with the concept of a stochastic simulation. Furthermore, there is a minimum possible value for D that depends on the simulation length; for a simulation of J days, D cannot be below 1/J.

Table 4 presents results at selected points over the full range of both D and A, using the new method. The values of D and A achieved with this new method agree with the target values within 0.02 in nearly all cases, and within 0.01 much of the time. Thus, if D is requested to be 0.25, it will nearly always fall between 0.23 and 0.27 for any sizeable simulation. The same holds for A, at least for A values greater than -0.5. Large negative A targets do not match quite as well, unless a correction factor is included in the algorithm. Such a correction can be implemented fairly easily, but in practice should not be necessary since such large negative A values are not normally seen in human behavior patterns.

Some other small but reproducible effects may be seen. For example, if a very large and positive autocorrelation is requested, it is achieved but the target D statistic becomes slightly larger than requested (by about 0.02 for A=1). This effect is negligible for A<0.5, which means it is unlikely to be an issue for human behavior simulations. If it were deemed to be important, one could compensate for it by suppressing the target D value in such cases.

5) *Performance over various simulation lengths*
The method has been tested successfully over a wide range of simulation lengths, ranging from a few days up to six years. Table 5 presents some results from these simulations. For all lengths over 30 days, the match for both D and A is very good. For very short simulations, it is difficult to precisely target these statistics. For one thing, the sample mean of the X-scores for any individual does not necessarily come close to the target mean score T, when only a few scores are drawn. For another, it is very difficult to target particular autocorrelations merely by rearranging the order of the values. In fact, for three data points the autocorrelation cannot be positive, no matter what their values or how they are rearranged. For any simulation below one week in length, the autocorrelation step is nearly irrelevant, although there is no harm in allowing it to rearrange the scores. For long simulations the performance is always good, with D and A extremely close to the target values for simulations of six years in length.

6) *Varying targets for D and/or A within a simulation*
In certain applications the user might wish to vary D or A over time. For example, different day-types might each have their own targets, or perhaps D or A might change with seasons or with age over a long simulation period. The new method is easily extended to such a situation. Basically, the method would be applied separately to each set of days with a distinct D. For each set, define a distribution of target T scores, pick one for each person (keeping the percentile the same for all sets), and pick enough X-scores for the given set of days. The reordering would be done within each distinct set of days, to prevent mixing X-scores from different distributions. The final time series would then merge the vectors for the various sets of days, according to the calendar sequence.

The implementation of multiple targets for A is extremely easy. A new beta distribution is required every day since its parameters depend on both the target autocorrelation and on the rank of the X-score assigned to the prior day, and this latter quantity changes every day. Instead of supplying the target autocorrelation as a scalar, use a vector indexed by the day number, and use A(j) everywhere that A is currently used.

While it is not difficult to vary D and/or A by day-type, there is no evidence in the southern California study data that this effect is significant. Therefore, for simplicity, the basic explanation of the new method does not include this possibility directly. However, nothing is fundamentally different if these extensions are used.

7) *Movement of X-scores across day-types*
The basic method does not distinguish differing D and A targets for differing day-types, as discussed in the previous subsection. But even if A depended on day-type, the X-scores could be moved freely across day-types during the reordering step, as long as D and T did not change. This is because the X-scores are independently randomly sampled, and as long as the distribution remains the same, the scores can be interchanged.

As discussed in subsection (1) above, this is one of the advantages in using X-scores that are based on relative rankings rather than employing the original variable. The same distribution of rankings exists on all day-types, although the distribution of the original key variable will differ across day-types (if it did not, there is no reason to separate the day-types). The proposed method recognizes this difference through the differing diary pools. For example, an X score of 0.25 may correspond to 40 minutes of outdoor time on a weekday, but correspond to 70 minutes of outdoor time on a weekend. The reason why the reordering has an overall null effect on the mean and variance of the key variable is that it is just as likely for an X score of 0.25 to be shifted from a weekday to a weekend as vice versa. Therefore, over a large enough sample of persons, the distribution of X-scores before reordering and after reordering are indistinguishable.

8) *The frequency distribution for relatively rare diary events*
One concern with many of the existing longitudinal diary assembly methods currently used in exposure models is that they limit the within-person variance (and thereby induce behavioral habits) by selecting relatively few different one-day diaries for each simulated individual. This leads to the forced re-use of each of the selected diaries many times. Thus, a model that selects only eight diaries to represent one year must use each diary an average of 45 times. For such methods, each particular kind of diary event will occur with the correct overall frequency in the population as a whole, but the frequency within individuals is highly distorted.

As an example, suppose the pollutant of concern is ozone. The combination of high breathing ventilation rate, outdoor activity, and warm daytime conditions will lead to high ozone exposure. Then a relatively rare event like a long distance run (for example, a marathon) is significant to the exposure model. Under the model where only eight diaries are used, if a long distance run occurs at all (which is not likely), it occurs every time the diary is reused. This leads to a situation where the vast majority of the population have no such events, and a small number have (say) 45 such events packed into one year (or even one season), with no one having only a few such events.

With the new method, if the diary pool contains one diary with a long distance run (and hence much outdoor time on that day), this diary might be selected not at all or perhaps once, for a person whose target T has little outdoor time. For persons with larger T , this diary might be chosen a handful of times in a year. For a person whose target T matches this diary closely, it might be picked a couple of dozen times. The point is that the population has a quasi-continuous frequency distribution for this event, rather than a discontinuous one (having it occur either never or at least 45 times). Thus, the proposed method better reproduces the variance in exposure across the population.

9) *Ease of use*
The proposed method places a minimal burden on the user in terms of required input. Beyond the definitions of cohorts and diary pools, which are always required (either as user input or hard-coded into the model), the new method only requires the designation of the key variable and the targets for D and A. The various beta distributions are constructed by the model code from these inputs without further user intervention.

Summary

The new method is very flexible and succeeds in reproducing target D and A values over the entire possible range, for any choice of key variable. The D statistic of diaries assembled by the new method is independent of the length of the simulation, unlike most existing diary assembly methods. The new method avoids forced repetitions of the same activity diary from one day to another, and therefore allows for some events to occur uniquely or rarely on a given longitudinal diary. It imposes as much habitual behavior as is requested through the D and A statistics, no more and no less. The method is relatively simple to implement in computer models, requiring the ability to sort lists and to draw random numbers from beta distributions. A great advantage over many other methods is that the computer code for generating the vectors of X-scores does not depend on the choice of cohorts or diary pools.

References

Anderssen, N., D. R. Jacobs, S. Sidney, D. E. Bild, B. Sternfeld, M. L. Slattery, and P. Hannan, 1996, "Change and Secular Trends in Physical Activity Patterns in Young Adults", Amer. J. Epidemiology 143:351-362.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel, 1994, *Time Series Analysis: Forecasting and Control*, Prentice Hall, Englewood Cliffs, NJ, 598 pp.

DeBourdeaudhuij, I., J. Sallis, and C. Vandelanotte, 2002, "Tracking an Explanation of Physical Activity in Young Adults over a Seven Year Period", *Res. Q. Exer. Sport* 73:376-385.

Graham, S. E. and T. McCurdy, 2004, "Developing Meaningful Cohorts for Human Exposure Models", *J. Exposure Anal. Environ. Epidem.* 14: 23-43.

Hogg, R. V. and A. T. Craig, 1995, *Introduction to Mathematical Statistics*, Prentice Hall, Englewood Cliffs, NJ, 564 pp.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1994, *Continuous Univariate Distributions - 2* , John Wiley and Sons, New York, NY, 306 pp.

Kelder, S. H., C. L. Perry, K.-I. Klepp, and L. L. Lyttle, 1994, "Longitudinal Tracking of Adolescent Smoking, Physical Activity, and Food Choice Behaviors", *Amer. J. Public Health* 84:1121-1126.

MacIntosh, D., 2001, "Refinements to EPA/NERL's Aggregate SHEDS-Pesticides Model", EPA Contract OD-5537-NTTX, Research Triangle Park, North Carolina.

McCurdy, T., G. Glen, L. Smith, and Y. Lakkadi, 2000, “The National Exposure Research Laboratory’s Consolidated Human Activity Database” *J. Exposure Anal. Environ. Epidem.* 10: 566-578.

Schwab, M., A. McDermott, and J. Spengler, 1992, "Using Longitudinal Data to Understand Children's Activity Patterns in an Exposure Context" *Environ. Intern.* 18:173-189.

Xue, J., T. McCurdy, J. Spengler, and H. Özkaynak, 2004, "Understanding Variability in the Time Spent in Selected Locations for 7-12 year old Children" *J. Exposure Anal. Environ. Epidem.* 14 : 222-233.

SUPPLEMENT

1) Statistical properties of longitudinal diaries

Consider a set of longitudinal diaries for P persons, each diary covering the same J days. For this analysis we will assume that there are no time gaps, so that all days are consecutive. Let ‘j’ be an index that runs over simulation days, and let ‘i’ be an index that runs over persons. Consider just one variable and one cohort of persons, so all persons share the same pool of available diaries on any given day. Let t_{ij} be the value of the variable of interest on day ‘j’ of the longitudinal diary for person ‘i’. Note that in this analysis, variance calculations use division by the number of data points, without the convention of subtracting one to account for degrees of freedom (Hogg and Craig (1995), Box et al. (1994)).

Let μ_i be the average value for the given variable for person ‘i’, so for i=1,...,P we have

$$\mu_i = (1/J) \sum_{j=1}^{J} t_{ij} \qquad (1\text{-}1)$$

where J is the number of days in the simulation. There is also an intra-personal (within-person) variance for ‘t’ which may differ from one person to another:

$$\sigma_i^2 = (1/J) \sum_{j=1}^{J} (t_{ij} - \mu_i)^2 = (1/J) \sum_{j=1}^{J} t_{ij}^2 - \mu_i^2 \qquad (1\text{-}2)$$

For convenience, define V^2 as

$$V^2 = 1/(JP) \sum_{i=1}^{P} \sum_{j=1}^{J} t_{ij}^2 . \qquad (1\text{-}3)$$

The mean for the variable μ_i over all persons is given by

$$\mu = (1/P) \sum_{i=1}^{P} \mu_i = 1/(JP) \sum_{i=1}^{P} \sum_{j=1}^{J} t_{ij} \qquad (1\text{-}4)$$

which is also the mean of all the t_{ij}. The total variance of the t_{ij} is given by

$$\sigma^2 = 1/(JP) \sum_{i=1}^{P} \sum_{j=1}^{J} (t_{ij} - \mu)^2 = 1/(JP) \sum_{i=1}^{P} \sum_{j=1}^{J} t_{ij}^2 - \mu^2 = V^2 - \mu^2 \qquad (1\text{-}5)$$

The mean of the intra-personal variances across all persons is given by

$$\sigma_w^2 = (1/P) \sum^{P} \sigma_i^2 = 1/(JP) \sum^{P} \sum^{J} (t_{ij} - \mu_i)^2 = V^2 - (1/P) \sum^{P} \mu_i^2 \qquad (1\text{-}6)$$

$$i=1 \qquad i=1 \;\; j=1 \qquad i=1$$

where the subscript ‘w’ stands for ‘within-person’. There is also an inter-person (between person) variance, which is the variance in the personal means μ_i

$$\sigma_b^2 = (1/P) \sum_{i=1}^{P} (\mu_i - \mu)^2 = (1/P) \sum_{i=1}^{P} \mu_i^2 - \mu^2 \tag{1-7}$$

where ‘b’ stands for ‘between-persons’. In brief, σ_w^2 is the mean of the intra-personal variances, while σ_b^2 is the variance of the intra-personal means. An important result is that

$$\sigma_w^2 + \sigma_b^2 = V^2 - \mu^2 = \sigma^2 \tag{1-8}$$

which follows from the three prior equations. Thus, for a given set of longitudinal diaries, σ_w^2, σ_b^2 and σ^2 are tied together by equation (1-8). This has important implications when targeting variance. For a given set of diary pools from which the longitudinal dairies are to be constructed, the total variance σ^2 can be calculated. This means that in longitudinal diary construction there is a direct trade-off between σ_w^2 and σ_b^2 ; one can only be made larger if the other is made smaller, given that the diaries are to be sampled in an unbiased manner.

This is starkly illustrated by considering two extreme approaches to assembling longitudinal diaries. If, for each person, one simply chooses a single diary and reuses it each day, then $\sigma_w^2 = 0$ and σ_b^2 is maximized. Alternatively, if a new diary is chosen at random every day for each person, each individual tends to have a similar σ_i^2, and σ^2 is comprised mostly of σ_w^2, while σ_b^2 tends to zero (particularly for large J).

The population variability in typical measures of long-term exposure like annual average daily dose (ADD) or lifetime average daily dose (LADD) is proportional to σ_b^2. The mean exposure will be correct for any unbiased method of longitudinal diary construction. But for a fixed mean, a higher variance implies that the high end of the exposure distribution will be at higher values (and also that the low end is at lower values). The high-end exposures are often of interest, and the estimates of these exposures will be sensitive to the method of constructing longitudinal diaries. Hence it is important that the method be matched to experimental data as far as possible.

Define D as

$$D = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2) = \sigma_b^2 / \sigma^2 \tag{1-9}$$

This definition is similar to the definition of ICC used by Xue et. al. (2004). The value of D may range from zero (when $\sigma_b^2 = 0$) up to one (when $\sigma_w^2 = 0$). Using equations (1-7) and (1-8), the expression for D becomes

$$P$$

$$D \;=\; 1/(P\,\sigma^2) \sum_{i=1} \mu_i^2 \;-\; \mu^2/\sigma^2 \qquad (1\text{-}10)$$

So this statistic reflects the distribution of the personal means μ_i. It does not reflect any patterns or ordering of the $t_{i\,j}$ within a diary. Note that it is possible to interchange two or more days (interchange $t_{i\,j}$ and $t_{i\,k}$ for two days 'j' and 'k'), without changing μ_i or σ_i^2 (or μ or σ^2).

Thus, longitudinal diary construction can be separated into two problems, the first being the selection of the set of $t_{i\,j}$ values without any particular regard to day order, and the second being to reorder them to match the patterns expected within individual longitudinal diaries. These patterns are summarized by autocorrelation statistics, discussed in section 3 below.

2) Specifying the parameters of the beta distributions for the T_i and X-scores

To complete the description of the proposed method, formulas for the parameters of the beta distributions are required. Each person simulated is assigned a personal target score T_i, which is the mean of the distribution from which their X-scores are drawn. For this analysis, the $t_{i\,j}$ are the X-scores. There are two constraints to be met. The set of selected values $t_{i\,j}$ should produce a sample D statistic close to the requested value, and the set of $t_{i\,j}$ (across all persons) should be as uniformly distributed between zero and one as is possible.

A beta distribution with parameters 'a' and 'b', bounded by zero and one, has a probability density function (pdf) given by

$$p(x) = \Gamma(a+b)\;\; x^{\,a-1}\,(1-x)^{\,b-1} \,/\, [\,\Gamma(a)\;\;\Gamma(b)\,] \qquad (2\text{-}1)$$

which has a mean of

$$\mu\,(a,b) \;=\; a\,/\,(a+b) \qquad (2\text{-}2)$$

and a variance

$$\sigma^2\,(a,b) \;=\; a\,b\,/\,[(a+b)^2\,(1+a+b)] \qquad (2\text{-}3)$$

(see for example Johnson, Kotz, and Balakrishnan, 1994). Replacing 'a' and 'b' by the mean μ and the sum S (where S = a+b) results in

$$p(x) = \Gamma(S)\; x^{\,\mu S - 1}\,(1-x)^{\,S - \mu S - 1} \;/\, [\,\Gamma(\mu\,S)\;\;\Gamma(S - \mu\,S)\,] \qquad (2\text{-}4)$$

$$\sigma^2\,(\mu,S) = \;\mu\,(1-\mu)\,/\,(1+S)\,. \qquad (2\text{-}5)$$

Except for the beta distribution that is used for selecting the personal targets T_i, all the beta distributions used in this approach have bounds zero and one, so the above formulas apply.

For the T_i, the bounds of the beta distributions are at (1/2-w/2) and (1/2+w/2), where 'w' is a function of 'D' and may range from zero to one. These beta distributions are symmetric about their midpoint, so 'a' = 'b' = α . The pdf in such cases is

$$p(x) = \Gamma(2\alpha) (1 - (2x-1)^2 / w^2)^{\alpha-1} / [w \Gamma(\alpha)^2 2^{2\alpha-2}] , \quad \text{for } (1/2-w/2) < x < (1/2+w/2) \qquad (2\text{-}6)$$

and the statistics for this distribution are mean

$$\mu = 1/2 \qquad (2\text{-}7)$$

which is obvious from the symmetry, and variance

$$\sigma^2 = w^2 / (4 + 8 \alpha) . \qquad (2\text{-}8)$$

For a particular person with a target score T_i , the beta distribution from which their X-scores are drawn has a mean T_i and a variance which follows from (2-5):

$$\sigma_i^2 = T_i (1 - T_i) / (1 + S_i) \qquad (2\text{-}9)$$

where S_i is the sum of the 'a' and 'b' parameters for that particular person. For a sample of size J drawn from this distribution, the square of the standard error of the mean is σ_i^2 / J. Also, the expected value of the sample variance is

$$s_i^2 = (J-1) \sigma_i^2 / J . \qquad (2\text{-}10)$$

The within-person variance σ_w^2 is the mean across persons of the s_i^2. In the limit of a large simulated population, this is the same as the weighted average over T_i .

$$\begin{aligned} \sigma_w^2 &= \int p(T_i) s_i^2 d T_i \\ &= ((J-1)/J) \int p(T_i) T_i (1 - T_i) / (1 + S_i) d T_i . \qquad (2\text{-}11) \end{aligned}$$

This integral has a simple solution if the denominator can be factored out, which is possible when the sum of the parameters of the beta distribution for the X-scores, which is S_i, is the same for all persons (or all T_i). Assume that such a solution exists that also meets all the other constraints; that is, assume S_i is equal to a constant S for all persons. The remaining terms in the integral consist of the difference between the first and second moments of T_i about the origin. The first moment is the mean (which is 1/2), while the second moment about the origin is the variance plus the square of the mean. The variance of the T_i is given by equation (2-8). Hence

$$\sigma_w^2 = [(J-1)/(J + J S)] [1/2 - (1/4 + w^2 / (4 + 8 \alpha))]$$

$$= [(J-1)/(J + J S)] [1/4 - w^2 / (4 + 8 \alpha)] . \qquad (2\text{-}12)$$

Now consider the between-person variance σ_b^2. It can be interpreted as the second moment about the overall mean μ. (see equation 1-7). Here the mean of the T_i is 1/2 by equation (2-7). For one value of T_i, if several persons share this T_i then the expected variance in μ_i for this subgroup is the square of the standard error of the mean. Each person is assigned J X-scores, one per simulation day. The standard error of the mean of these scores is given by $\sigma_i / J^{1/2}$, hence the expected variance in μ_i for persons sharing the same T_i is σ_i^2 / J, which by equation (2-9) is $T_i (1 - T_i) / (J + J S_i)$. To evaluate σ_b^2, the variance in μ_i about T_i must be converted to the second moment of μ_i about the overall population mean of 1/2.

Hence,

$$(2^{nd} \text{ moment of } \mu_i \text{ about 1/2 for given } T_i) = \int p(\mu_i) (\mu_i - 1/2)^2 \, d\mu_i$$

$$= \int p(\mu_i) (\mu_i^2 - \mu_i + 1/4) \, d\mu_i$$

$$= \int p(\mu_i) \mu_i^2 \, d\mu_i - T_i + 1/4 \qquad (2\text{-}13)$$

which follows since $\int p(\mu_i) \mu_i \, d\mu_i = T_i$ and also $\int p(\mu_i) 1/4 \, d\mu_i = 1/4$.

The variance in μ_i is the second moment about the mean T_i, or

$$\sigma_i^2 / J = \int p(\mu_i) (\mu_i - T_i)^2 \, d\mu_i$$

$$= \int p(\mu_i) \mu_i^2 \, d\mu_i - \int p(\mu_i) \, 2 \mu_i T_i \, d\mu_i + \int p(\mu_i) T_i^2 \, d\mu_i$$

$$= \int p(\mu_i) \mu_i^2 \, d\mu_i - 2 T_i^2 + T_i^2 \qquad (2\text{-}14)$$

Substituting this expression into (2-13) gives

$$(2^{nd} \text{ moment of } \mu_i \text{ about } 1/2 \text{ for given } T_i) = \sigma_i^{\,2} / J + T_i^{\,2} - T_i + 1/4$$

$$= \sigma_i^{\,2} / J + (T_i - 1/2)^2$$

$$= T_i (1 - T_i) / (J + J S_i) + (T_i - 1/2)^2 . \qquad (2\text{-}15)$$

For a large simulated population, σ_b^2 is the mean of this quantity over all T_i, namely

$$\sigma_b^2 = \int p(T_i) [T_i (1 - T_i) / (J + J S_i) + (T_i - 1/2)^2] \, d T_i \qquad (2\text{-}16)$$

where $p(T_i)$ is given by equation (2-6). This integral can be split in two; the first part is the same integral as in (2-11), while the second part is just the variance in T_i, which is given by equation (2-8). As for σ_w^2, the first integral can be solved by assuming S_i is constant for all persons. So

$$\sigma_b^2 = [1/(J + J S)] [1/4 - w^2 / (4 + 8 \alpha)] + w^2 / (4 + 8 \alpha) . \qquad (2\text{-}17)$$

Collectively, the X-scores for all the simulated persons should be as close to uniformly distributed as possible, to ensure no net bias in the usage of the diaries. In general this cannot be achieved exactly, but it is possible to ensure that the X-scores collectively have the same mean and variance as a uniform distribution, which for a uniform bounded by zero and one is

$$\text{mean of X-scores} = 1/2 , \qquad (2\text{-}18)$$

$$\text{variance of X-scores} = 1/12 . \qquad (2\text{-}19)$$

The mean will be 1/2 by the symmetry of the T_i distribution about 1/2. The collective variance of the X-scores is related to σ_b^2 and σ_w^2 by equation (1-8). Hence

$$\sigma_b^2 + \sigma_w^2 = 1/12 . \qquad (2\text{-}20)$$

Substituting in the expressions from (2-12) and (2-17), one obtains

$$[1/(1+S)] [1/4 - w^2 / (4 + 8 \alpha)] + w^2 / (4 + 8 \alpha) = 1/12 \qquad (2\text{-}21)$$

which when solved for S results in

$$S = 2 / (1 - 3 w^2 / (1 + 2 \alpha)) . \qquad (2\text{-}22)$$

This result permits the specification of the parameters for the beta distribution for person 'i' from which all their X-scores are drawn. This distribution has a mean of T_i. Remembering that

S = a+b, equation (2-2) can be written as

$$a = S\ T_i = \ 2\ T_i\ / (1 - 3\ w^2 / (1 + 2\ \alpha))\ . \tag{2-23}$$

Therefore

$$b = S\text{-}a\ = 2\ (\ 1 - T_i\ ) / (1 - 3\ w^2 / (1 + 2\ \alpha))\ \ . \tag{2-24}$$

From equation (2-6), the parameters w and α define the beta distribution from which all the personal target scores T_i are drawn. This distribution must match the requirements of the D statistic. To simplify the equations, define a new parameter

$$D^{\#}\ =\ 3\ w^2 / (1 + 2\ \alpha) \tag{2-25}$$

and rewrite equation (2-17) in terms of this new parameter:

$$\sigma_b^2\ =\ [1/(J + J\ (2/(1\text{-}D^{\#})))]\ [1/4 - D^{\#}/12]\ +\ D^{\#}/12. \tag{2-26}$$

which can be solved for $D^{\#}$ in terms of σ_b^2

$$D^{\#}\ =\ (12\ J\ \sigma_b^2\ \text{-}1\ ) / (\ J - 1)\ . \tag{2-27}$$

Also, equation (1-9) together with (2-20) give

$$D = \sigma_b^2\ / (\sigma_b^2 + \sigma_w^2\ )\ =\ 12\ \sigma_b^2\ \ . \tag{2-28}$$

Hence

$$D^{\#}\ =\ (J\ D - 1\ ) / (J - 1\ ) \tag{2-29}$$

As J becomes large, $D^{\#}$ approaches D. Therefore, $D^{\#}$ may be seen as a modified D statistic that accounts for the effects of short simulation periods. Since the user specifies the simulation length J and the diversity statistic D directly, $D^{\#}$ is therefore also specified. However, equation (2-25) still contains two unknowns (w and α). Thus, there is no unique solution.

Let R be the square root of $D^{\#}$

$$R\ =\ (D^{\#}\ )^{1/2}\ . \tag{2-30}$$

It is found empirically that the following relationship between α and R gives a nearly uniform distribution of X-scores:

$$\alpha = 1 - (4/5)\ [4\ R\ (1 - R)]^3 \tag{2-31}$$

In practice, the above formulas give a nearly uniform collective distribution of X-scores, and hence a nearly uniform usage of the available diaries. For example, suppose one year time series are generated for a large number of simulated persons, using a pool of 100 diaries. (For this purpose, neglect the effects of altering diary pools throughout the year.) Strict uniformity would result in each diary being assigned an average of 3.65 times per person (365 days divided by 100 diaries). The above formulae using beta distributions result in each diary being used between 3.50 and 4.0 times per person. Furthermore, both the mean and variance of the key variable on the assembled time series match the mean and variance seen in the diary pool. If even better uniformity in diary usage is desired, it is possible to use a smoothing function on the X-scores, at a slight cost in departing from strict beta distributions. This is usually not necessary and is not detailed here.

Unlike the other equations in this derivation, there is no necessity to use equation (2-31) when implementing this method. Any functions for α and 'w' that produce valid beta distributions and satisfy equation (2-25) may be used. Another choice which is simpler than (2-31) is

$$\alpha = 1 \tag{2-32}$$

whereupon equation (2-25) reduces to

$$w^2 = D^{\#} . \tag{2-33}$$

This choice results in a uniform distribution of the targets T_i between the limits $(1/2\text{-}D^{\#}/2)$ and $(1/2+D^{\#}/2)$. However, while the D statistic is matched, and the mean and variance of the X-scores match those of a uniform distribution, overall the X-scores are slightly less uniformly distributed than is obtained by using equation (2-31). The choice of functions for α and 'w' could be based on preferences for statistics other than D; for example, one might wish to match statistics on the distribution of the T_i targets themselves.

3) Method of reordering the diaries to match a target value of A

The second step in the proposed method for constructing longitudinal activity diaries is the reordering of the selected X-scores to match a target value for 'A'. It should be noted that autocorrelation is hard to measure on short time series. Box, Jenkins and Reinsel (1994) recommend a minimum of 50 data points to adequately characterize the autocorrelation of a time series. The method described below does a reasonable job for 30 days or more. The method can be applied to shorter time series, but the results will not match the target autocorrelation as closely as for longer simulations.

For purposes of autocorrelation, the ranks that matter are the ranks relative to the other days in the same time series. These ranks may differ substantially from the original X-scores. Note that the X-scores are uniformly distributed across persons and hence the mean (across persons) is 1/2, but the mean within a time series for a given person is μ_i . Hence the ranking of X-scores across persons may differ substantially from the ranking within persons.

To start the process of targeting the desired overall autocorrelation A, assign a target autocorrelation a_i to each simulated individual. These targets can be drawn from any distribution that has a mean of A, provided that all a_i are between -1 and 1.

Sort the X-scores within each simulated individual's time series and rank them from smallest to largest. Suppose there are J days in the simulation period. Recall that some extra X-scores (approximately 3%) should be selected for each person. The extra ones are needed to prevent a severe loss of degrees of freedom towards the end of each individual's reordering. Let K be the number of X-scores selected per person, including the extras. When sorted and ranked, the set of available ranks will be the integers from 1 to K. For example, rank 1 will correspond to the lowest of the X-scores assigned to this person, rank 2 is the second lowest X-score, and so on. Ties will not generally occur, as the X-scores are real numbers selected from continuous distributions; ties are ignored in practice. The goal is to reorder these ranks in a stochastic manner that will (on average) reproduce the requested autocorrelation. The reordering process will stop once J values are selected, any extras are discarded.

Let R_j be the rank assigned to day 'j' by this reordering process. The lag-one autocorrelation 'a_i' of the time series for person 'i' is the ratio of the lag-one covariance to the variance, or

$$a_i \;=\; E\,[\,(R_j - \rho)\,(\,R_{j+1} - \rho)\,]\;/\;E\,[\,(R_j - \rho)^2\,] \qquad (3\text{-}1)$$

where ρ is the mean of the ranks. Here E [arg] means the expected value of the argument 'arg'. There is a slight difference in the autocorrelation of the entire set of K ranks as compared to the autocorrelation of the J ranks of the selected subset, although this difference is quite small for J close to K. One difficulty is that while the latter is a measurable output from the diary assembly process, it is the former that is accessible during the reordering process. Thus, the ranks, means, and variances in equation (3-1) and subsequent equations apply to the full list of K values.

The denominator in equation (3-1) is the variance of the set of integers from 1 to K, which is $(K^2-1)/12$. Hence

$$a_i \;= (12\,/\,(K^2 - 1))\;\;E\,[(R_j - \rho)\,(\,R_{j+1} - \rho)\,]. \qquad (3\text{-}2)$$

The expectation value in equation (3-2) can be evaluated if we have the conditional probability $p(R_{j+1}\,|\,R_j\,)$; that is, the probability for each rank being chosen on day 'j+1', given the rank R_j chosen on day 'j'.

The conditional probability distribution $p(R_{j+1}\,|\,R_j\,)$ is a discrete distribution, since the set of ranks is discrete. However, the number of ranks is often in the hundreds, and it is more convenient to use a continuous probability distribution. Thus, a beta distribution for p(y|x) is developed, where 'x' and 'y' are continuous variables ranging from zero to one that are mapped onto the ranks:

$$R_j = \text{ceil}(K\ x), \quad \text{and} \ \ R_{j+1} = \text{ceil}(K\ y) \tag{3-3}$$

where 'ceil' is the ceiling or least integer function that rounds up to the next integer. To invert these relationships, note that on average the ceiling function adds 1/2 to the argument, so the mean values of x and y that correspond to given ranks are

$$x = (R_j - 1/2) / K, \quad \text{and} \quad y = (R_{j+1} - 1/2) / K \tag{3-4}$$

We wish to select the parameters for a beta distribution that give the selection probabilities for R_{j+1}, based on the value of R_j and the other constants in equation (3-2). The expected value appearing in equation (3-2) is given by weighting the sum over all outcomes by the probability of occurrence:

$$E\ [(R_j - \rho)\ (\ R_{j+1} - \rho)\ ] = \ \Sigma\ \Sigma\ \ (R_j - \rho)\ (\ R_{j+1} - \rho)\ p(R_j)\ p(R_{j+1}\ |\ R_j) \tag{3-5}$$

where one sum is over all R_j from 1 to K and the other is over all R_{j+1} from 1 to K. All values for R_j should be equally likely, that is, $p(R_j)$ =1/K for all cases. Replace the sum over all R_{j+1} by an integral over all y, with R_{j+1} replaced by (Ky+1/2):

$$E\ [(R_j - \rho)\ (\ R_{j+1} - \rho)\ ] = \ (1/K)\ \ \Sigma\ (R_j - \rho)\ \ \int p(y|x)\ (Ky+1/2- \rho)\ dy \tag{3-6}$$

The integral over 'y' consists of two parts. Factoring out the K, the first part is the mean value of 'y' for the given 'x', which can be symbolized as E(y|x). The second part equals (1/2-ρ), since the integrand is independent of y, and ∫p(y|x) dy = 1. Also note that for the integers from 1 to K, the mean is ρ = (K+1)/2, so (1/2-ρ) = -K/2. Thus,

$$E\ [(R_j - \rho)\ (\ R_{j+1} - \rho)\ ]\ = \ \Sigma\ (R_j - \rho)\ (\ E(y|x) - 1/2) \tag{3-7}$$

The value of E(y|x) will depend on the parameters of the beta distribution. As in section 2, let 'a' and 'b' be the parameters of the beta distribution and let S = a+b. Consider the following:

$$a = S/2 - S\ w/2 + S\ w\ x\ . \tag{3-8}$$

Then

$$b = S\text{-}a \ = \ S/2 + S\ w/2 - S\ w\ x. \tag{3-9}$$

The mean of a beta distribution bounded by zero and one is given by equation (2-2), therefore

$$E\,(y|x) = a\,/\,(a+b) = a\,/S = 1/2 - w/2 + w\,x\,. \tag{3-10}$$

Thus, equation (3-7) becomes

$$E\,[(R_j - \rho)\,(R_{j+1} - \rho)\,] = \Sigma\,(R_j - K/2 - 1/2)\,w\,(x-1/2). \tag{3-11}$$

Replacing x by $(R_j -1/2)/K$ and noting that the sums evaluate to

$$\Sigma\,R_j^{\,2} = K\,(K+1)\,(2K+1)/6, \quad \Sigma\,R_j = K\,(K+1)/2, \quad \Sigma\,1 = K, \tag{3-12}$$

then equation (3-11) can be expanded and evaluated to give

$$E\,[(R_j - \rho)\,(R_{j+1} - \rho)\,] = w\,(K^2 - 1)\,/12. \tag{3-13}$$

With this choice of the beta distribution, equation (3-2) reduces to the very simple form

$$w = a_i. \tag{3-14}$$

To completely specify the parameters of the beta distribution, a form for the sum of parameters S = a+b must be given. The second requirement is that the distribution of ‘y’ be essentially uniform, when averaged over all values of ‘x’. In practice, this condition cannot be met exactly. Instead, a reasonable match can be made by matching the first few moments of the distribution for ‘y’ to the moments of a uniform distribution.

The kth moment about zero of a uniform distribution from zero to one is

$$m_k = \int x^k\ p(x)\,dx = 1/\,(k+1) \tag{3-15}$$

since p(x) =1 for a uniform. The moments of the ‘y’ values are

$$\begin{aligned} E(y^k) &= \int\int\ y^k\,p(y|x)\,p(x)\,dy\,dx \\ &= \int p(x)\,dx\ \ \int y^k\,p(y|x)\,dy\,. \end{aligned} \tag{3-16}$$

The second integral is the kth moment of the beta distribution p(y|x). The first moment is given by equation (3-10). The second moment of a beta distribution (Johnson, Kotz, and Balakrishnan, 1994) is

$$\begin{aligned} M_2 &= a\,(a+1)\,/\,(\,S\,(S+1)\,) \\ &= (S/2 - S\,w\,/2 + S\,w\,x\,)\,(1+S/2 - S\,w/2 + S\,w\,x)\,/\,[S\,(S+1)\,] \end{aligned}$$

$$= [\ (1\text{-}w)(2+S\text{-}S\ w)/4 + (w + S\ w - s\ w^2\ )\ x + S\ w^2\ x^2\ ]\ /\ (S+1). \qquad (3\text{-}17)$$

Hence the first moment of 'y' is

$$E(y\ ) = \int p(x)\ (1/2 - w/2 + w\ x)\ dx$$

$$= 1/2 - w/2 + w\ (1/2)$$

$$= 1/2 \qquad (3\text{-}18)$$

which agrees with the first moment m_1 of a uniform (0,1) distribution. For the second moment, equation (3-17) must be integrated over x from zero to one, giving

$$E(y^2\ ) = [(1\text{-}w)(2+S\text{-}S\ w)/4 + (w+S\ w - S\ w^2)/2 + S\ w^2/3\ ]\ /\ (S+1)$$

$$= (6 + 3\ S + S\ w^2\ )\ /\ (12 + 12\ S) \qquad (3\text{-}19)$$

Matching $E(y^2)$ to the second moment m_2 of a uniform (0,1) distribution (which is 1/3) and solving for S gives

$$S = 2\ /\ (1\text{-}\ w^2) \qquad (3\text{-}20)$$

Matching the third moments $m_3 = E\ (y^3)$ results in the same relationship $S = 2\ /\ (1\text{-}\ w^2)$. Moments higher than this generally will not match.

To summarize, the parameter values should be

$$w = a_i$$

$$S = 2\ /\ (1\text{-}\ w^2). \qquad (3\text{-}21)$$

The preceding development is in terms of the target autocorrelation 'a_i' that is specific to one individual 'i'. The population statistic A is the mean of the a_i across persons. An examination of the data used in Xue et al. (2004) indicates that people within the same cohort may differ greatly in their personal autocorrelations. For four different choices of the key variable, the standard deviation of a_i across persons was 0.20. A symmetric beta distribution centered on A with a standard deviation of 0.20 was chosen for the results reported here. The bounds on this beta are (A-1/2) to (A+1/2), provided these do not extend past 1 or -1. If A is less than -1/2 or greater than 1/2, the beta distribution is "squeezed" symmetrically until the bounds are within limits. Other choices of the key variable or data from other studies may lead to alternative choices for the distribution of 'a_i' .

The proposed method could allow differing autocorrelations for different points in the time series. For example, suppose that there is one autocorrelation for the case where both days 'j' and 'j+1' are of the same day-type, and another if the days are of differing day-types. The average autocorrelation is the weighted average. The user would specify the overall target autocorrelations A_j for each day-type. For each individual, a target a_{ij} is required for each A_j. Since a new beta distribution is generated every day, one merely replaces a_i by a_{ij} in equations (3-21), so that 'w' and 's' become functions of 'j'. Note that there are few data sets extensive enough to determine if this effect is significant. Furthermore, the stability of each autocorrelation target will decrease when it is applied to fewer and fewer days. Hence, the derivation does not emphasize this possibility.

4) Mapping the X-scores back to activity diaries

For the first day of the simulation, select any of the ranks from 1 to K at random. For each subsequent day, a beta distribution with parameters determined by (3-21) and (3-8) is used to select the next rank. The beta distribution will return a real number between zero and one; call this value 'y'. Convert this to a rank R from 1 to K by

$$R_{j+1} = \text{ceil}(K\ y) \tag{4-1}$$

where 'ceil' is the ceiling function. The only complication is if this rank has already been assigned to a prior day, in which case the nearest rank that has not already been used is assigned instead. The X-score corresponding to this rank is recorded (call it x_{j+1}), and the assigned rank is used to adjust the parameters of the beta distribution to be used for the next day. Continue until J values have been assigned.

To connect the time series of X-scores with actual diaries, the pool of available diaries for each day must be identified. If there are D_{j+1} available diaries in the pool for simulation day 'j+1', then use the diary at position d_{j+1} in the sorted list of available diaries, where

$$d_{j+1} = \text{ceil}(D_{j+1}\ x_{j+1}). \tag{4-2}$$

5) Possible Modifications

There are several reasons why the derivation of the parameters needed to match a target autocorrelation yields an approximate, but not analytically exact, solution. First, and most importantly for short simulations, the correspondence between the discrete nature of the ranks and the continuous beta distribution can become a difficulty. Mathematically, this means that larger segments of x and y space map onto a single rank. The implicit assumption in the derivation is that ranks can be mapped to the midpoint of these segments, and vice versa. This is a good approximation as long as the probability is not rapidly changing within each segment, which is the case when each segment is small (meaning many days in the simulation).

Secondly, the beta distribution for reordering may select the same rank on two days in a row, in which case the second rank must be shifted away from the first, which lowers autocorrelation. In fact, anytime selected rank R_{j+1} has been used before, the result must be shifted to the nearest unused rank. However, if this is not the same rank as R_j, then the shift is equally likely to move the rank closer to R_j as moving it further away, so the net effect on autocorrelation is small.

Additionally, near the end of the simulation, there are relatively few unused ranks, and in practice these ranks may be near to each other. So when examined in detail, the time series may show a tendency for autocorrelation to increase toward the end.

The final point is that within each time series the rankings for the J selected X-scores will differ from the rankings with respect to all K of the X-scores. Usually, this is not a problem since the mean and variance of the subset are close to the mean and variance of the larger set. In exceptional cases, the autocorrelation measured on the original rankings (based on K) may differ from the autocorrelation based on the rankings within the subset; this could happen when the omitted X-scores are congregated near one end of the ranking scale.

Accounting for the above factors may be possible by modifying the proposed method, though at a cost of complicating the approach. However, in total, these effects tend to be small for simulations over 30 days in length. Also, some of the potential problems have a tendency to cancel out. It is found in simulations that D and A are usually within 0.02 of the requested value, an excellent agreement.

APPENDIX D. UNCERTAINTY ANALYSIS OF RESIDENTIAL AIR EXCHANGE RATE DISTRIBUTIONS

This page intentionally left blank.

MEMORANDUM

To: John Langstaff, EPA OAQPS

From: Jonathan Cohen, Arlene Rosenbaum, ICF International

Date: June 5, 2006

Re: Uncertainty analysis of residential air exchange rate distributions

This memorandum describes our assessment of some of the sources of the uncertainty of city-specific distributions of residential air exchange rates that were fitted to the available study data. City-specific distributions for use with the APEX ozone model were developed for 12 modeling cities, as detailed in the memorandum by Cohen, Mallya and Rosenbaum, 2005[7] (Appendix A of this report). In the first part of the memorandum, we analyze the between-city uncertainty by examining the variation of the geometric means and standard deviations across cities and studies. In the second part of the memorandum, we assess the within-city uncertainty by using a bootstrap distribution to estimate the effects of sampling variation on the fitted geometric means and standard deviations for each city. The bootstrap distributions assess the uncertainty due to random sampling variation but do not address uncertainties due to the lack of representativeness of the available study data, the matching of the study locations to the modeled cities, and the variation in the lengths of the AER monitoring periods.

Variation of geometric means and standard deviations across cities and studies

The memorandum by Cohen, Mallya and Rosenbaum, 2005[8] (Appendix A of this report) describes the analysis of residential air exchange rate (AER) data that were obtained from seven studies. The AER data were subset by location, outside temperature range, and the A/C type, as defined by the presence or absence of an air conditioner (central or window). In each case we chose to fit a log-normal distribution to the AER data, so that the logarithm of the AER for a given city, temperature range, and A/C type is assumed to be normally distributed. If the AER data has geometric mean GM and geometric standard deviation GSD, then the logarithm of the AER is assumed to have a normal distribution with mean log(GM) and standard deviation log(GSD).

Table D-1 shows the assignment of the AER data to the 12 modeled cities. Note that for Atlanta, GA and Washington DC, the Research Triangle Park, NC data for houses with A/C was used to represent the AER distributions for houses with A/C, and the non-California data for houses

[7] Cohen, J., H. Mallya, and A. Rosenbaum. 2005. Memorandum to John Langstaff. *EPA 68D01052, Work Assignment 3-08. Analysis of Air Exchange Rate Data*. September 30, 2005.
[8] *Op. Cit.*

without A/C was used to represent the AER distributions for houses without A/C. Sacramento, CA AER distributions were estimated using the AER data from the inland California counties of Sacramento, Riverside, and San Bernardino; these combined data are referred to by the City Name "Inland California." St Louis, MO AER distributions were estimated using the AER data from all states except for California and so are referred to be the City Name "Outside California."

Table D-1. Assignment of Residential AER distributions to modeled cities

Modeled city	AER distribution
Atlanta, GA, A/C	Research Triangle Park, A/C only
Atlanta, GA, no A/C	All non-California, no A/C ("Outside California")
Boston, MA	New York
Chicago, IL	New York
Cleveland, OH	New York
Detroit, MI	New York
Houston, TX	Houston
Los Angeles, CA	Los Angeles
New York, NY	New York
Philadelphia, PA	New York
Sacramento	Inland parts of Los Angeles ("Inland California")
St. Louis	All non-California ("Outside California")
Washington, DC, A/C	Research Triangle Park, A/C only
Washington, DC, no A/C	All non-California, no A/C ("Outside California")

It is evident from Table D-1 that for some of the modeled cities, some potentially large uncertainty was introduced because we modeled their AER distributions using available data from another city or group of cities thought to be representative of the first city on the basis of geography and other characteristics. This was necessary for cities where we did not have any or sufficient AER data measured in the same city that also included the necessary temperature and A/C type information. One way to assess the impact of these assignments on the uncertainty of the AER distributions is to evaluate the variation of the fitted log-normal distributions across the cities with AER data. In this manner we can examine the effect on the AER distribution if a different allocation of study data to the modeled cities had been used.

Even for the cities where we have study AER data, there is uncertainty about the fitted AER distributions. First, the studies used different measurement and residence selection methods. In some cases the residences were selected by a random sampling method designed to represent the entire population. In other cases the residences were selected to represent sub-populations. For example, for the RTP study, the residences belong to two specific cohorts: a mostly Caucasian, non-smoking group aged at least 50 years having cardiac defibrillators living in Chapel Hill; a group of non-smoking, African Americans aged at least 50 years with controlled hypertension living in a low-to-moderate SES neighborhood in Raleigh. The TEACH study was restricted to residences of inner-city high school students. The RIOPA study was a random sample for Los Angeles, but was designed to preferentially sample locations near major air toxics sources for Elizabeth, NJ and Houston TX. Furthermore, some of the studies focused on different towns or cities within the larger metropolitan areas, so that, for example, the Los Angeles data from the Avol study was only measured in Lancaster, Lake Gregory, Riverside, and San Dimas but the Los Angeles data from the Wilson studies were measured in multiple cities in Southern California. One way to assess the uncertainty of the AER distributions due to variations of study methodologies and study sampling locations is to evaluate the variation of the fitted log-normal distributions within each modeled city across the different studies.

We evaluated the variation between cities, and the variation within cities and between studies, by tabulating and plotting the AER distributions for all the study/city combinations. Since the original analyses by Cohen, Mallya and Rosenbaum, 2005 clearly showed that the AER distribution depends strongly on the outside temperature and the A/C type (whether or not the residence has air conditioning), this analysis was stratified by the outside temperature range and the A/C type. Otherwise, study or city differences would have been confounded by the temperature and A/C type differences and you would not be able to tell how much of the AER difference was due to the variation of temperature and A/C type across cities or studies. In order to be able to compare cities and studies we could not use different temperature ranges for the different modeled cities as we did for the original AER distribution modeling. For these analyses we stratified the temperature into the ranges <= 10, 10-20, 20-25, and >25 ºC and categorized the A/C type as “Central or Window A/C” versus ‘No A/C,” giving 8 temperature and A/C type strata.

Table D-2 shows the geometric means and standard deviations by city and study. These geometric mean and standard deviation pairs are plotted in Figure D-1 through D-8. Each figure shows the variation across cities and studies for a given temperature range and A/C type. The results for a city with only one available study are shown with a blank study name. For cities with multiple studies, results are shown for the individual studies and the city overall distribution is designated by a blank value for the study name.

Table D-2. Geometric means and standard deviations by city and study.

A/C Type	Temperature	City	Study*	N	Geo Mean	Geo Std Dev**
Central or Room A/C	<= 10	Houston		2	0.32	1.80
Central or Room A/C	<= 10	Los Angeles		5	0.62	1.51
Central or Room A/C	<= 10	Los Angeles	Avol	2	0.72	1.22
Central or Room A/C	<= 10	Los Angeles	RIOPA	1	0.31	
Central or Room A/C	<= 10	Los Angeles	Wilson 1991	2	0.77	1.12
Central or Room A/C	<= 10	New York City		20	0.71	2.02
Central or Room A/C	<= 10	Research Triangle Park		157	0.96	1.81
Central or Room A/C	<= 10	Sacramento		3	0.38	1.82
Central or Room A/C	<= 10	San Francisco		2	0.43	1.00

Table D-2. Geometric means and standard deviations by city and study.

A/C Type	Temperature	City	Study*	N	Geo Mean	Geo Std Dev**
Central or Room A/C	<= 10	Stockton		7	0.48	1.64
Central or Room A/C	10-20	Arcata		1	0.17	
Central or Room A/C	10-20	Bakersfield		2	0.36	1.34
Central or Room A/C	10-20	Fresno		8	0.30	1.62
Central or Room A/C	10-20	Houston		13	0.42	2.19
Central or Room A/C	10-20	Los Angeles		716	0.59	1.90
Central or Room A/C	10-20	Los Angeles	Avol	33	0.48	1.87
Central or Room A/C	10-20	Los Angeles	RIOPA	11	0.60	1.87
Central or Room A/C	10-20	Los Angeles	TEACH	1	0.68	
Central or Room A/C	10-20	Los Angeles	Wilson 1984	634	0.59	1.89
Central or Room A/C	10-20	Los Angeles	Wilson 1991	37	0.64	2.11
Central or Room A/C	10-20	New York City		5	1.36	2.34
Central or Room A/C	10-20	New York City	RIOPA	4	1.20	2.53
Central or Room A/C	10-20	New York City	TEACH	1	2.26	
Central or Room A/C	10-20	Redding		1	0.31	
Central or Room A/C	10-20	Research Triangle Park		320	0.56	1.91
Central or Room A/C	10-20	Sacramento		7	0.26	1.67
Central or Room A/C	10-20	San Diego		23	0.41	1.55
Central or Room A/C	10-20	San Francisco		5	0.42	1.25
Central or Room A/C	10-20	Santa Maria		1	0.23	
Central or Room A/C	10-20	Stockton		4	0.73	1.42
Central or Room A/C	20-25	Houston		20	0.47	1.94
Central or Room A/C	20-25	Los Angeles		273	1.10	2.36
Central or Room A/C	20-25	Los Angeles	Avol	32	0.61	1.95
Central or Room A/C	20-25	Los Angeles	RIOPA	26	0.90	2.42
Central or Room A/C	20-25	Los Angeles	Wilson 1984	215	1.23	2.33
Central or Room A/C	20-25	New York City		37	1.11	2.74
Central or Room A/C	20-25	New York City	RIOPA	20	0.93	2.91
Central or Room A/C	20-25	New York City	TEACH	17	1.37	2.52
Central or Room A/C	20-25	Red Bluff		2	0.61	3.20
Central or Room A/C	20-25	Research Triangle Park		196	0.40	1.89
Central or Room A/C	> 25	Houston		79	0.43	2.17
Central or Room A/C	> 25	Los Angeles		114	0.72	2.60
Central or Room A/C	> 25	Los Angeles	Avol	25	0.37	3.10
Central or Room A/C	> 25	Los Angeles	RIOPA	10	0.94	1.71
Central or Room A/C	> 25	Los Angeles	Wilson 1984	79	0.86	2.33
Central or Room A/C	> 25	New York City		19	1.24	2.18
Central or Room A/C	> 25	New York City	RIOPA	14	1.23	2.28
Central or Room A/C	> 25	New York City	TEACH	5	1.29	2.04
Central or Room A/C	> 25	Research Triangle Park		145	0.38	1.71
No A/C	<= 10	Houston		13	0.66	1.68
No A/C	<= 10	Los Angeles		18	0.54	3.09
No A/C	<= 10	Los Angeles	Avol	14	0.51	3.60
No A/C	<= 10	Los Angeles	RIOPA	2	0.72	1.11
No A/C	<= 10	Los Angeles	Wilson 1991	2	0.60	1.00
No A/C	<= 10	New York City		48	1.02	2.14

Table D-2. Geometric means and standard deviations by city and study.

A/C Type	Temperature	City	Study*	N	Geo Mean	Geo Std Dev**
No A/C	<= 10	New York City	RIOPA	44	1.04	2.20
No A/C	<= 10	New York City	TEACH	4	0.79	1.28
No A/C	<= 10	Sacramento		3	0.58	1.30
No A/C	<= 10	San Francisco		9	0.39	1.42
No A/C	10-20	Bakersfield		1	0.85	
No A/C	10-20	Fresno		4	0.90	2.42
No A/C	10-20	Houston		28	0.63	2.92
No A/C	10-20	Los Angeles		390	0.75	2.09
No A/C	10-20	Los Angeles	Avol	23	0.78	2.55
No A/C	10-20	Los Angeles	RIOPA	87	0.78	1.96
No A/C	10-20	Los Angeles	TEACH	9	2.32	2.05
No A/C	10-20	Los Angeles	Wilson 1984	241	0.70	2.06
No A/C	10-20	Los Angeles	Wilson 1991	30	0.75	1.82
No A/C	10-20	New York City		59	0.79	2.04
No A/C	10-20	Sacramento		1	1.09	
No A/C	10-20	San Diego		49	0.47	1.95
No A/C	10-20	San Francisco		15	0.34	3.05
No A/C	10-20	Santa Maria		2	0.27	1.23
No A/C	20-25	Houston		10	0.92	2.41
No A/C	20-25	Los Angeles		148	1.37	2.28
No A/C	20-25	Los Angeles	Avol	19	0.95	1.87
No A/C	20-25	Los Angeles	RIOPA	38	1.30	2.11
No A/C	20-25	Los Angeles	Wilson 1984	91	1.52	2.40
No A/C	20-25	New York City		26	1.62	2.24
No A/C	20-25	New York City	RIOPA	19	1.50	2.30
No A/C	20-25	New York City	TEACH	7	1.99	2.11
No A/C	20-25	Red Bluff		1	0.55	
No A/C	> 25	Houston		2	0.92	3.96
No A/C	> 25	Los Angeles		25	0.99	1.97
No A/C	> 25	Los Angeles	Avol	6	1.56	1.36
No A/C	> 25	Los Angeles	RIOPA	4	1.33	1.37
No A/C	> 25	Los Angeles	TEACH	3	0.86	1.02
No A/C	> 25	Los Angeles	Wilson 1984	12	0.74	2.29
No A/C	> 25	New York City		6	1.54	1.65
No A/C	> 25	New York City	RIOPA	3	1.73	2.00
No A/C	> 25	New York City	TEACH	3	1.37	1.38

* For a given city, if AER data were available from only one study, then the study name is missing. If AER data were available for two or more studies, then the overall city distribution is shown in the row where the study name is missing, and the distributions by study and city are shown in the rows with a specific study name.

** The geometric standard deviation is undefined if the sample size equals 1.

In general, there is a relatively wide variation across different cities. This implies that the AER modeling results would be very different if the matching of modeled cities to study cities was changed, although a sensitivity study using the APEX model would be needed to assess the impact on the ozone exposure estimates. In particular the ozone exposure estimates may be sensitive to the assumption that the St. Louis AER distributions can be represented by the combined non-California AER data. One way to address this is to perform a Monte Carlo analysis where the first stage is to randomly select a city outside of California, the second stage picks the A/C type, and the third stage picks the AER value from the assigned distribution for the city, A/C type and temperature range. Note that this will result in a very different distribution to the current approach that fits a single log-normal distribution to all the non-California data for a given temperature range and A/C type. The current approach weights each data point equally, so that cities like New York with most of the data values get the greatest statistical weight. The Monte Carlo approach gives the same total statistical weight for each city and fits a mixture of log-normal distributions rather than a single distribution.

In general, there is also some variation within studies for the same city, but this is much smaller than the variation across cities. This finding tends to support the approach of combining different studies. Note that the graphs can be deceptive in this regard because some of the data points are based on very small sample sizes (N) ; those data points are less precise and the differences would not be statistically significant. For example, for the No A/C data in the range 10-20 ºC, the Los Angeles TEACH study had a geometric mean of 2.32 based on only nine AER values, but the overall geometric mean, based on 390 values, was 0.75 and the geometric means for the Los Angeles Avol, RIOPA, Wilson 1984, and Wilson 1991 studies were each close to 0.75. One noticeable case where the studies show big differences for the same city is for the A/C houses in Los Angeles in the range 20-25 ºC where the study geometric means are 0.61 (Avol, N=32), 0.90 (RIOPA, N=26) and 1.23 (Wilson 1984, N=215).

Bootstrap analyses

The 39 AER subsets defined in the Cohen, Mallya, and Rosenbaum, 2005 memorandum (Appendix A of this report) and their allocation to the 12 modeled cities are shown in Table D-3. To make the distributions sufficiently precise in each AER subset and still capture the variation across temperature and A/C type, different modeled cities were assigned different temperature range and A/C type groupings. Therefore these temperature range groupings are sometimes different to those used to develop Table D-2 and Figure D-1 through D-8.

Table D-3. AER subsets by city, A/C type, and temperature range.

Subset City Name	Study Cities	Represents Modeled Cities:	A/C Type	Temperature Range (ºC)
Houston	Houston	Houston, TX	Central or Room A/C	<=20
Houston	Houston	Houston, TX	Central or Room A/C	20-25
Houston	Houston	Houston, TX	Central or Room A/C	25-30
Houston	Houston	Houston, TX	Central or Room A/C	>30
Houston	Houston	Houston, TX	No A/C	<=10
Houston	Houston	Houston, TX	No A/C	10-20
	Houston	Houston, TX	No A/C	>20
Inland California	Sacramento, Riverside, and San Bernardino counties, CA	Sacramento, CA	Central or Room A/C	<=25
Inland California	Sacramento, Riverside, and San Bernardino counties, CA	Sacramento, CA	Central or Room A/C	>25
Inland California	Sacramento, Riverside, and San Bernardino counties, CA	Sacramento, CA	No A/C	<=10

Table D-3. AER subsets by city, A/C type, and temperature range.

Subset City Name	Study Cities	Represents Modeled Cities:	A/C Type	Temperature Range (ºC)
Inland California	Sacramento, Riverside, and San Bernardino counties, CA	Sacramento, CA	No A/C	10-20
Inland California	Sacramento, Riverside, and San Bernardino counties, CA	Sacramento, CA	No A/C	20-25
Inland California	Sacramento, Riverside, and San Bernardino counties, CA	Sacramento, CA	No A/C	>25
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	Central or Room A/C	<=20
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	Central or Room A/C	20-25
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	Central or Room A/C	25-30
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	Central or Room A/C	>30
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	No A/C	<=10
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	No A/C	10-20
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	No A/C	20-25
Los Angeles	Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties, CA	Los Angeles, CA	No A/C	>25
New York City	New York, NY	Boston, MA, Chicago, IL, Cleveland, OH, Detroit, MI, New York, NY, Philadelphia, PA	Central or Room A/C	<=10
New York City	New York, NY	Boston, MA, Chicago, IL, Cleveland, OH, Detroit, MI, New York, NY, Philadelphia, PA	Central or Room A/C	10-25
New York City	New York, NY	Boston, MA, Chicago, IL, Cleveland, OH, Detroit, MI,	Central or Room A/C	>25

Table D-3. AER subsets by city, A/C type, and temperature range.

Subset City Name	Study Cities	Represents Modeled Cities:	A/C Type	Temperature Range (ºC)
		New York, NY, Philadelphia, PA		
New York City	New York, NY	Boston, MA, Chicago, IL, Cleveland, OH, Detroit, MI, New York, NY, Philadelphia, PA	No A/C	<=10
New York City	New York, NY	Boston, MA, Chicago, IL, Cleveland, OH, Detroit, MI, New York, NY, Philadelphia, PA	No A/C	10-20
New York City	New York, NY	Boston, MA, Chicago, IL, Cleveland, OH, Detroit, MI, New York, NY, Philadelphia, PA	No A/C	>20
Outside California	Cities outside CA	St. Louis, MO	Central or Room A/C	<=10
Outside California	Cities outside CA	St. Louis, MO	Central or Room A/C	10-20
Outside California	Cities outside CA	St. Louis, MO	Central or Room A/C	20-25
Outside California	Cities outside CA	St. Louis, MO	Central or Room A/C	25-30
Outside California	Cities outside CA	St. Louis, MO	Central or Room A/C	>30
Outside California	Cities outside CA	St. Louis, MO Atlanta, GA Washington DC	No A/C	<=10
Outside California	Cities outside CA	St. Louis, MO Atlanta, GA Washington DC	No A/C	10-20
Outside California	Cities outside CA	St. Louis, MO Atlanta, GA Washington DC	No A/C	>20
Research Triangle Park	Research Triangle Park, NC	Atlanta, GA Washington DC	Central or Room A/C	<=10
Research Triangle Park	Research Triangle Park, NC	Atlanta, GA Washington DC	Central or Room A/C	10-20
Research Triangle Park	Research Triangle Park, NC	Atlanta, GA Washington DC	Central or Room A/C	20-25
Research Triangle Park	Research Triangle Park, NC	Atlanta, GA Washington DC	Central or Room A/C	>25

The GM and GSD values that define the fitted log-normal distributions for these 39 AER subsets are shown in Table D-4. Examples of these pairs are also plotted in Figures D-9 through D-19, to be further described below. Each of the example figures D-9 through D-19 corresponds to a single GM/GSD “Original Data” pair. The GM and GSD values for the “Original Data” are at the intersection of the horizontal and vertical lines that are parallel to the x- and y-axes in the figures.

Table D-4. Geometric means and standard deviations for AER subsets by city, A/C type, and temperature range.

Subset City Name	A/C Type	Temperature Range (°C)	N	Geometric Mean	Geometric Standard Deviation
Houston	Central or Room A/C	<=20	15	0.4075	2.1135
Houston	Central or Room A/C	20-25	20	0.4675	1.9381
Houston	Central or Room A/C	25-30	65	0.4221	2.2579
Houston	Central or Room A/C	>30	14	0.4989	1.7174
Houston	No A/C	<=10	13	0.6557	1.6794
Houston	No A/C	10-20	28	0.6254	2.9162
	No A/C	>20	12	0.9161	2.4512
Inland California	Central or Room A/C	<=25	226	0.5033	1.9210
Inland California	Central or Room A/C	>25	83	0.8299	2.3534
Inland California	No A/C	<=10	17	0.5256	3.1920
Inland California	No A/C	10-20	52	0.6649	2.1743
Inland California	No A/C	20-25	13	1.0536	1.7110
Inland California	No A/C	>25	14	0.8271	2.2646
Los Angeles	Central or Room A/C	<=20	721	0.5894	1.8948
Los Angeles	Central or Room A/C	20-25	273	1.1003	2.3648
Los Angeles	Central or Room A/C	25-30	102	0.8128	2.4151
Los Angeles	Central or Room A/C	>30	12	0.2664	2.7899
Los Angeles	No A/C	<=10	18	0.5427	3.0872
Los Angeles	No A/C	10-20	390	0.7470	2.0852
Los Angeles	No A/C	20-25	148	1.3718	2.2828
Los Angeles	No A/C	>25	25	0.9884	1.9666
New York City	Central or Room A/C	<=10	20	0.7108	2.0184
New York City	Central or Room A/C	10-25	42	1.1392	2.6773
New York City	Central or Room A/C	>25	19	1.2435	2.1768
New York City	No A/C	<=10	48	1.0165	2.1382
New York City	No A/C	10-20	59	0.7909	2.0417
New York City	No A/C	>20	32	1.6062	2.1189
Outside California	Central or Room A/C	<=10	179	0.9185	1.8589
Outside California	Central or Room A/C	10-20	338	0.5636	1.9396
Outside California	Central or Room A/C	20-25	253	0.4676	2.2011
Outside California	Central or Room A/C	25-30	219	0.4235	2.0373

Table D-4. Geometric means and standard deviations for AER subsets by city, A/C type, and temperature range.

Subset City Name	A/C Type	Temperature Range (°C)	N	Geometric Mean	Geometric Standard Deviation
Outside California	Central or Room A/C	>30	24	0.5667	1.9447
Outside California	No A/C	<=10	61	0.9258	2.0836
Outside California	No A/C	10-20	87	0.7333	2.3299
Outside California	No A/C	>20	44	1.3782	2.2757
Research Triangle Park	Central or Room A/C	<=10	157	0.9617	1.8094
Research Triangle Park	Central or Room A/C	10-20	320	0.5624	1.9058
Research Triangle Park	Central or Room A/C	20-25	196	0.3970	1.8887
Research Triangle Park	Central or Room A/C	>25	145	0.3803	1.7092

To evaluate the uncertainty of the GM and GSD values, a bootstrap simulation was performed, as follows. Suppose that a given AER subset has N values. A bootstrap sample is obtained by sampling N times at random with replacement from the N AER values. The first AER value in the bootstrap sample is selected randomly from the N values, so that each of the N values is equally likely. The second, third, …, N'th values in the bootstrap sample are also selected randomly from the N values, so that for each selection, each of the N values is equally likely. The same value can be selected more than once. Using this bootstrap sample, the geometric mean and geometric standard deviation of the N values in the bootstrap sample was calculated. This pair of values is plotted as one of the points in a figure for that AER subset. 1,000 bootstrap samples were randomly generated for each AER subset, producing a set of 1,000 geometric mean and geometric standard deviation pairs, which were plotted in example Figures D-9 through D-19.

The bootstrap distributions display the part of the uncertainty of the GM and GSD that is entirely due to random sampling variation. The analysis is based on the assumption that the study AER data are a random sample from the population distribution of AER values for the given city, temperature range, and A/C type. On that basis, the 1,000 bootstrap GM and GSD pairs estimate the variation of the GM and GSD across all possible samples of N values from the population. Since each GM, GSD pair uniquely defines a fitted log-normal distribution, the pairs also estimate the uncertainty of the fitted log-normal distribution. The choice of 1,000 was made as a compromise between having enough pairs to accurately estimate the GM, GSD distribution and not having too many pairs so that the graph appears as a smudge of overlapped points. Note that even if there were infinitely many bootstrap pairs, the uncertainty distribution would still be an estimate of the true uncertainty because the N is finite, so that the empirical distribution of the N measured AER values does not equal the unknown population distribution.

In most cases the uncertainty distribution appears to be a roughly circular or elliptical geometric mean and standard deviation region. The size of the region depends upon the sample size and on the variability of the AER values; the region will be smallest when the sample size N is large and/or the variability is small, so that there are a large number of values that are all close together.

The bootstrap analyses show that the geometric standard deviation uncertainty for a given CSA/air-conditioning-status/temperature-range combination tends to have a range of at most from “fitted GSD-1.0 hr^{-1}” to “fitted GSD+1.0 hr^{-1}”, but the intervals based on larger AER sample sizes are frequently much narrower. The ranges for the geometric means tend to be approximately from “fitted GM-0.5 hr^{-1}” to “fitted GM+0.5 hr^{-1}”, but in some cases were much smaller.

The bootstrap analysis only evaluates the uncertainty due to the random sampling. It does not account for the uncertainty due to the lack of representativeness, which in turn is due to the fact that the samples were not always random samples from the entire population of residences in a city, and were sometimes used to represent different cities. Since only the GM and GSD were used, the bootstrap analyses does not account for uncertainties about the true distributional shape, which may not necessarily be log-normal. Furthermore, the bootstrap uncertainty does not account for the effect of the calendar year (possible trends in AER values) or of the uncertainty due to the AER measurement period; the distributions were intended to represent distributions of 24 hour average AER values although the study AER data were measured over a variety of measurement periods.

To use the bootstrap distributions to estimate the impact of sample size on the fitted distributions, a Monte Carlo approach could be used with the APEX model. Instead of using the Original Data distributions, a bootstrap GM, GSD pair could be selected at random and the AER value could be selected randomly from the log-normal distribution with the bootstrap GM and GSD.

Figure D-1

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: Central or Room A/C
Temperature Range: <= 10 Degrees Celsius

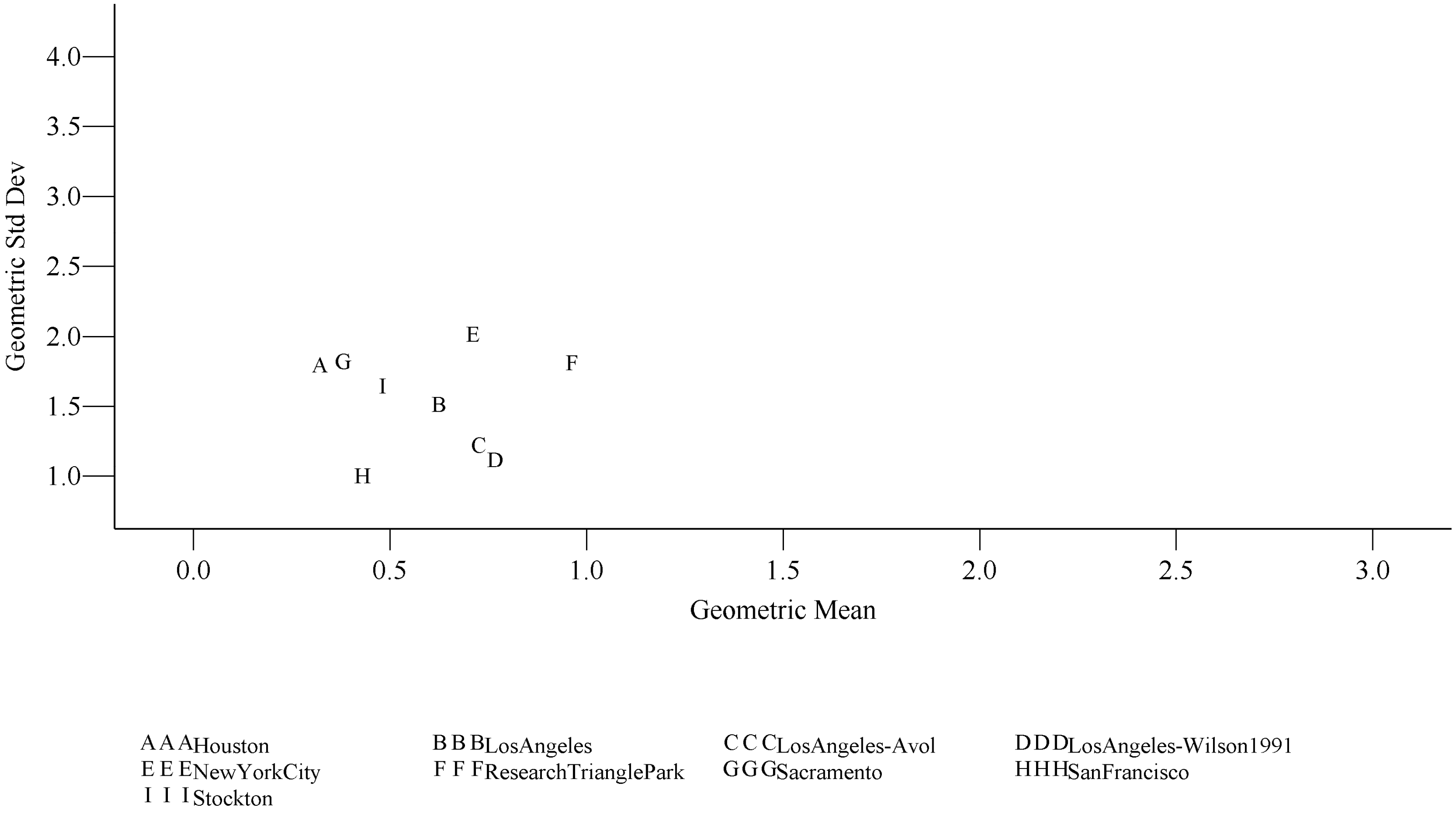


Figure D-2

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: Central or Room A/C
Temperature Range: 10-20 Degrees Celsius

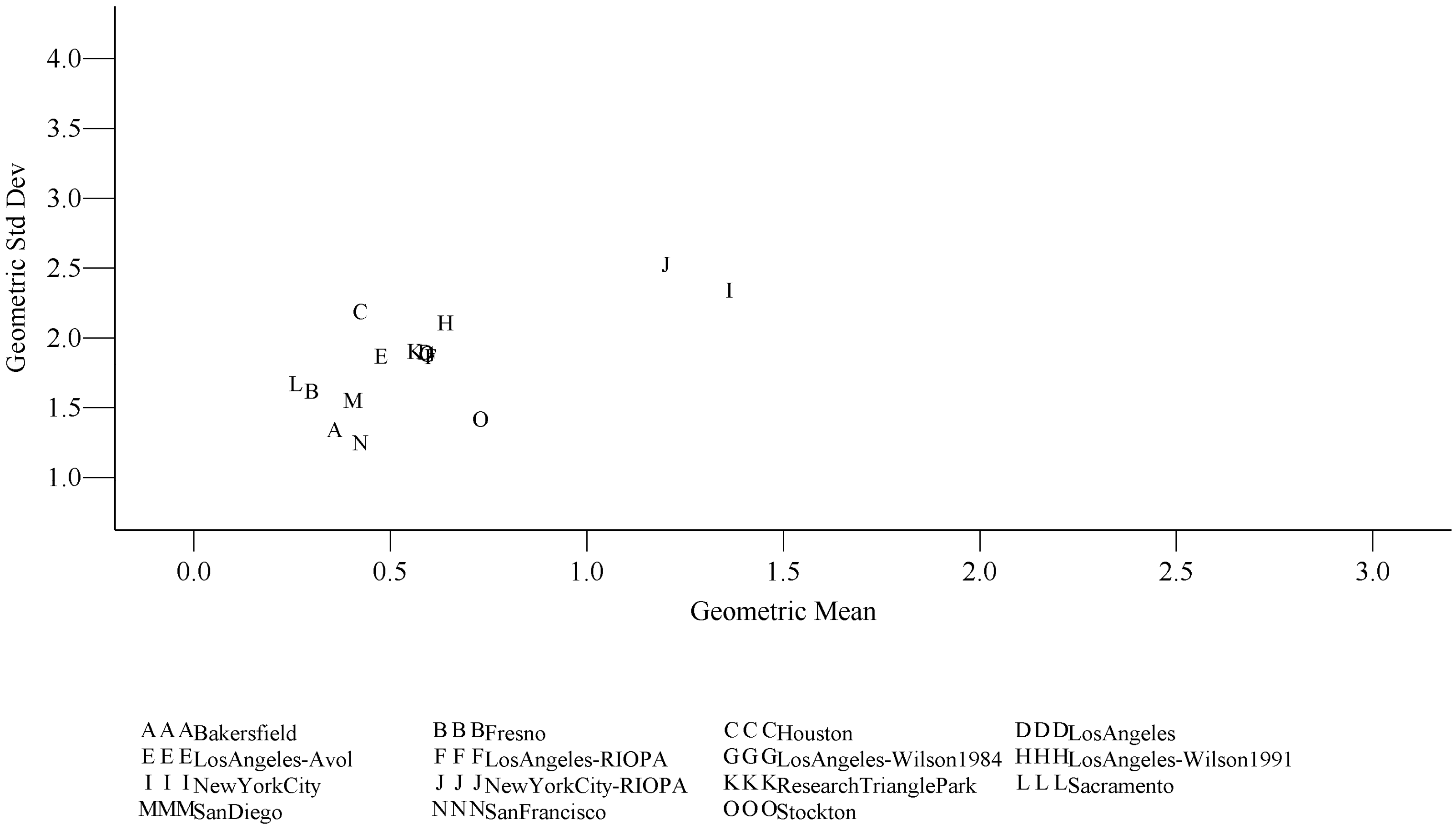


Figure D-3

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: Central or Room A/C
Temperature Range: 20-25 Degrees Celsius

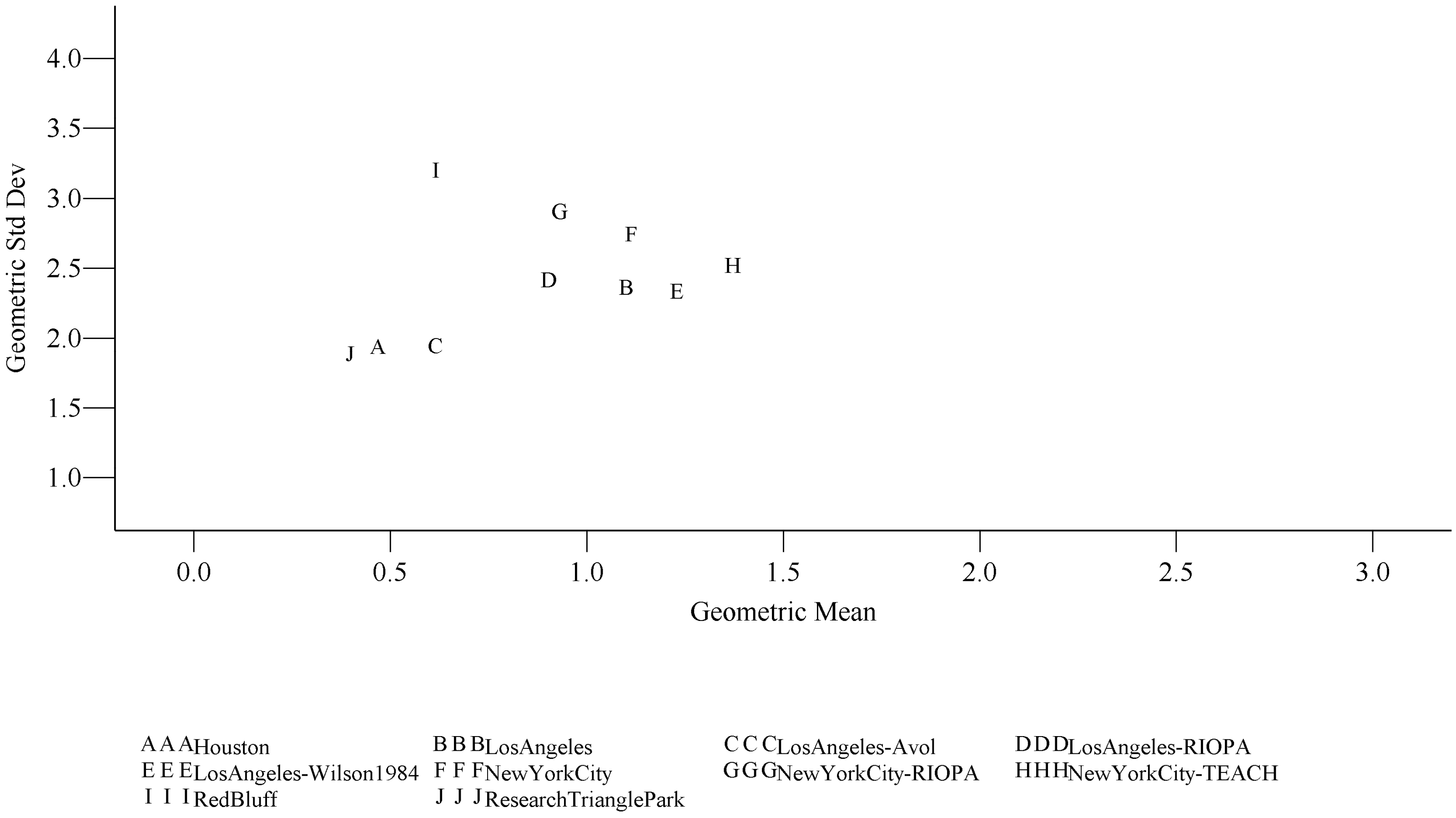


Figure D-4

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: Central or Room A/C
Temperature Range: > 25 Degrees Celsius

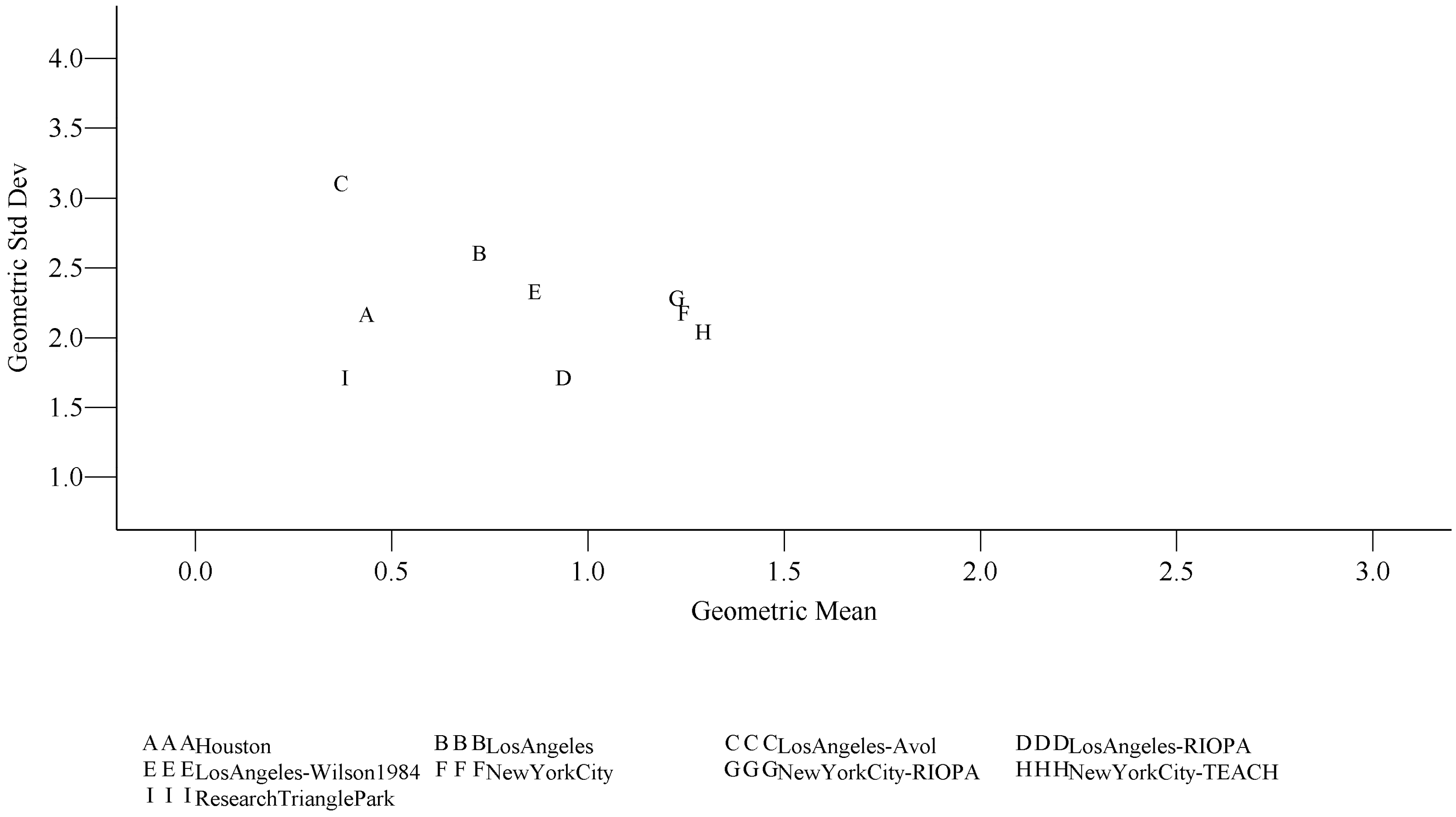


Figure D-5

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: No A/C
Temperature Range: <= 10 Degrees Celsius

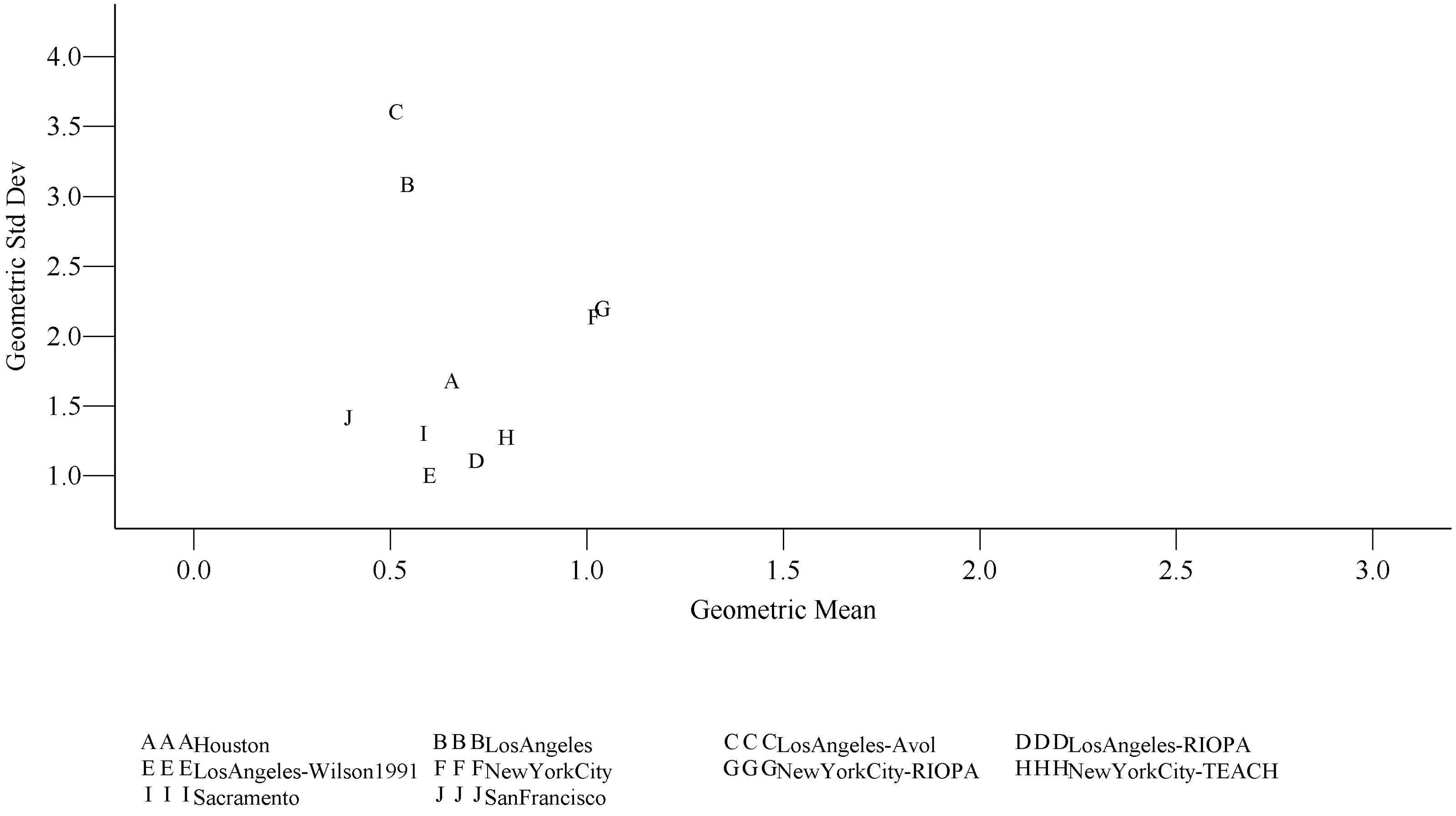


Figure D-6

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: No A/C
Temperature Range: 10-20 Degrees Celsius

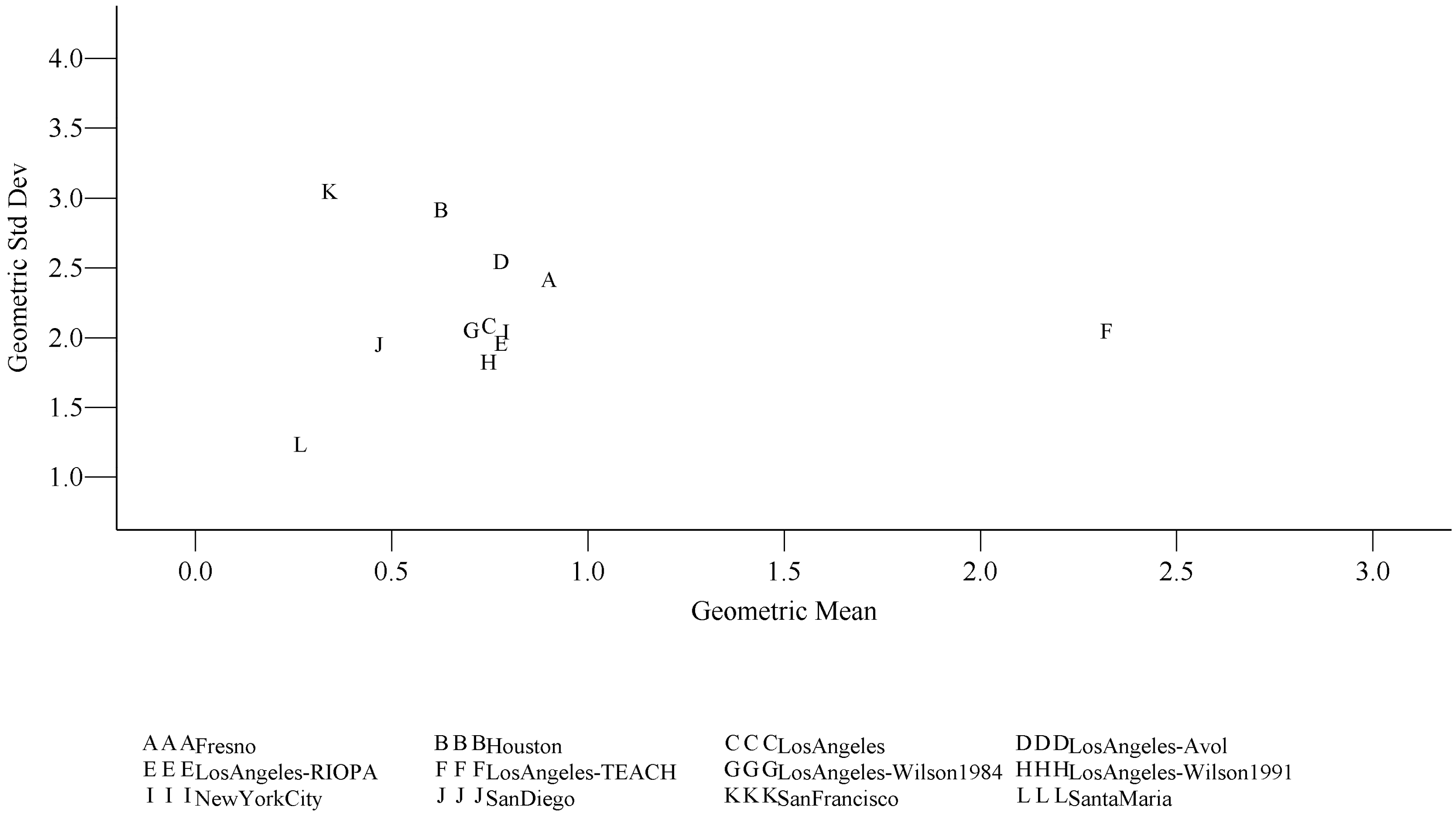


Figure D-7

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: No A/C
Temperature Range: 20-25 Degrees Celsius

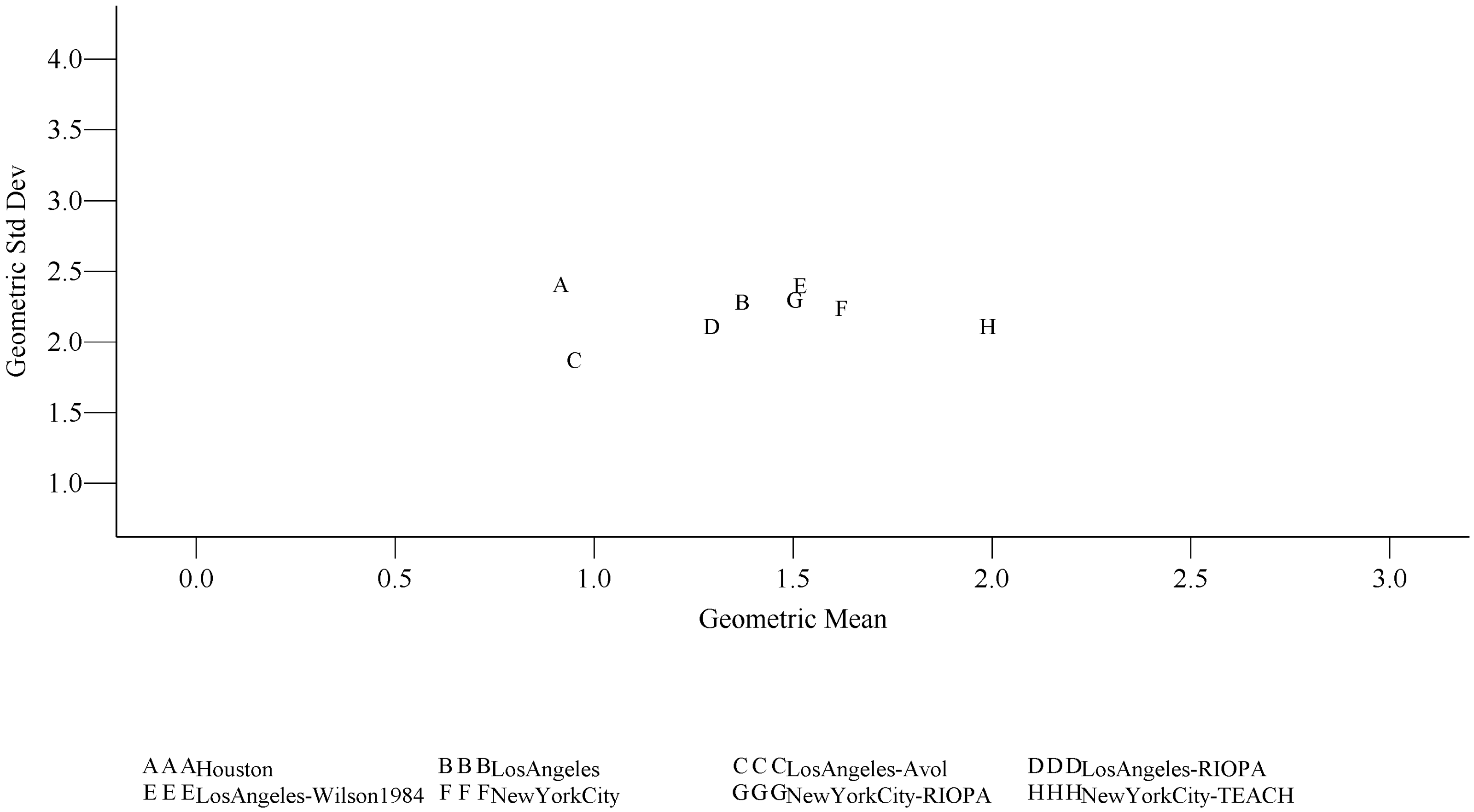


Figure D-8

Geometric mean and standard deviation of air exchange rate
For different cities and studies
Air Conditioner Type: No A/C
Temperature Range: > 25 Degrees Celsius

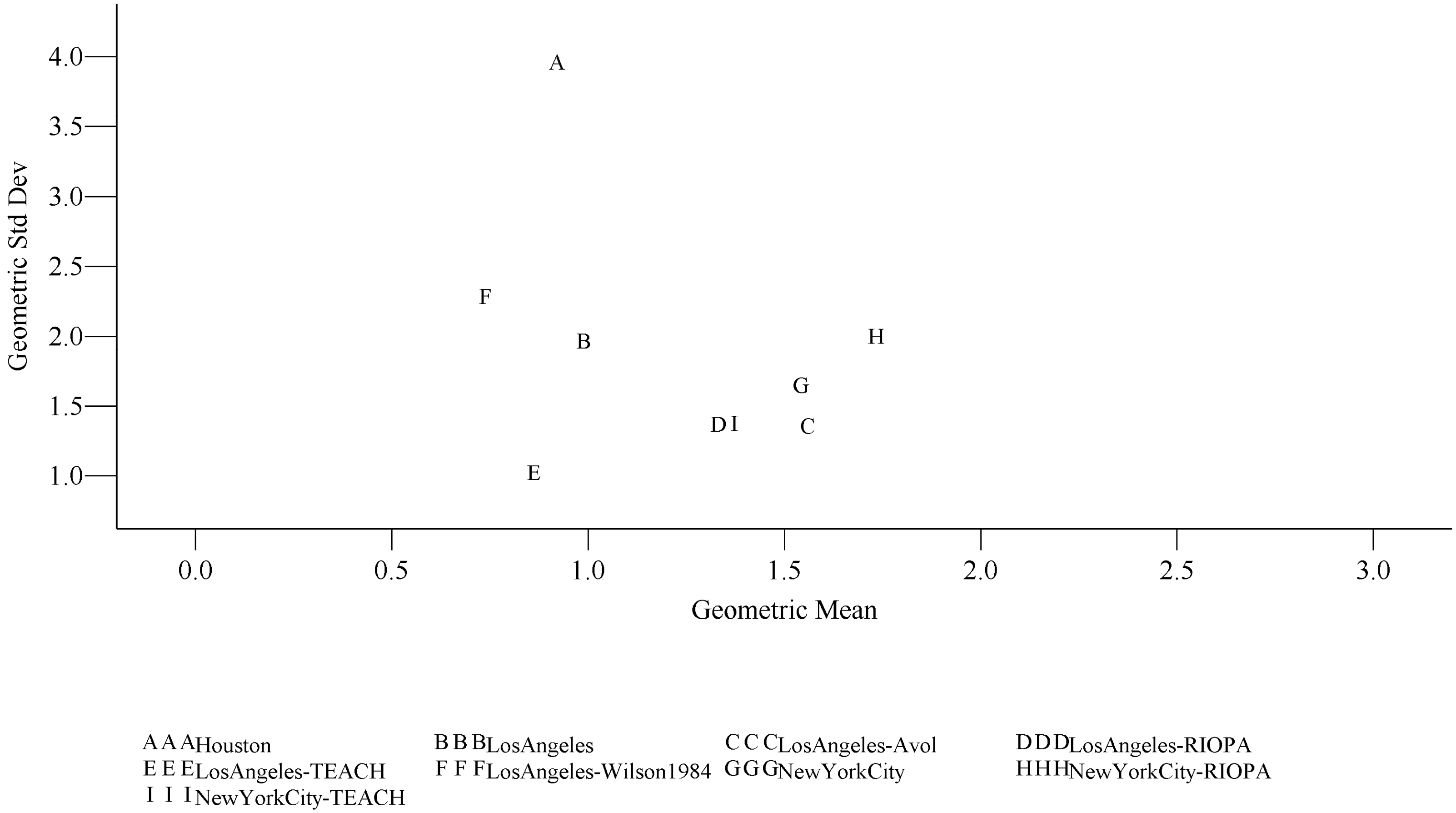


Figure D-9

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Houston
Air Conditioner Type: Central or Room A/C
Temperature Range: 20-25 Degrees Celsius

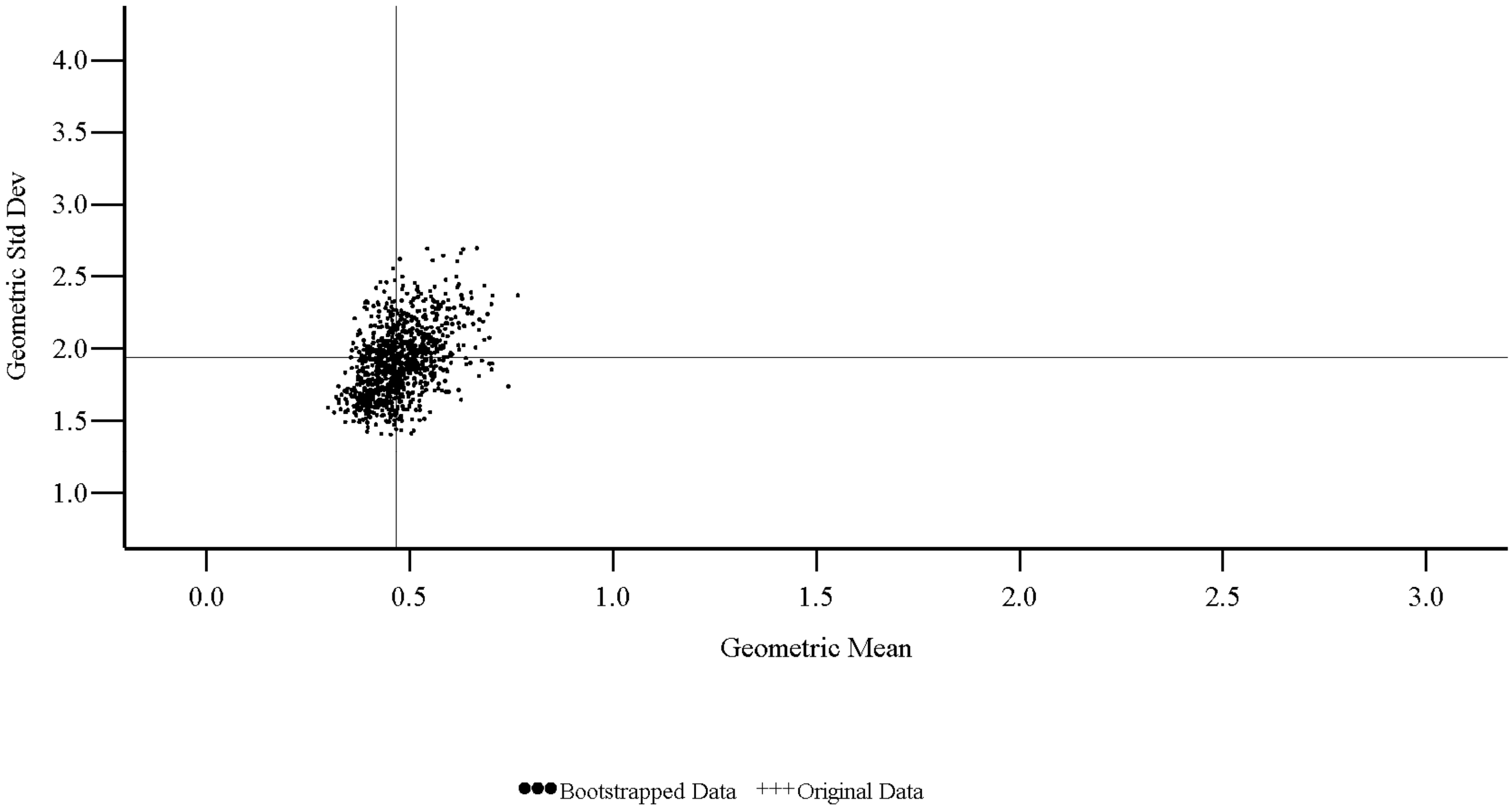


Figure D-10

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Houston
Air Conditioner Type: No A/C
Temperature Range: >20 Degrees Celsius

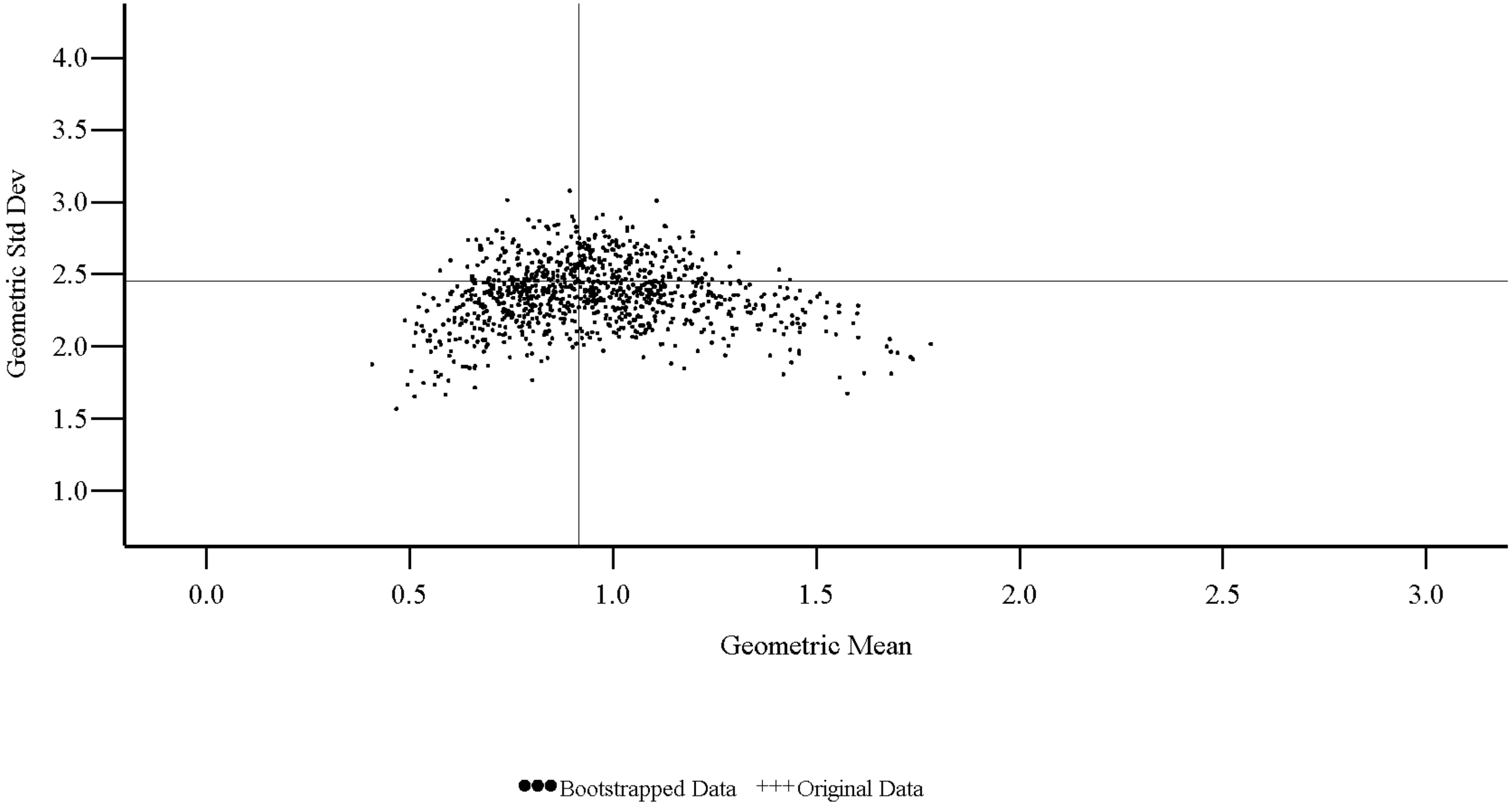


Figure D-11

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Inland California
Air Conditioner Type: Central or Room A/C
Temperature Range: <=25 Degrees Celsius

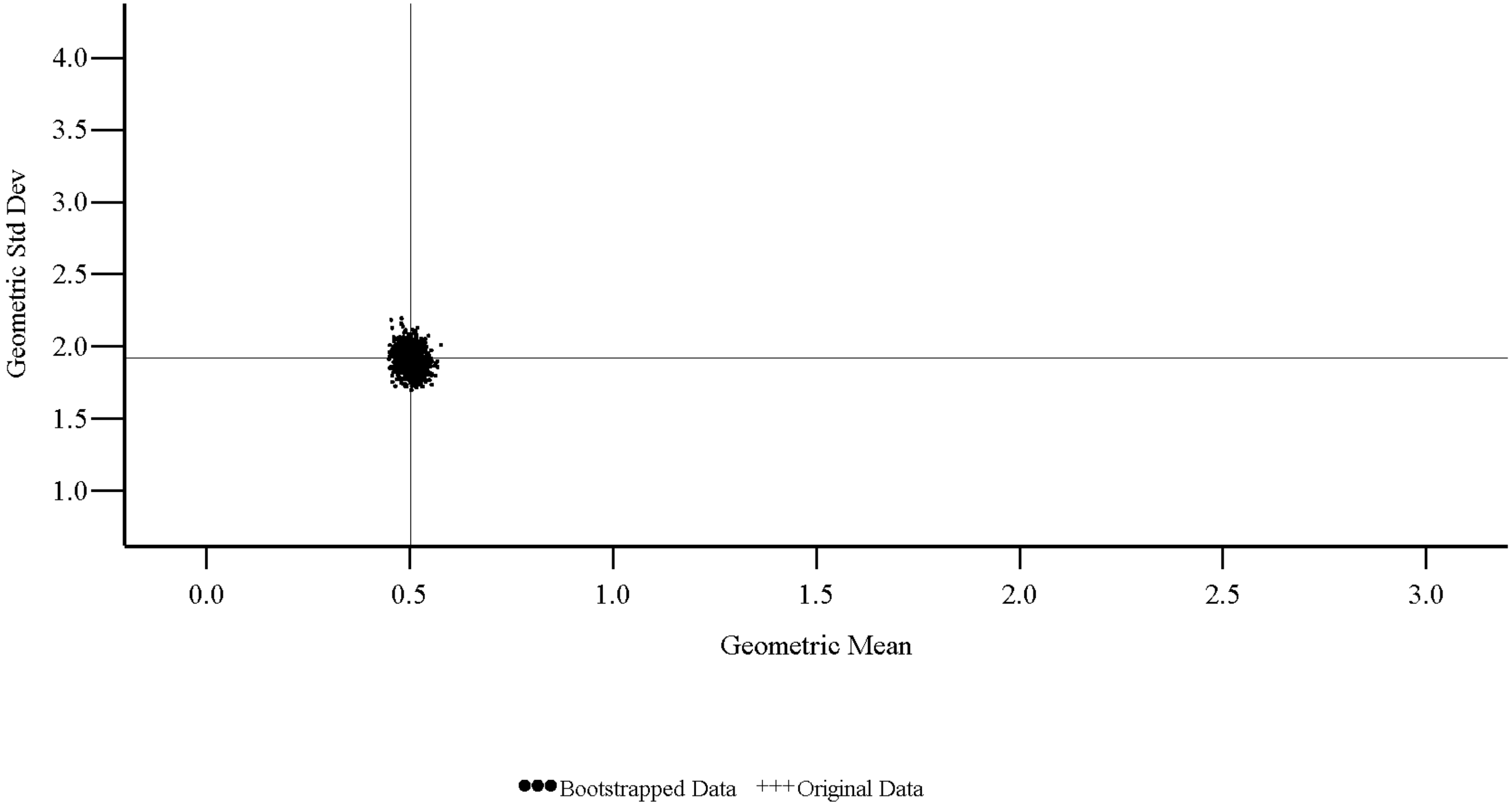


Figure D-12

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Inland California
Air Conditioner Type: No A/C
Temperature Range: 20-25 Degrees Celsius

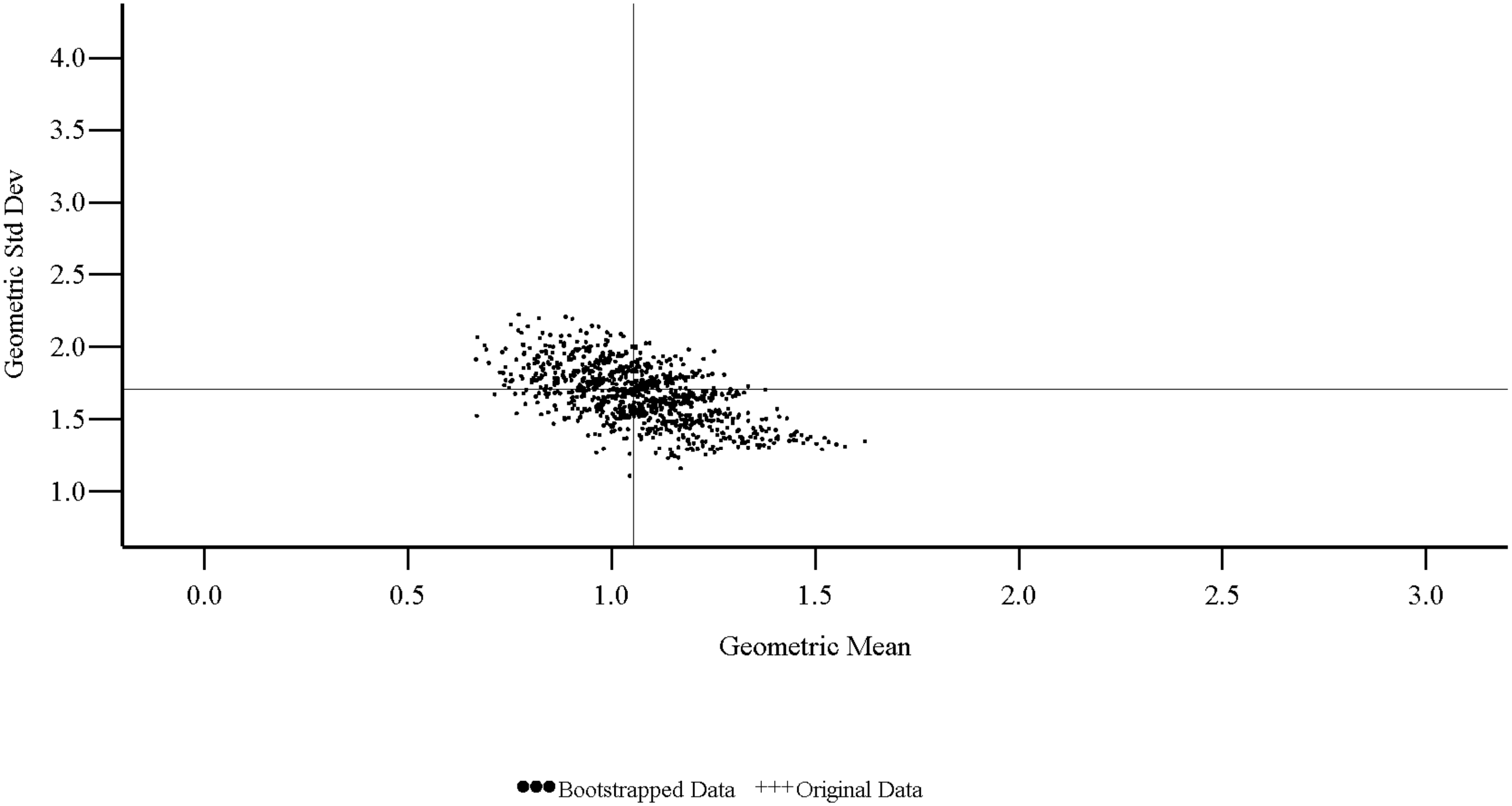


Figure D-13

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Los Angeles
Air Conditioner Type: Central or Room A/C
Temperature Range: 20-25 Degrees Celsius

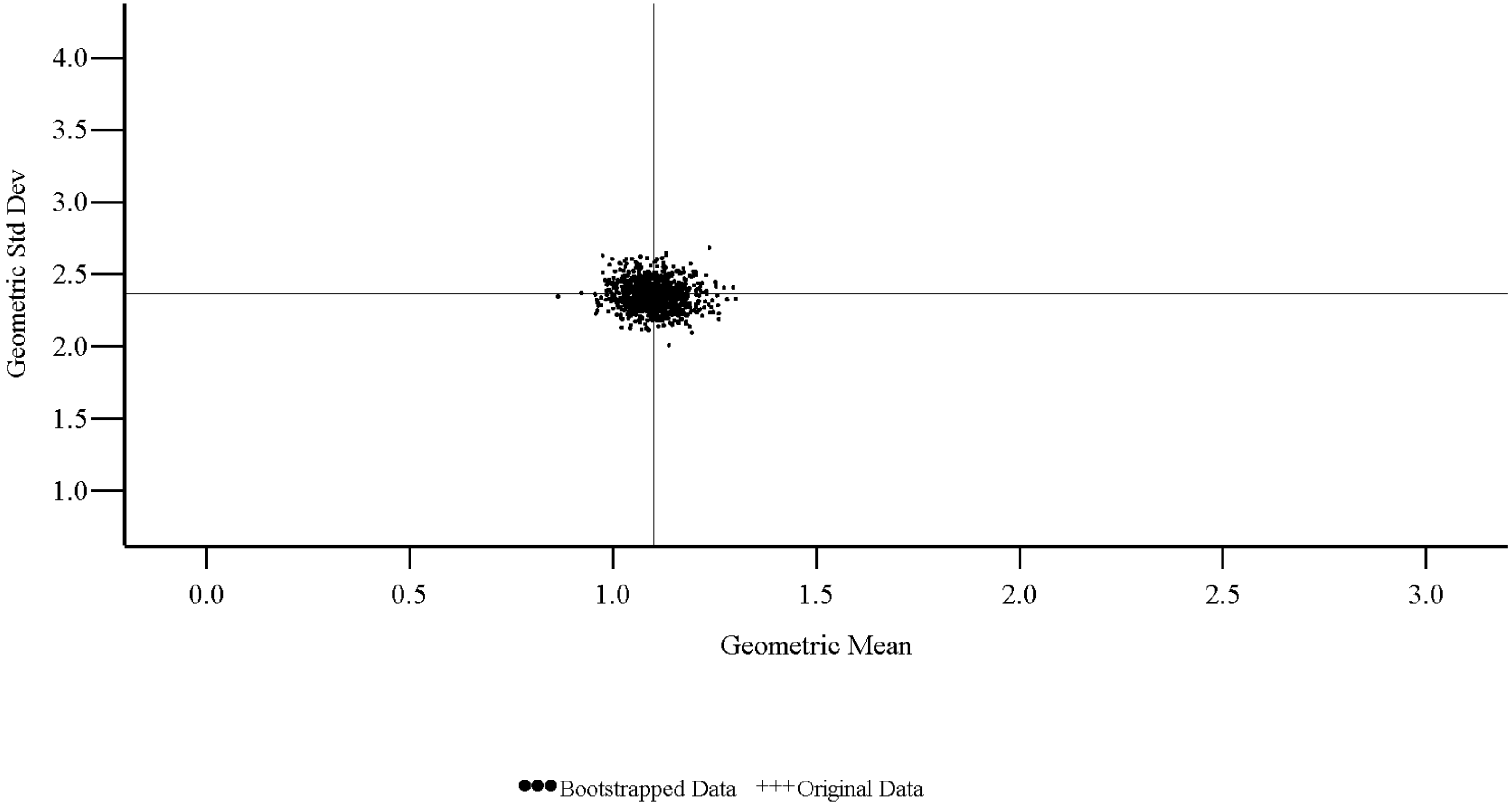


Figure D-14

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Los Angeles
Air Conditioner Type: No A/C
Temperature Range: 20-25 Degrees Celsius

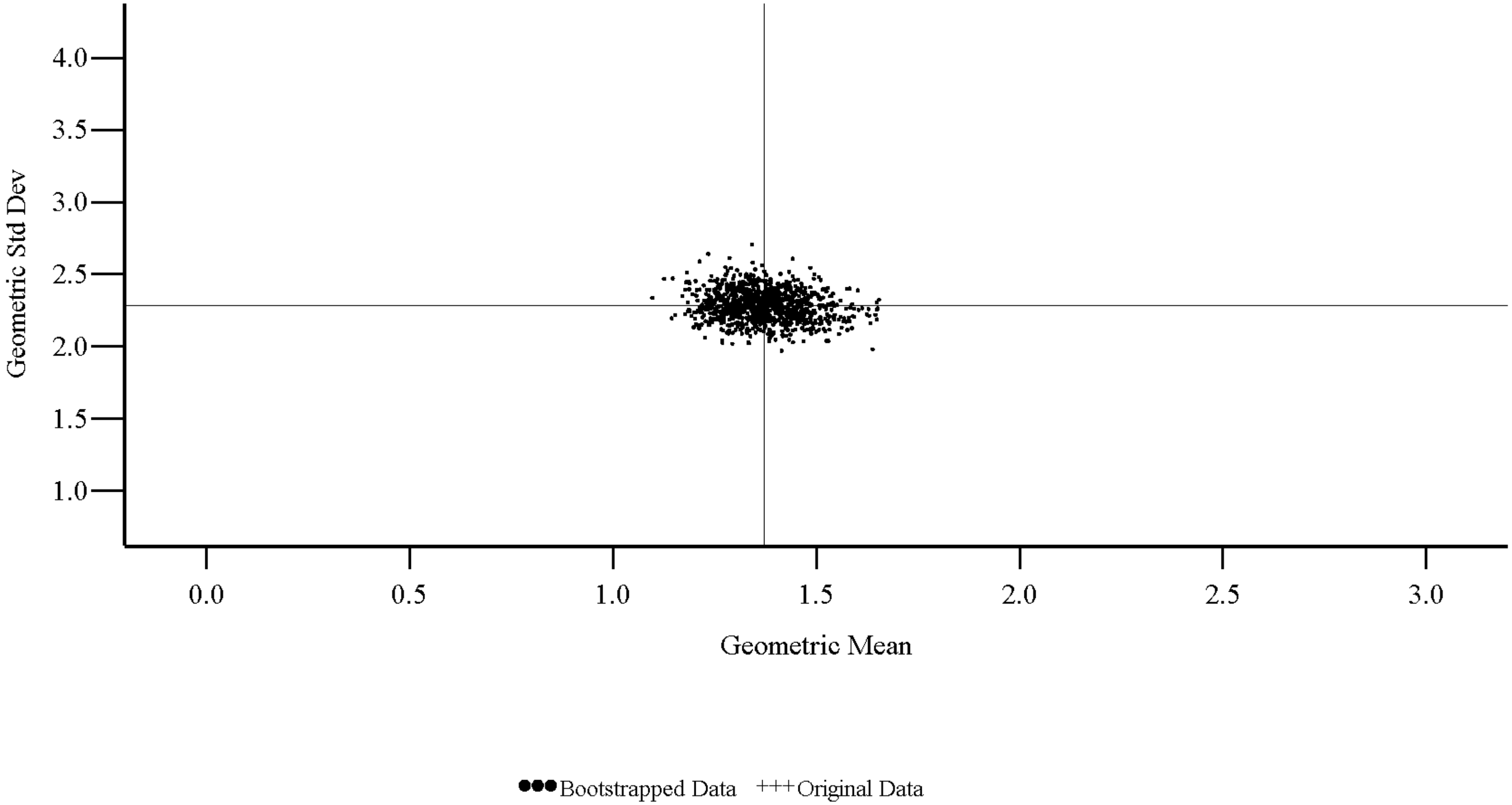


Figure D-15

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: New York City
Air Conditioner Type: Central or Room A/C
Temperature Range: 10-25 Degrees Celsius

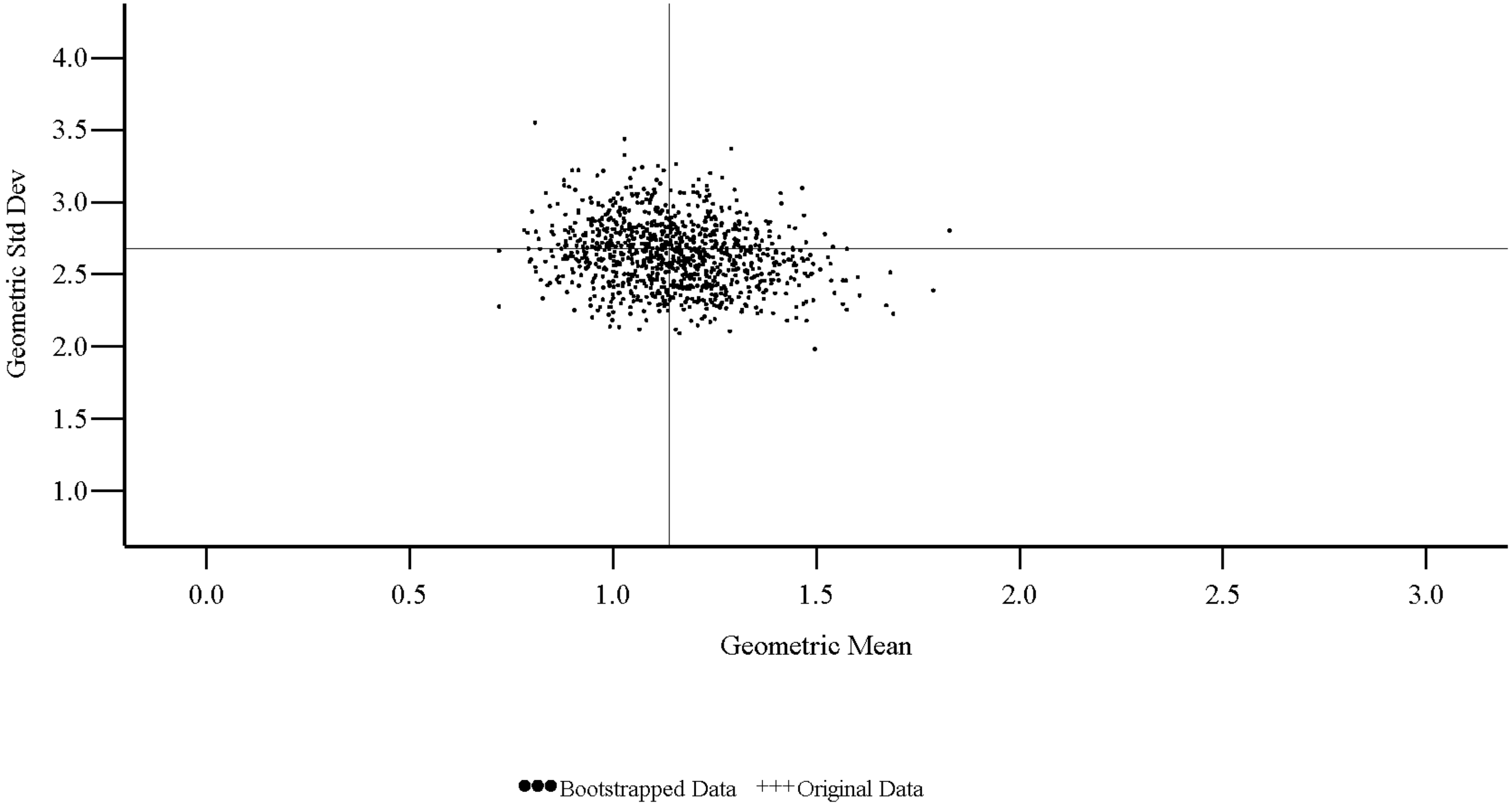


Figure D-16

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: New York City
Air Conditioner Type: No A/C
Temperature Range: >20 Degrees Celsius

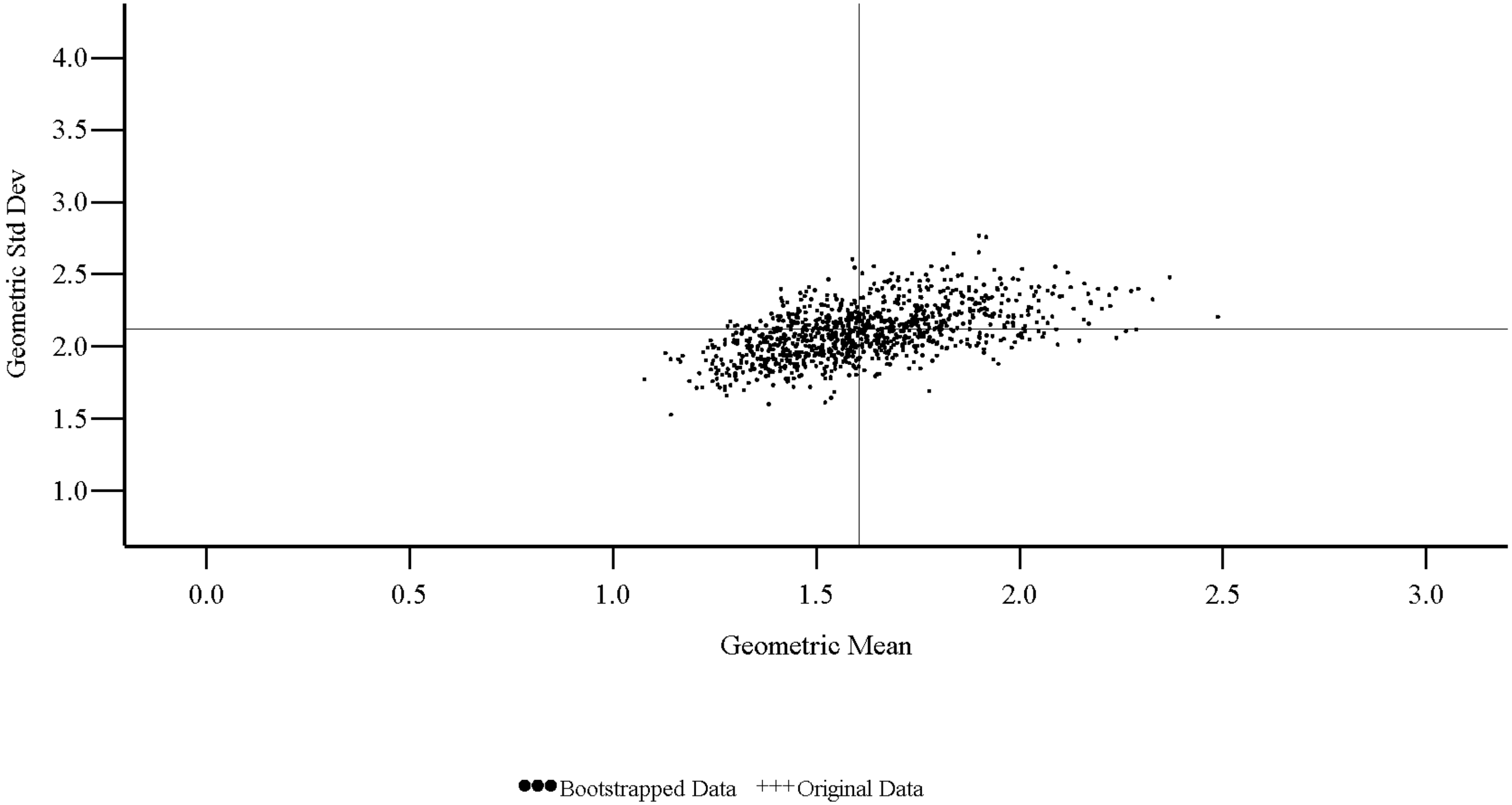


Figure D-17

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Outside California
Air Conditioner Type: Central or Room A/C
Temperature Range: 20-25 Degrees Celsius

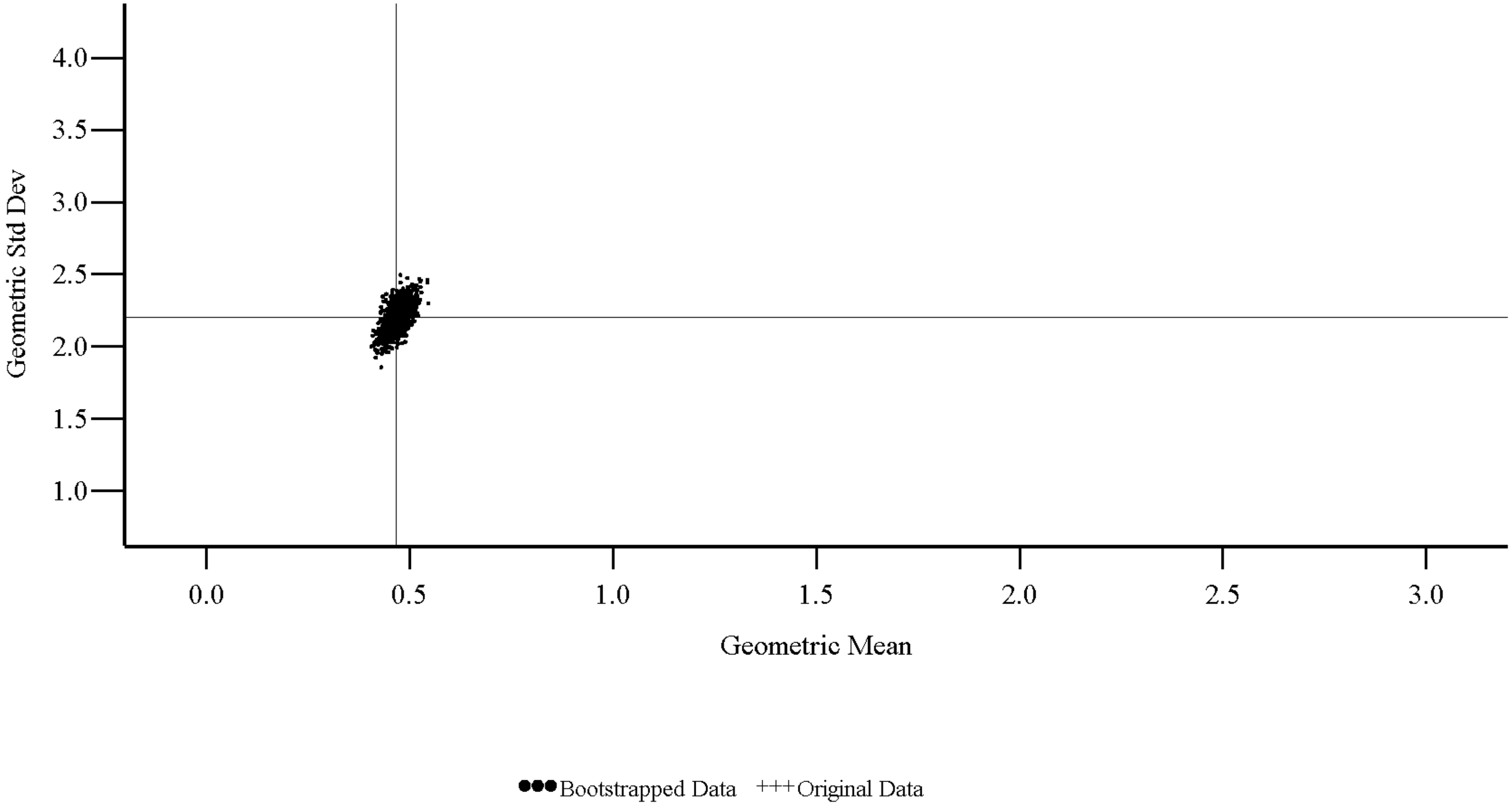


Figure D-18

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Outside California
Air Conditioner Type: No A/C
Temperature Range: >20 Degrees Celsius

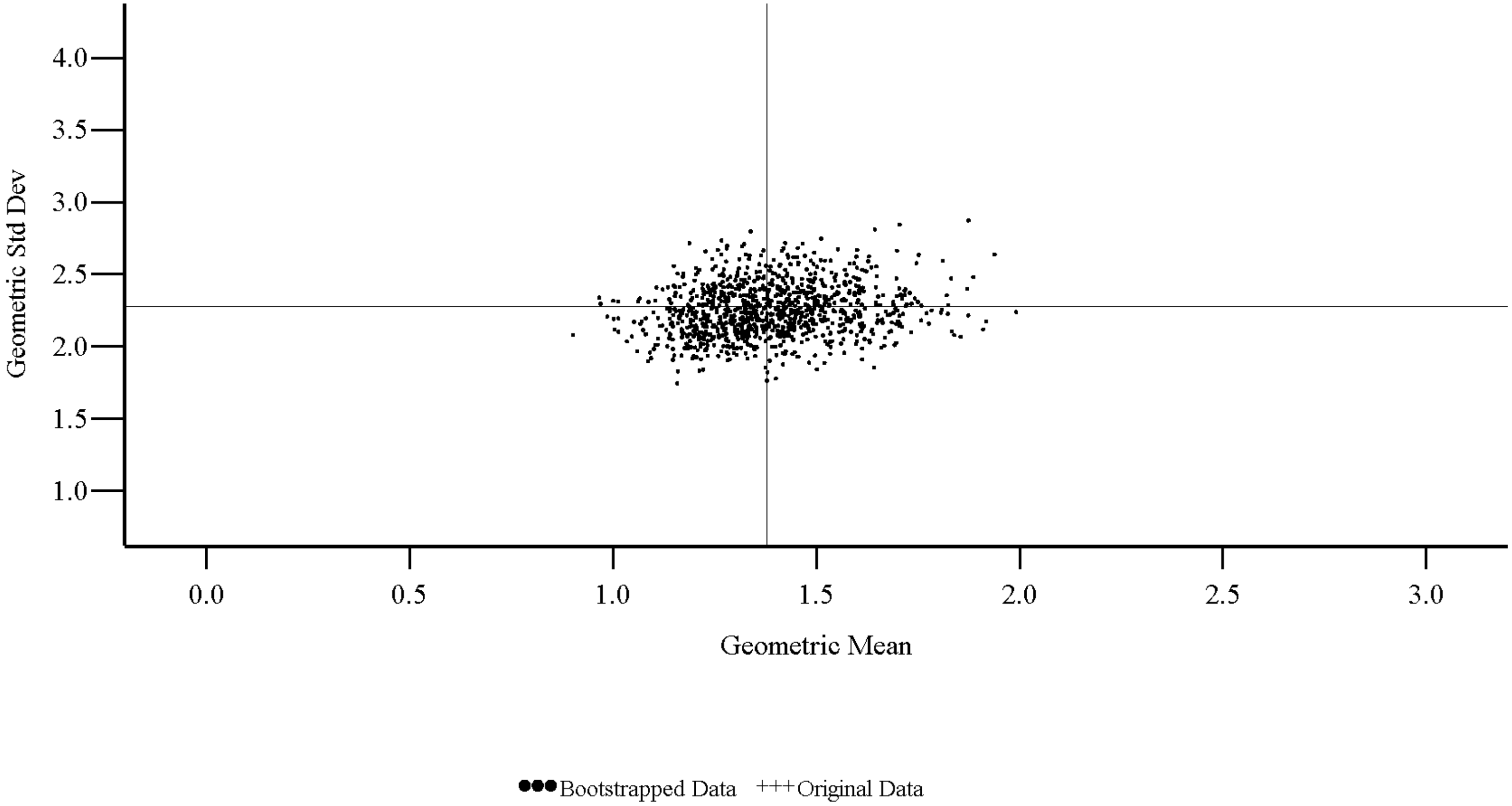
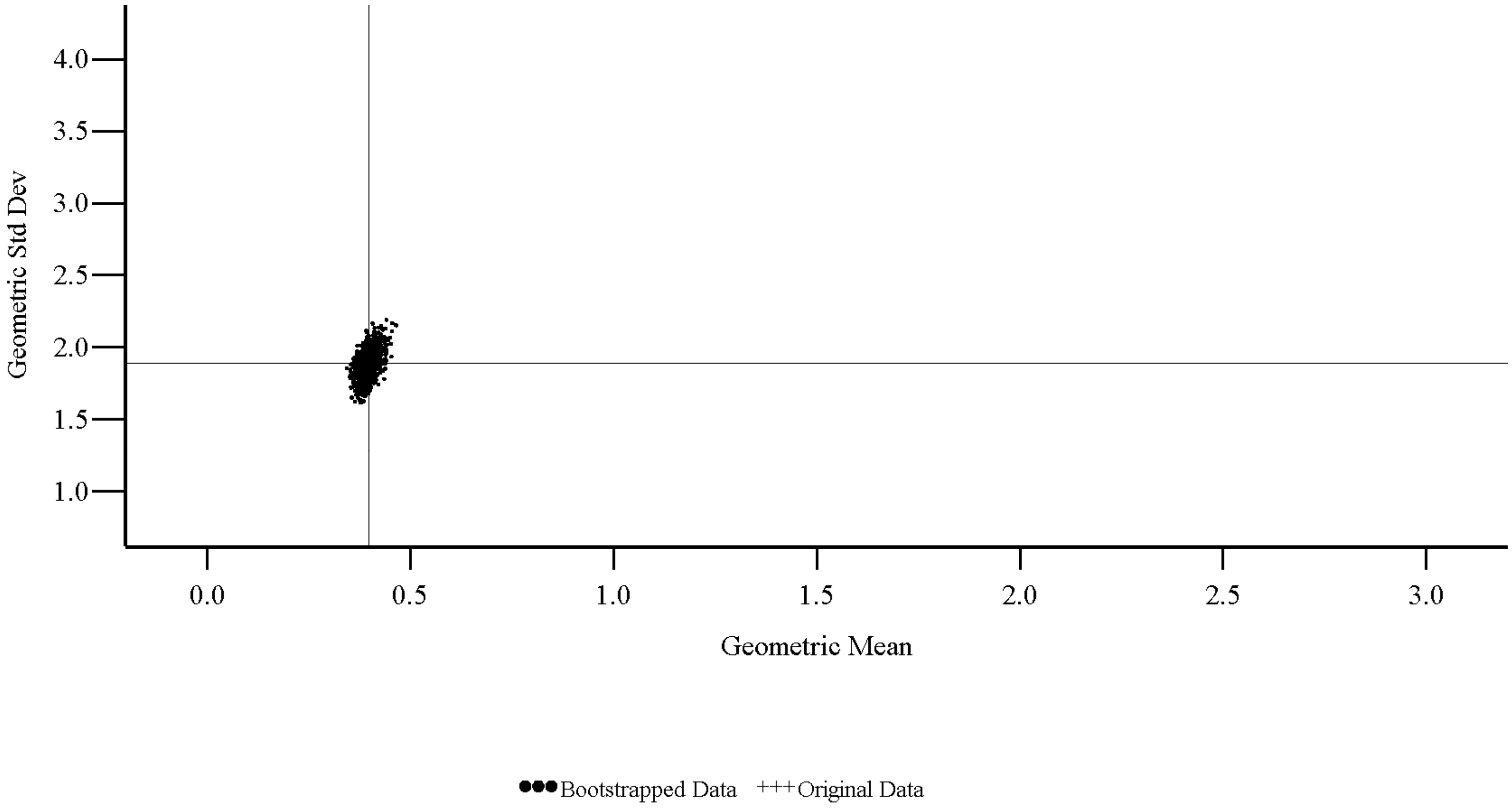


Figure D-19

Geometric mean and standard deviation of air exchange rate
Bootstrapped distributions for different cities
City: Research Triangle Park
Air Conditioner Type: Central or Room A/C
Temperature Range: 20-25 Degrees Celsius



APPENDIX E. DISTRIBUTIONS OF AIR EXCHANGE RATE AVERAGES OVER MULTIPLE DAYS

This page intentionally left blank.

ICF INTERNATIONAL

MEMORANDUM

To: John Langstaff, EPA OAQPS
From: Jonathan Cohen, Arlene Rosenbaum, ICF International
Date: June 8, 2006
Re: Distributions of air exchange rate averages over multiple days

As detailed in the memorandum by Cohen, Mallya and Rosenbaum, 2005[9] (Appendix A of this report) we have proposed to use the APEX model to simulate the residential air exchange rate (AER) using different log-normal distributions for each combination of outside temperature range and the air conditioner type, defined as the presence or absence of an air conditioner (central or room).

Although the averaging periods for the air exchange rates in the study databases varied from one day to seven days, our analyses did not take the measurement duration into account and treated the data as if they were a set of statistically independent daily averages. In this memorandum we present some analyses of the Research Triangle Park Panel Study that show extremely strong correlations between consecutive 24-hour air exchange rates measured at the same house. This provides support for the simplified approach of treating all averaging periods as if they were 24-hour averages.

In the current version of the APEX model, there are several options for stratification of time periods with respect to AER distributions, and for when to re-sample from a distribution for a given stratum. The options selected for this current set of simulations resulted in a uniform AER for each 24-hour period and re-sampling of the 24-hour AER for each simulated day. This re-sampling for each simulated day implies that the simulated AERs on consecutive days in the same microenvironment are statistically independent. Although we have not identified sufficient data to test the assumption of uniform AERs throughout a 24-hour period, the analyses described in this memorandum suggest that AERs on consecutive days are highly correlated. Therefore, we performed sensitivity simulations to assess the impact of the assumption of temporally independent air exchange rates, but found little difference between APEX predictions for the two scenarios (i.e., temporally independent and autocorrelated air exchange rates).

[9] Cohen, J., H. Mallya, and A. Rosenbaum. 2005. Memorandum to John Langstaff. *EPA 68D01052, Work Assignment 3-08. Analysis of Air Exchange Rate Data*. September 30, 2005.

Distributions of multi-day averages from the RTP Panel Study

The RTP Panel study included measurements of 24-hour averages at 38 residences for up to four periods of at least seven days. These periods were in different seasons and/or calendar years. Daily outside temperatures were also provided. All the residences had either window or room air conditioners or both. We used these data to compare the distributions of daily averages taken over 1, 2, 3, .. 7 days.

The analysis is made more complicated because the previous analyses showed the dependence of the air exchange rate on the outside temperature, and the daily temperatures often varied considerably. Two alternative approaches were employed to group consecutive days. For the first approach, A, we sorted the data by the HOUSE_ID number and date and began a new group of days for each new HOUSE_ID and whenever the sorted measurement days on the same HOUSE_ID were 30 days or more apart. In most cases, a home was measured over four different seasons for seven days, potentially giving 38 × 4 = 152 groups; the actual number of groups was 124. For the second approach, B, we again sorted the data by the HOUSE_ID number and date, but this time we began a new group of days for each new HOUSE_ID and whenever the sorted measurement days on the same HOUSE_ID were 30 days or more apart or were for different temperature ranges. We used the same four temperature ranges chosen for the analysis in the Cohen, Mallya, and Rosenbaum, 2005, memorandum (Appendix A): <= 10, 10-20, 20-25, and > 25 ºC. For example, if the first week of measurements on a given HOUSE_ID had the first three days in the <= 10 ºC range, the next day in the 10-20 ºC range, and the last three days in the <= 10 ºC range, then the first approach would treat this as a single group of days. The second approach would treat this as three groups of days, i.e., the first three days, the fourth day, and the last three days. Using the first approach, the days in each group can be in different temperature ranges. Using the second approach, every day in a group is in the same temperature range. Using the first approach we treat groups of days as being independent following a transition to a different house or season. Using the second approach we treat groups of days as being independent following a transition to a different house or season or temperature range.

To evaluate the distributions of multi-day air exchange rate (AER) averages, we averaged the AERs over consecutive days in each group. To obtain a set of one-day averages, we took the AERs for the first day of each group. To obtain a set of two-day averages, we took the average AER over the first two days from each group. We continued this process to obtain three-, four-, five-, six-, and seven-day averages. There were insufficiently representative data for averaging periods longer then seven days. Averages over non-consecutive days were excluded. Each averaging period was assigned the temperature range using the average of the daily temperatures for the averaging period. Using Approach A, some or all of the days in the averaging period might be in different temperature ranges than the overall average. . Using Approach B, every day is in the same temperature range as the overall average. For each averaging period and temperature range, we calculated the mean, standard deviation, and variance of the period average AER and of its natural logarithm. Note than the geometric mean equals *e* raised to the power Mean log (AER) and the geometric standard deviation equals *e* raised to the power Std Dev log (AER). The results are shown in Tables E-1 (Approach A) and E-2 (Approach B).

Table E-1. Distribution of AER averaged over K days and its logarithm. Groups defined by Approach A.

Temperature (ºC)	K	Groups	Mean AER	Mean log(AER)	Std Dev AER	Std Dev log(AER)	Variance AER	Variance log(AER)
<= 10	1	35	1.109	-0.066	0.741	0.568	0.549	0.322
<= 10	2	30	1.149	-0.009	0.689	0.542	0.474	0.294
<= 10	3	28	1.065	-0.088	0.663	0.546	0.440	0.298
<= 10	4	28	1.081	-0.090	0.690	0.584	0.476	0.341
<= 10	5	24	1.103	-0.082	0.754	0.598	0.568	0.358
<= 10	6	24	1.098	-0.083	0.753	0.589	0.567	0.347
<= 10	7	29	1.054	-0.109	0.704	0.556	0.496	0.309
10-20	1	48	0.652	-0.659	0.417	0.791	0.174	0.625
10-20	2	55	0.654	-0.598	0.411	0.607	0.169	0.368
10-20	3	51	0.641	-0.622	0.416	0.603	0.173	0.363
10-20	4	50	0.683	-0.564	0.440	0.619	0.194	0.384
10-20	5	53	0.686	-0.546	0.419	0.596	0.175	0.355
10-20	6	49	0.677	-0.533	0.379	0.544	0.144	0.296
10-20	7	34	0.638	-0.593	0.343	0.555	0.118	0.308
20-25	1	32	0.500	-1.005	0.528	0.760	0.279	0.577
20-25	2	28	0.484	-0.972	0.509	0.623	0.259	0.388
20-25	3	27	0.495	-0.933	0.491	0.604	0.241	0.365
20-25	4	17	0.536	-0.905	0.623	0.652	0.389	0.425
20-25	5	17	0.543	-0.905	0.672	0.649	0.452	0.421
20-25	6	17	0.529	-0.899	0.608	0.617	0.370	0.381
20-25	7	14	0.571	-0.889	0.745	0.683	0.555	0.466
> 25	1	9	0.470	-1.058	0.423	0.857	0.179	0.734
> 25	2	11	0.412	-1.123	0.314	0.742	0.098	0.551
> 25	3	12	0.411	-1.036	0.243	0.582	0.059	0.339
> 25	4	23	0.385	-1.044	0.176	0.429	0.031	0.184
> 25	5	23	0.390	-1.028	0.175	0.425	0.031	0.181
> 25	6	23	0.399	-1.010	0.193	0.435	0.037	0.189
> 25	7	17	0.438	-0.950	0.248	0.506	0.061	0.256

Using both approaches, Tables E-1 and E-2 show that the mean values for the AER and its logarithm are approximately constant for the same temperature range but different averaging periods. This is expected if the daily AER values all have the same statistical distribution, regardless of whether or not they are independent. More interesting is the observation that the standard deviations and variances are also approximately constant for the same temperature range but different averaging periods, except for the data at > 25 ºC where the standard deviations and variances tend to decrease as the length of the averaging period increases. If the daily AER values were statistically independent, then the variance of an average over K days is given by Var / K, where Var is the variance of a single daily AER value. Clearly this formula does not apply. Since the variance is approximately constant for different values of K in the same temperature range (except for the relatively limited data at > 25 ºC), this shows that the daily AER values are strongly correlated. Of course the correlation is not perfect, since otherwise the AER for a given day would be identical to the AER for the next day, if the temperature range were the same, which did not occur.

Table E-2. Distribution of AER averaged over K days and its logarithm. Groups defined by Approach B.

Temperature (ºC)	K	Groups	Mean AER	Mean log(AER)	Std Dev AER	Std Dev log(AER)	Variance AER	Variance log(AER)
<= 10	1	62	1.125	-0.081	0.832	0.610	0.692	0.372
<= 10	2	41	1.059	-0.063	0.595	0.481	0.355	0.231
<= 10	3	32	1.104	-0.040	0.643	0.530	0.413	0.281
<= 10	4	17	1.292	0.115	0.768	0.531	0.590	0.282
<= 10	5	5	1.534	0.264	1.087	0.608	1.182	0.370
10-20	1	109	0.778	-0.482	0.579	0.721	0.336	0.520
10-20	2	81	0.702	-0.532	0.451	0.603	0.204	0.363
10-20	3	63	0.684	-0.540	0.409	0.580	0.167	0.336
10-20	4	27	0.650	-0.626	0.414	0.663	0.171	0.440
10-20	5	22	0.629	-0.660	0.417	0.654	0.174	0.428
10-20	6	12	0.614	-0.679	0.418	0.638	0.175	0.407
10-20	7	6	0.720	-0.587	0.529	0.816	0.280	0.667
20-25	1	107	0.514	-0.915	0.518	0.639	0.269	0.409
20-25	2	63	0.511	-0.930	0.584	0.603	0.341	0.364
20-25	3	23	0.577	-0.837	0.641	0.659	0.411	0.434
20-25	4	3	1.308	-0.484	1.810	1.479	3.277	2.187
> 25	1	54	0.488	-0.949	0.448	0.626	0.201	0.392
> 25	2	32	0.486	-0.900	0.351	0.595	0.123	0.354
> 25	3	23	0.427	-0.970	0.218	0.506	0.048	0.256
> 25	4	12	0.401	-1.029	0.207	0.509	0.043	0.259
> 25	5	12	0.410	-1.003	0.207	0.507	0.043	0.257
> 25	6	6	0.341	-1.164	0.129	0.510	0.017	0.261
> 25	7	6	0.346	-1.144	0.125	0.494	0.016	0.244

These arguments suggest that, based on the RTP Panel study data, to a reasonable approximation, the distribution of an AER measurement does not depend upon the length of the averaging period for the measurement, although it does depend upon the average temperature. This supports the methodology used in the Cohen, Mallya, and Rosenbaum, 2005, analyses that did not take into account the length of the averaging period.

The above argument suggests that the assumption that daily AER values are statistically independent is not justified. Statistical modeling of the correlation structure between consecutive daily AER values is not easy because of the problem of accounting for temperature effects, since temperatures vary from day to day. In the next section we present some statistical models of the daily AER values from the RTP Panel Study.

Statistical models of AER auto-correlations from the RTP Panel Study

We used the MIXED procedure from SAS to fit several mixed models with fixed effects and random effects to the daily values of AER and log(AER). The fixed effects are the population average values of AER or log(AER), and are assumed to depend upon the temperature range. The random effects have expected values of zero and define the correlations between pairs of measurements from the same Group, where the Groups are defined either using Approach A or Approach B above. As described above, a Group is a period of up to 14 consecutive days of measurements at the same house. For these mixed model analyses we included periods with one or more missing days. For all the statistical models, we assume that AER values in different Groups are statistically independent, which implies that data from different houses or in different seasons are independent.

The main statistical model for AER was defined as follows:

$$\begin{aligned} \text{AER} = \quad & \text{Mean(Temp Range)} + \text{A(Group, Temp Range)} \\ & + \text{B(Group, Day Number)} + \text{Error(Group, Day Number)} \end{aligned}$$

Mean(Temp Range) is the fixed effects term. There is a different overall mean value for each of the four temperature ranges.

A(Group, Temp Range) is the random effect of temperature. For each Group, four error terms are independently drawn from four different normal distributions, one for each temperature range. These normal distributions all have mean zero, but may have different variances. Because of this term, there is a correlation between AER values measured in the same Group of days for a pair of days in the same temperature range.

B(Group, Day Number) is the repeated effects term. The day number is defined so that the first day of a Group has day number 1, the next calendar day has day number 2, and so on. In some cases AER's were missing for some of the day numbers. B(Group, Day Number) is a normally distributed error term for each AER measurement. The expected value (i.e., the mean) is zero. The variance is V. The covariance between B(Group g, day i) and B(Group h, day j) is zero for days in different Groups g and h, and equals $V \times \exp(d \times |i-j|)$ for days in the same Group. V and d are fitted parameters. This is a first order auto-regressive model. Because of this term, there is a correlation between AER values measured in the same Group of days, and the correlation decreases if the days are further apart.

Finally, Error(Group, Day Number) is the Residual Error term. There is one such error term for every AER measurement, and all these terms are independently drawn from the same normal distribution, with mean 0 and variance W.

We can summarize this rather complicated model as follows. The AER measurements are uncorrelated if they are from different Groups. If they are in the same Group, they have a correlation that decreases with the day difference, and they have an additional correlation if they are in the same temperature range.

Probably the most interesting parameter for these models is the parameter d, which defines the strength of the auto-correlation between pairs of days. This parameter d lies between -1 (perfect negative correlation) and +1 (perfect positive correlation) although values exactly equal to +1 or -1 are impossible for a stationary model. Negative values of d would be unusual since they would imply a tendency for a high AER day to be followed by a low AER day, and vice versa. The case d=0 is for no auto-correlation.

Table E-3 gives the fitted values of d for various versions of the model. The variants considered were:

- model AER or log(AER)
- include or exclude the term A(Group, Temp Range) (the "random" statement in the SAS code)
- use Approach A or Approach B to define the Groups

Since Approach B forces the temperature ranges to be the same for very day in a Group, the random temperature effect term is difficult to distinguish from the other terms. Therefore this term was not fitted using Approach B.

Table E-3. Autoregressive parameter d for various statistical models for the RTP Panel Study AERs.

Dependent variable	Include A(Group, Temp Range)?	Approach	d
AER	Yes	A	0.80
AER	No	A	0.82
AER	No	B	0.80
Log(AER)	Yes	A	0.87
Log(AER)	No	A	0.87
Log(AER)	No	B	0.85

In all cases, the parameter d is 0.8 or above, showing the very strong correlations between AER measurements on consecutive or almost consecutive days.

Impact of accounting for daily average AER auto-correlation

In the current version of the APEX model, there are several options for stratification of time periods with respect to AER distributions, and for when to re-sample from a distribution for a given stratum. The options selected for this current set of simulations resulted in a uniform AER for each 24-hour period and re-sampling of the 24-hour AER for each simulated day. This re-sampling for each simulated day implies that the simulated AERs on consecutive days in the same microenvironment are statistically independent. Although we have not identified sufficient data to test the assumption of uniform AERs throughout a 24-hour period, the analyses described in this memorandum suggest that AERs on consecutive days are highly correlated.

Therefore, in order to determine if bias was introduced into the APEX estimates with respect to either the magnitudes or variability of exposure concentrations by implicitly assuming uncorrelated air exchange rates, we re-ran the 2002 base case simulations using the option to not re-sample the AERs. For this option APEX selects a single AER for each microenvironment/stratum combination and uses it throughout the simulation.

The comparison of the two scenarios indicates little difference in APEX predictions, probably because the AERs pertain only to indoor microenvironments and for the base cases most exposure to elevated concentrations occurs in the “other outdoors” microenvironment. Figures E-1 and E-2 below present the comparison for exceedances of 8-hour average concentration during moderate exertion for active person in Boston and Houston, respectively.

Figure E-1

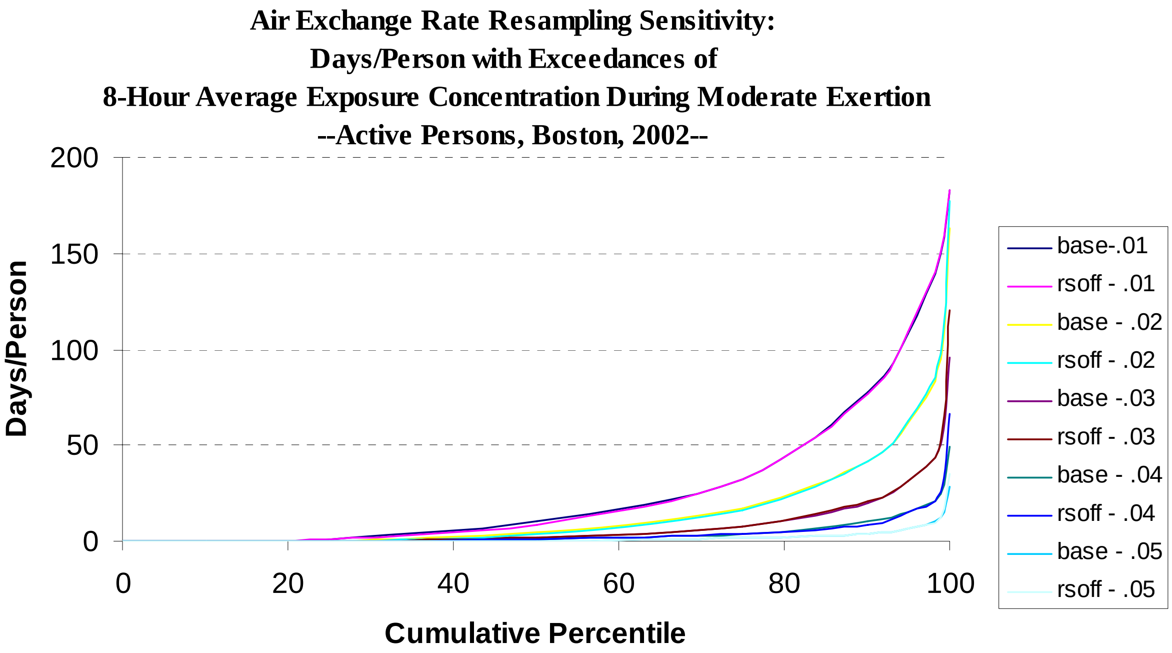
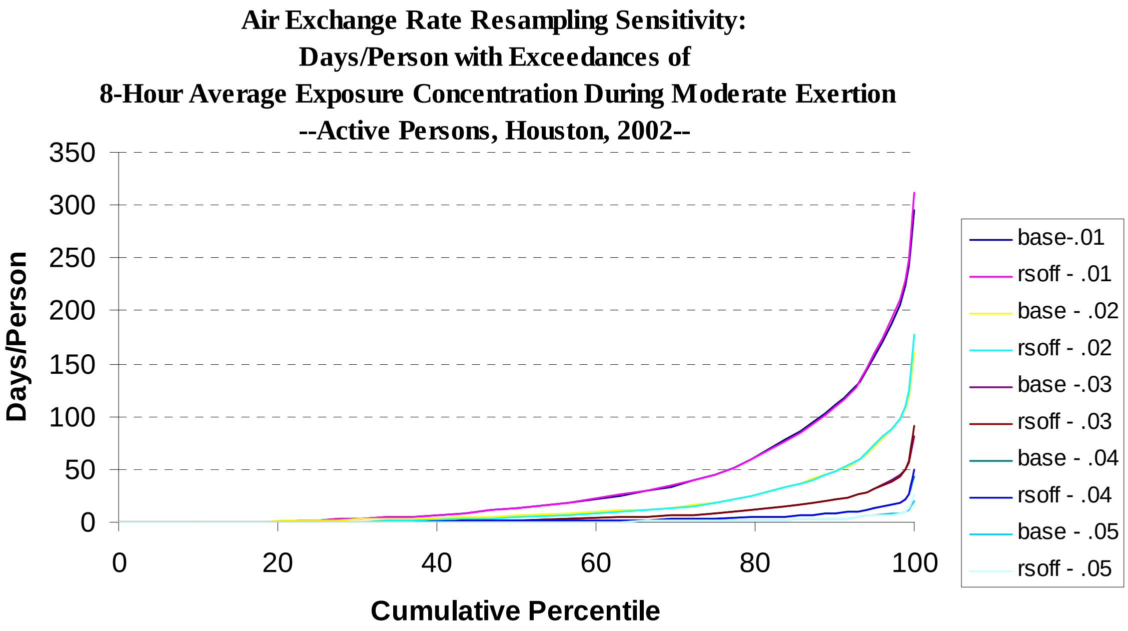


Figure E-2



APPENDIX F. PREVALENCE OF RESIDENTIAL AIR CONDITIONERS AND THE EFFECTS OF SWAMP COOLERS

This page intentionally left blank.

ICF INTERNATIONAL

MEMORANDUM

To: John Langstaff, EPA OAQPS
From: Jonathan Cohen, Arlene Rosenbaum, ICF International
Date: June 6, 2006
Re: Prevalence of residential air conditioners and the effects of swamp coolers

---

This memorandum describes our analysis of the prevalence of air conditioners in the 12 cities being modeled using the APEX. As detailed in the memorandum by Cohen, Mallya and Rosenbaum, 2005[10] (Appendix A of this report) we have proposed to use the APEX model to simulate the residential air exchange rate (AER) using different log-normal distributions for each combination of outside temperature range and the air conditioner type, defined as the presence or absence of an air conditioner (central or room). For each modeled city, the presence or absence of an air conditioner is simulated randomly using the probability that a residence has an air conditioner, which is the air conditioner prevalence. Our proposed approach used city-specific data from the American Housing Survey of 2003. In this memorandum we present uncertainty estimates in the form of confidence intervals for the air conditioner prevalence. We compare these with confidence intervals developed from the Energy Information Administration's Residential Energy Consumption Survey of 2001.

Some residences use evaporative coolers, also known as "swamp" coolers, for cooling. Although both the housing surveys specifically exclude swamp coolers from their definitions of an air conditioner, it is plausible that the AER distributions might also depend upon the presence of a swamp cooler. To evaluate this issue, we also present a comparison of the AER distributions with and without swamp coolers using the available data from three of the AER studies.

American Housing Survey air conditioner prevalence

Data from the American Housing Survey for 2003, a continuous Census Bureau survey of selected cities, (http://www.census.gov/hhes/www/housing/ahs/ahs.html) was used to estimate the air conditioner prevalence in the memorandum by Cohen, Mallya, and Rosenbaum, 2005 (Appendix A of this report). The survey questions ask whether the housing unit has central and room air conditioners, central air conditioners only, room air conditioners only, or no air conditioners. The following definition was used to define air conditioning. Note that evaporative or "swamp" coolers are specifically excluded.

[10] Cohen, J., H. Mallya, and A. Rosenbaum. 2005. Memorandum to John Langstaff. *EPA 68D01052, Work Assignment 3-08. Analysis of Air Exchange Rate Data*. September 30, 2005.

> “*Air conditioning.* Air conditioning is defined as the cooling of air by a refrigeration unit; excluded are evaporative coolers, fans, or blowers that are not connected to a refrigeration unit. A room air-conditioning unit is an individual air conditioner that is installed in a window or an outside wall and generally intended to cool one room, although it may sometimes be used to cool several rooms. A central system is a central installation that air conditions the entire housing unit or major portions of it. In an apartment building, a central system may cool all apartments in the building; each apartment may have its own central system; or there may be several systems, each providing central air conditioning for a group of apartments. A central installation with individual room controls is a central air-conditioning system.[11]”

Table F-1 shows the prevalence estimates calculated using the data and survey weights from the American Housing Survey, together with 95 % confidence intervals. The confidence intervals were computed using the formulas provided in the “Source and Accuracy Statement for the 2003 AHS-N Data Chart” (http://www.census.gov/hhes/www/housing/ahs/03dtchrt/source.html):

$$\text{Standard Error}\,(P) = \sqrt{\frac{3850\,P\,(1-P)}{N}},$$

$$\text{Confidence Interval}\,(P) = P \pm 1.96 \times \text{Standard Error}\,(P)$$

where P is the estimated percentage and N is the estimated total number of housing units.

Table F-1. Prevalence estimates from the American Housing Survey with 95 % Confidence Intervals.

AHS Survey	Percentage (A/Cs)	Housing Units	Std Error	Lower Bound	Upper Bound
Atlanta, 2003	97.01	797,687	1.18	94.69	99.33
Boston, 2003	85.23	1,056,874	2.14	81.03	89.43
Chicago, 2003	87.09	2,253,540	1.39	84.37	89.81
Cleveland, 2003	74.64	637,081	3.38	68.01	81.27
Detroit, 2003	81.41	1,877,178	1.76	77.96	84.86
Houston, 2003	98.7	1,106,268	0.67	97.39	100.01
Los Angeles, 2003	55.05	3,296,819	1.70	51.72	58.38
New York, 2003	81.57	3,575,019	1.27	79.08	84.06
Philadelphia, 2003	90.61	1,943,492	1.30	88.07	93.15
Sacramento, 2003	94.63	524,252	1.93	90.84	98.42
St. Louis, 2003	95.53	592,086	1.67	92.26	98.80
Washington DC, 2003	96.47	1,305,811	1.00	94.51	98.43

[11] Codebook for the American Housing Survey, Public Use File: 1997 and later. Version 1.77. December 2004.

Residential Energy Consumption Survey air conditioner prevalence

The Energy Information Administration's Residential Energy Consumption Survey of 2001 provides estimates of air conditioner prevalence for four states, nine Census Divisions and four Census Regions. Summary data were obtained from the website http://www.eia.doe.gov/emeu/recs/recs2001_hc/2001tblhp.html using the "Appliances" tables HC5-7b, HC5-9b, HC5-10b, HC5-11b, and HC5-12b. This survey uses the following definition of air conditioning (see http://www.eia.doe.gov/glossary/glossary_a.htm):

> **"Air conditioning:** Cooling and dehumidifying the air in an enclosed space by use of a refrigeration unit powered by electricity or natural gas. *Note*: Fans, blowers, and evaporative cooling systems ("swamp coolers") that are not connected to a refrigeration unit are excluded."

Note again that evaporative or "swamp" coolers are specifically excluded.

The relevant data extracted from the "Appliances" tables are presented in Table F-2. The RECS summary tables provide separate prevalence estimates for the households that have air conditioners and for the households that use their air conditioners. The estimates include the small number of households where the fuel for central air-conditioning equipment was something other than electricity.

Table F-2. Prevalence estimates from RECS 2001.

Area (US, State, Census Division, or Census Region)	Has A/C			Uses A/C	
	Prevalence (%)	Row RSE Factor (Has)	Column RSE Factor (Has or Uses)	Prevalence (%)	Row RSE Factor (Uses)
US	77.5	2.4	0.4	75.5	2.6
NY	68.6	2.4	1.2	66.5	2.6
CA	48.3	2.4	1.1	41.8	2.6
TX	96.7	2.4	1.4	95.9	2.6
FL	98.0	2.4	1.3	95.7	2.6
North East	71.6	3.3	1.0	70.2	3.4
Middle Atlantic	76.4	3.3	1.3	74.5	3.4
New England	58.4	3.3	1.6	58.4	3.4
Mid West	83.6	2.2	1.0	82.3	2.3
East North Central	79.8	2.2	1.2	78.5	2.3
West North Central	92.4	2.2	1.5	91.1	2.3
South	95.7	1.4	0.8	94.6	1.6
South Atlantic	95.0	1.4	1.1	93.7	1.6
East South Central	94.3	1.4	1.4	93.5	1.6
West South Central	97.5	1.4	1.5	97.0	1.6
West	45.7	7.1	1.0	41.0	7.2

Table F-2. Prevalence estimates from RECS 2001.

Area (US, State, Census Division, or Census Region)	Has A/C			Uses A/C	
	Prevalence (%)	Row RSE Factor (Has)	Column RSE Factor (Has or Uses)	Prevalence (%)	Row RSE Factor (Uses)
Mountain	50.9	7.1	1.7	47.7	7.2
Pacific	43.6	7.1	1.2	38.3	7.2

The row and column RSE factors are used to calculate 95 % confidence intervals, as follows:

$$\text{RSE} = \text{Relative Standard Error (\%)} = \text{Row RSE Factor} \times \text{Column RSE Factor}$$
$$\text{Standard Error } (P) = \text{RSE} \times P/100,$$
$$\text{Confidence Interval}\,(P) = P \pm 1.96 \times \text{ Standard Error } (P)$$

To apply these results, Table F-3 gives the Census Divisions and Regions for the 12 modeled cities. These Census groupings are defined as groups of States, together with Washington DC included in the South Atlantic Division.

Table F-3. Census Divisions and Regions for 12 cities.

City	Census Division	Census Region
Atlanta	South Atlantic	South
Boston	New England	Northeast
Chicago	East North Central	Midwest
Cleveland	East North Central	Midwest
Detroit	East North Central	Midwest
Houston	West South Central	South
Los Angeles	Pacific	West
New York	Middle Atlantic	Northeast
Philadelphia	Middle Atlantic	Northeast
Sacramento	Pacific	West
St. Louis	West North Central	Midwest
Washington DC	South Atlantic	South

The RECS prevalence estimates and confidence intervals are presented along with the AHS estimates in Tables F-4 and F-5. Table F-4 uses the RECS estimates of households owning air conditioners. Table F-5 uses the slightly lower RECS estimates of households using air conditioners. The agreement depends upon the city. For Boston and Sacramento, the two surveys give extremely different results, the AHS estimates being much higher. Good agreements between the AHS and RECS confidence intervals is found for Atlanta, Cleveland, Detroit, Houston, and Washington DC. Poor agreement (but not as bad as for Boston and Sacramento) with the AHS for either the Census Region or Census Division estimates is found for Chicago, Los Angeles, New York, Philadelphia, and St. Louis.

Since the AHS survey results are city-specific and were based on a more recent survey, we recommend using the AHS prevalence estimates for the APEX modeling.

Air conditioner prevalence versus use.

The approach taken in the Cohen, Mallya, and Rosenbaum, 2005, memorandum (Appendix A of this report) was to stratify the data based on the ownership of an air conditioner. It is very plausible that the AER is more directly related to the actual use of an air conditioner. The Avol and RIOPA studies provided data on the duration of air conditioner use during the AER measurement. However, in order to directly include air conditioner usage in the AER distributions for the APEX model, it would be necessary to know the residential probability of AER usage for each city. The AHS survey did not ask about air conditioner use. The RECS survey asked the following question about central air conditioner usage and also asked the same question about room or wall air conditioner usage:

> *"USECENAC* **Please look at Exhibit F-6. Which of the statements shown best describes the way your household** (if before July 1, insert **will use**; if between July 1 and August 30, insert **uses**; if after August 30, insert **used**) **the central air-conditioning system during the summer of 2001?**
>
> Not used at all 0
> Turned on only a few days or nights when really needed 1
> Turned on quite a bit 2
> Turned on just about all summer 3
> Other .... 5"

It is not clear exactly how these answers was used to define "Air Conditioner Not Used" for the summary tables. Furthermore, the RECS data was not city-specific and was an older survey than the AHS.

We do not recommend directly modeling the AER data as a function of air conditioner use instead of air conditioner ownership in view of the limited data on the prevalence of air conditioner use and the limited AER data where air conditioner use is also recorded.

Table F-4. AHS and RECS air conditioner prevalence estimates. RECS prevalence for households owning air conditioners.*

City	AHS %	AHS Lower	AHS Upper	Division %	Division Lower	Division Upper	Region %	Region Lower	Region Upper
Atlanta	97.01	94.69	99.33	95.00	92.13	97.87	95.70	93.60	97.80
Boston	85.23	81.03	89.43	58.40	52.36	64.44	71.60	66.97	76.23
Chicago	87.09	84.37	89.81	79.80	75.67	83.93	83.60	80.00	87.20
Cleveland	74.64	68.01	81.27	79.80	75.67	83.93	83.60	80.00	87.20
Detroit	81.41	77.96	84.86	79.80	75.67	83.93	83.60	80.00	87.20
Houston	98.70	97.39	100.01	97.50	93.49	101.51	95.70	93.60	97.80
Los Angeles	55.05	51.72	58.38	43.60	36.32	50.88	45.70	39.34	52.06
New York	81.57	79.08	84.06	76.40	69.98	82.82	71.60	66.97	76.23
Philadelphia	90.61	88.07	93.15	76.40	69.98	82.82	71.60	66.97	76.23
Sacramento	94.63	90.84	98.42	43.60	36.32	50.88	45.70	39.34	52.06
St. Louis	95.53	92.26	98.80	92.40	86.42	98.38	83.60	80.00	87.20
Washington DC	96.47	94.51	98.43	95.00	92.13	97.87	95.70	93.60	97.80

*AHS % = AHS estimated prevalence. [AHS Lower, AHS Upper] = 95 % confidence interval for AHS prevalence. Division % = RECS estimated prevalence based on the Census Division. [Division Lower, Division Upper] = 95 % confidence interval for RECS prevalence based on Census Division. Region % = RECS estimated prevalence based on the Census Region. [Region Lower, Region Upper] = 95 % confidence interval for RECS prevalence based on Census Region.

Table F-5. AHS and RECS air conditioner prevalence estimates. RECS prevalence for households using air conditioners.*

City	AHS %	AHS Lower	AHS Upper	Division %	Division Lower	Division Upper	Region %	Region Lower	Region Upper
Atlanta	97.01	94.69	99.33	93.70	90.47	96.93	94.60	92.23	96.97
Boston	85.23	81.03	89.43	58.40	52.17	64.63	70.20	65.52	74.88
Chicago	87.09	84.37	89.81	78.50	74.25	82.75	82.30	78.59	86.01
Cleveland	74.64	68.01	81.27	78.50	74.25	82.75	82.30	78.59	86.01
Detroit	81.41	77.96	84.86	78.50	74.25	82.75	82.30	78.59	86.01
Houston	98.70	97.39	100.01	97.00	92.44	101.56	94.60	92.23	96.97
Los Angeles	55.05	51.72	58.38	38.30	31.81	44.79	41.00	35.21	46.79

Table F-5. AHS and RECS air conditioner prevalence estimates. RECS prevalence for households using air conditioners.*

City	AHS %	AHS Lower	AHS Upper	Division %	Division Lower	Division Upper	Region %	Region Lower	Region Upper
New York	81.57	79.08	84.06	74.50	68.05	80.95	70.20	65.52	74.88
Philadelphia	90.61	88.07	93.15	74.50	68.05	80.95	70.20	65.52	74.88
Sacramento	94.63	90.84	98.42	38.30	31.81	44.79	41.00	35.21	46.79
St. Louis	95.53	92.26	98.80	91.10	84.94	97.26	82.30	78.59	86.01
Washington DC	96.47	94.51	98.43	93.70	90.47	96.93	94.60	92.23	96.97

*AHS % = AHS estimated prevalence. [AHS Lower, AHS Upper] = 95 % confidence interval for AHS prevalence. Division % = RECS estimated prevalence based on the Census Division. [Division Lower, Division Upper] = 95 % confidence interval for RECS prevalence based on Census Division. Region % = RECS estimated prevalence based on the Census Region. [Region Lower, Region Upper] = 95 % confidence interval for RECS prevalence based on Census Region.

Evaporative or “Swamp” Coolers

As discussed above, neither the AHS not the RECS provide estimates of the prevalence of swamp coolers, which were specifically excluded from the definition of air conditioning. This means that an AER model stratified by the presence or absence of swamp coolers is not feasible, unless another survey with swamp cooler prevalence estimates becomes available. Nevertheless, we used the data from the RIOPA (CA, NJ, and TX), Avol (Southern CA), and Wilson 1991 (CA) studies to evaluate the effects of swamp coolers on the AER. The results generally showed no statistically significant differences between the AER distributions for residences with or without swamp coolers, after stratifying by state, city, study, air conditioner type (presence or absence), and temperature range.

First we calculated summary statistics for the AERs for each combination of State (including All), city (including All), study (including All), A/C ownership (presence or absence of a central or room air conditioner), temperature range (<= 10. 10-20, 20-30, > 30 ºC), and swamp cooler ownership. The summary statistics included the average, standard deviation, variance, minimum, maximum, and various percentile values. Summary statistics of the natural logarithms of the AERs were also calculated.

Then we compared the distributions with and without swamp coolers. For each stratum, defined by a combination of State, city, study, A/C ownership, and temperature range, an F-Statistic was calculated to compare the mean values between groups using a one way analysis of variance (ANOVA). This test assumes that the AER or log(AER) values are normally distributed with a mean that may depend upon swamp cooler ownership and a constant variance..

The Kruskal-Wallis Statistics were also computed as a non-parametric test for equal group medians. It is equivalent to the more familiar Wilcoxon test, since there are only two groups compared (presence or absence of swamp coolers). The analysis is valid if the AER minus the group median has the same distribution for each group. (The test is also consistent under weaker assumptions against more general alternatives). Since the logarithm is a strictly increasing function and the test is non-parametric, the Kruskal-Wallis tests give identical results for AER and Log (AER).

In addition the Mood Statistics were computed as a non-parametric test to the scale statistics for each pair of groups. The scale statistic measures variation about the central value, which is a non-parametric generalization of the standard deviation. Specifically, suppose there is a total of N AER or log(AER) values, summing across both groups. These N values are ranked from 1 to N, and the j’th highest value is given a score of $\{j - (N+1)/2\}^2$. The Mood statistic uses a one way ANOVA statistic to compare the total scores for each group. Since the logarithm is a strictly increasing function and the test is non-parametric, the Mood tests give identical results for AER and Log (AER).

For each statistic (i.e., F, Kruskal-Wallis, Mood) a P-value was determined. P-values above 0.05 indicate cases where the two group means are not statistically significantly different at the 5 percent level. Most of the p-values were above 0.05, indicating no significant differences. (There are a few anomalous significant differences for very small sample sizes when the estimated variances are zero). This appears to justify treating a residence with a swamp cooler but no A/C the same as a residence without a swamp cooler and no A/C; and treating a residence with a swamp cooler and with A/C the same as a residence without a swamp cooler but with A/C. In other words, this analysis shows that there is no improvement in the statistical AER model if we also stratify by swamp cooler presence or absence, given that we already stratify by city, air conditioner presence or absence, and temperature range.

This page intentionally left blank.

APPENDIX G. ANALYSIS OF NHIS ASTHMA PREVALENCE DATA

This page intentionally left blank.

DRAFT MEMORANDUM

To: John Langstaff

From: Jonathan Cohen, Arlene Rosenbaum

Date: September 30, 2005

Re: EPA 68D01052, Work Assignment 3-08. Analysis of NHIS Asthma Prevalence Data

This memorandum describes our analysis of children's asthma prevalence data from the National Health Interview Survey (NHIS) for 2003. Asthma prevalence rates for children aged 0 to 17 years were calculated for each age, gender, and region. The regions defined by NHIS are "Midwest," "Northeast," "South," and "West." For this project, asthma prevalence was defined as the probability of a Yes response to the question CASHMEV: "Ever been told that … had asthma?" among those that responded Yes or No to this question. The responses were weighted to take into account the complex survey design of the NHIS survey. Standard errors and confidence intervals for the prevalence were calculated using a logistic model, taking into account the survey design. Prevalence curves showing the variation of asthma prevalence against age for a given gender and region were plotted. A scatterplot smoothing technique using the LOESS smoother was applied to smooth the prevalence curves and compute the standard errors and confidence intervals for the smoothed prevalence estimates. Logistic analysis of the prevalence curves shows statistically significant differences in prevalence by gender and by region. Therefore we did not combine the prevalence rates for different genders or regions.

Logistic Models

NHIS survey data for 2003 were provided by EPA. One obvious approach to calculate prevalence rates and their uncertainties for a given gender, region, and age is to calculate the proportion of Yes responses among the Yes and No responses for that demographic group, weighting each response by the survey weight. Although that approach was initially used, two problems are that the distributions of the estimated prevalence rates are not well approximated by normal distributions, and that the estimated confidence intervals based on the normal approximation often extend outside the [0, 1] interval. A better approach is to use a logistic transformation and fit a model of the form:

$$\text{Prob (asthma)} = \exp(\text{beta}) / (1 + \exp(\text{beta})),$$

where beta may depend on the explanatory variables for age, gender, or region. This is equivalent to the model:

Beta = logit {prob (asthma) } = log { prob (asthma) / [1 – prob (asthma)] }.

The distribution of the estimated values of beta is more closely approximated by a normal distribution than the distribution of the corresponding estimates of prob (asthma). By applying a logit transformation to the confidence intervals for beta, the corresponding confidence intervals for prob (asthma) will always be inside [0, 1]. Another advantage of the logistic modeling is that it can be used to compare alternative statistical models, such as models where the prevalence probability depends upon age, region, and gender, or on age and region but not gender.

A variety of logistic models for asthma prevalence were fit and compared, where the transformed probability variable beta is a given function of age, gender, and region. SAS's SURVEYLOGISTIC procedure was used to fit the logistic models, taking into account the NHIS survey weights and survey design (stratification and clustering).

The following Table G-1 lists the models fitted and their log-likelihood goodness-of-fit measures. 16 models were fitted. The Strata column lists the four possible stratifications: no stratification, by gender, by region, by region and gender. For example, "4. region, gender" means that separate prevalence estimates were made for each combination of region and gender. As another example, "2. gender" means that separate prevalence estimates were made for each gender, so that for each gender, the prevalence is assumed to be the same for each region. The prevalence estimates are independently calculated for each stratum.

Table G-1. Alternative logistic models for asthma prevalence.

Model	Description	Strata	- 2 Log Likelihood	DF
1	1. logit(prob) = linear in age	1. none	54168194.62	2
2	1. logit(prob) = linear in age	2. gender	53974657.17	4
3	1. logit(prob) = linear in age	3. region	54048602.57	8
4	1. logit(prob) = linear in age	4. region, gender	53837594.97	16
5	2. logit(prob) = quadratic in age	1. none	53958021.20	3
6	2. logit(prob) = quadratic in age	2. gender	53758240.99	6
7	2. logit(prob) = quadratic in age	3. region	53818198.13	12
8	2. logit(prob) = quadratic in age	4. region, gender	53593569.84	24
9	3. logit(prob) = cubic in age	1. none	53849072.76	4
10	3. logit(prob) = cubic in age	2. gender	53639181.24	8
11	3. logit(prob) = cubic in age	3. region	53694710.66	16
12	3. logit(prob) = cubic in age	4. region, gender	53441122.98	32

Model	Description	Strata	- 2 Log Likelihood	DF
13	4. logit(prob) = f(age)	1. none	53610093.48	18
14	4. logit(prob) = f(age)	2. gender	53226610.02	36
15	4. logit(prob) = f(age)	3. region	53099749.33	72
16	4. logit(prob) = f(age)	4. region, gender	52380000.19	144

The Description column describes how beta depends upon the age:

Linear in age:	Beta = $\alpha + \beta \times$ age, where α and β vary with the strata.
Quadratic in age:	Beta = $\alpha + \beta \times$ age + $\gamma \times$ age^2 where α β and γ vary with the strata.
Cubic in age:	Beta = $\alpha + \beta \times$ age + $\gamma \times$ age^2 + $\delta \times$ age^3 where α β, γ, and δ vary with the strata.
f(age)	Beta = arbitrary function of age, with different functions for different strata

The category f(age) is equivalent to making age one of the stratification variables, and is also equivalent to making beta a polynomial of degree 16 in age (since the maximum age for children is 17), with coefficients that may vary with the strata.

The fitted models are listed in order of complexity, where the simplest model (1) is an unstratified linear model in age and the most complex model (16) has a prevalence that is an arbitrary function of age, gender, and region. Model 16 is equivalent to calculating independent prevalence estimates for each of the 144 combinations of age, gender, and region.

Table G-1 also includes the -2 Log Likelihood, a goodness-of-fit measure, and the degrees of freedom, DF, which is the total number of estimated parameters. Two models can be compared using their -2 Log Likelihood values; lower values are preferred. If the first model is a special case of the second model, then the approximate statistical significance of the first model is estimated by comparing the difference in the -2 Log Likelihood values with a chi-squared random variable with r degrees of freedom, where r is the difference in the DF. This is a likelihood ratio test. For all pairs of models from Table G-1, all the differences are at least 70,000 and the likelihood ratios are all extremely statistically significant at levels well below 5 percent. Therefore the model 16 is clearly preferred and was used to model the prevalences.

The SURVEYLOGISTIC model predictions are tabulated in Table G-2 below and plotted in Figures 1 and 3 below. Also shown in Table G-2 and in Figures 2 and 4 are results for smoothed curves calculated using a LOESS scatterplot smoother, as discussed below.

The SURVEYLOGISTIC procedure produces estimates of the beta values and their 95 % confidence intervals for each combination of age, region, and gender. Applying the inverse logit transformation,

Prob (asthma) = exp(beta) / (1 + exp(beta)), converted the beta values and 95 % confidence intervals into predictions and 95 % confidence intervals for the prevalence, as shown in Table G-2 and Figures 1 and 3. The standard error for the prevalence was estimated as

$$\text{Std Error } \{\text{Prob (asthma)}\} = \text{Std Error (beta)} \times \exp(-\text{ beta}) / (1 + \exp(\text{beta}))^2,$$

which follows from the delta method (a first order Taylor series approximation).

Loess Smoother

The estimated prevalence curves shows that the prevalence is not a smooth function of age. The linear, quadratic, and cubic functions of age modeled by SURVEYLOGISTIC were one strategy for smoothing the curves, but they did not provide a good fit to the data. One reason for this might be due to the attempt to fit a global regression curve to all the age groups, which means that the predictions for age A are affected by data for very different ages. We instead chose to use a local regression approach that separately fits a regression curve to each age A and its neighboring ages, giving a regression weight of 1 to the age A, and lower weights to the neighboring ages using a tri-weight function:

$$\text{Weight} = \{1 - [\ |\text{age} - A| / q\ ]^{3}\},\ \text{ where } |\ \text{age} - A| <= q.$$

The parameter q defines the number of points in the neighborhood of the age a. Instead of calling q the smoothing parameter, SAS defines the smoothing parameter as the proportion of points in each neighborhood. We fitted a quadratic function of age to each age neighborhood, separately for each gender and region combination. We fitted these local regression curves to the beta values, the logits of the asthma prevalence estimates, and then converted them back to estimated prevalence rates by applying the inverse logit function exp(beta) / (1 + exp(beta)). In addition to the tri-weight variable, each beta value was assigned a weight of 1 / [std error (beta)]^2, to account for their uncertainties.

The SAS LOESS procedure was applied to estimate smoothed curves for beta, the logit of the prevalence, as a function of age, separately for each region and gender. We fitted curves using the choices 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 for the smoothing parameter in an effort to determine the optimum choice based on various regression diagnostics.[12,13]

[12] Two outlier cases were adjusted to avoid wild variations in the "smoothed" curves: For the West region, males, age 0, there were 97 children surveyed that all gave No answers to the asthma question, leading to an estimated value of -15.2029 for beta with a standard error of 0.14. For the Northeast region, females, age 0, there were 29 children surveyed that all gave No answers to the asthma question, leading to an estimated value of -15.2029 for beta with a standard error of 0.19. In both cases the raw probability of asthma equals zero, so the corresponding estimated beta would be negative infinity, but SAS's software gives -15.2029 instead. To reduce the impact of these outlier cases, we replaced their estimated standard errors by 4, which is approximately four times the maximum standard error for all other region, gender, and age combinations.

[13] With only 18 points, a smoothing parameter of 0.2 cannot be used because the weight function assigns zero weights to all ages except age A, and a quadratic model cannot be uniquely fitted to a single value. A smoothing parameter of 0.3 also cannot be used because that choice assigns a neighborhood of 5 points only (0.3 × 18 = 5,

Quantities predicted in these smoothing parameter tests were the predicted value, standard error, confidence interval lower bound and confidence interval upper bound for the betas, and the corresponding values for the prevalence rates.

The polygonal curves joining values for different ages show the predicted values with vertical lines indicating the confidence intervals in Figures 3 and 4 for smoothing parameters 0 (i.e., no smoothing) and 0.5, respectively. Note that the confidence intervals are not symmetric about the predicted values because of the inverse logit transformation.

Note that in our application of LOESS, we used weights of 1 / [std error (beta)] ^{2}, so that $\sigma^2 = 1$ for this application. The LOESS procedure estimates σ^2 from the weighted sum of squares. Since in our application we assume $\sigma^2 = 1$, we multiplied the estimated standard errors by 1 / estimated σ, and adjusted the widths of the confidence intervals by the same factor.

Additionally, because the true value of σ equals 1, the best choices of smoothing parameter should give residual standard errors close to one. Using this criterion the best choice varies with gender and region between smoothing parameters 0.4 (3 cases), 0.5 (2 cases), 0.6 (1 case), and 0.7 (1 case).

As a further regression diagnostic the residual errors from the LOESS model were divided by std error (beta) to make their variances approximately constant. These approximately studentized residuals, 'student,' should be approximately normally distributed with a mean of zero and a variance of $\sigma^2 = 1$. To test this assumption, normal probability plots of the residuals were created for each smoothing parameter, combining all the studentized residuals across genders, regions, and ages. The plots for smoothing parameters seem to be equally straight for each smoothing parameter.

The final regression diagnostic is a plot of the studentized residuals against the smoothed beta values. Ideally there should be no obvious pattern and an average studentized residual close to zero. The plots indeed showed no unusual patterns, and the results for smoothing parameters 0.5 and 0.6 seem to showed a fitted LOESS close to the studentized residual equals zero line.

The regression diagnostics suggested the choice of smoothing parameter as 0.4 or 0.5. Normal probability plots did not suggest any preferred choices. The plots of residuals against smoothed predictions suggest the choices of 0.5 or 0.6. We therefore chose the final value of 0.5. These predictions, standard errors, and confidence intervals are presented in tabular form below as Table G-2.

rounded down), of which the two outside ages have assigned weight zero, making the local quadratic model fit exactly at every point except for the end points (ages 0, 1, 16 and 17). Usually one uses a smoothing parameter below one so that not all the data are used for the local regression at a given x value.

Figure 1. Raw asthma prevalence rates by age and gender for each region
region=Midwest

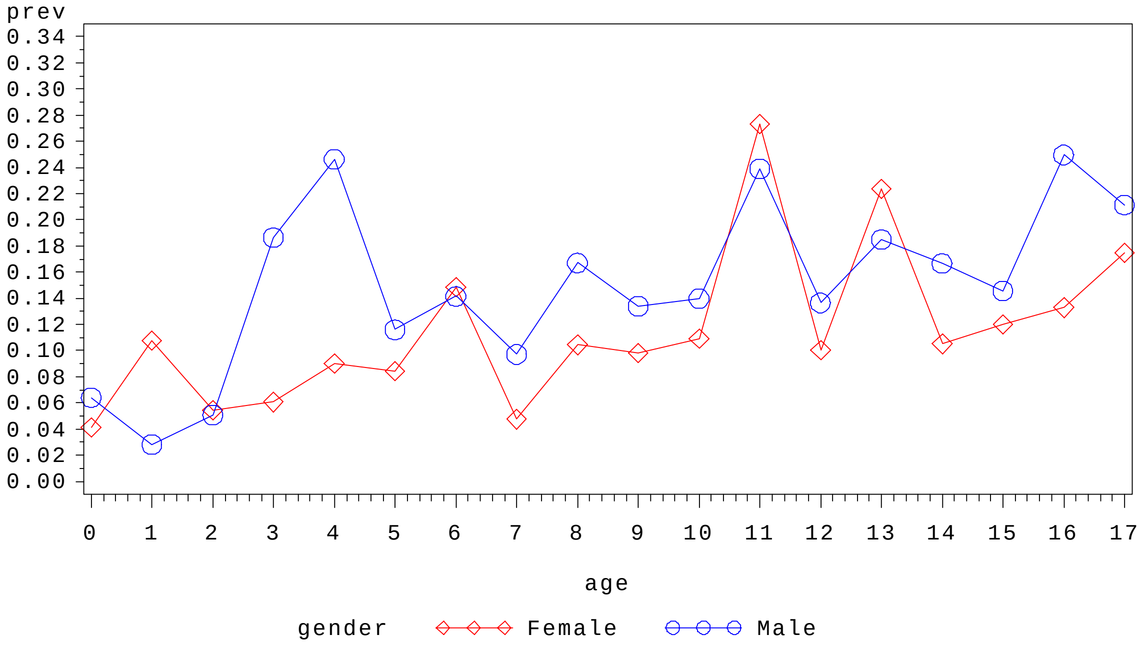


Figure 1. Raw asthma prevalence rates by age and gender for each region
region=Northeast

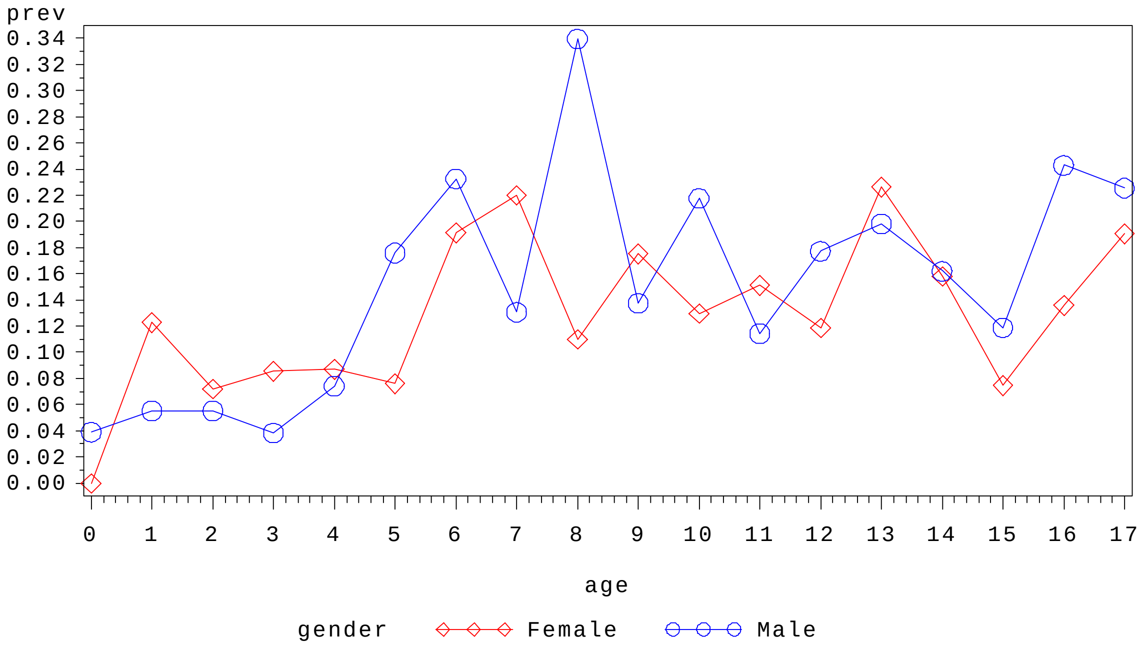


Figure 1. Raw asthma prevalence rates by age and gender for each region
region=South

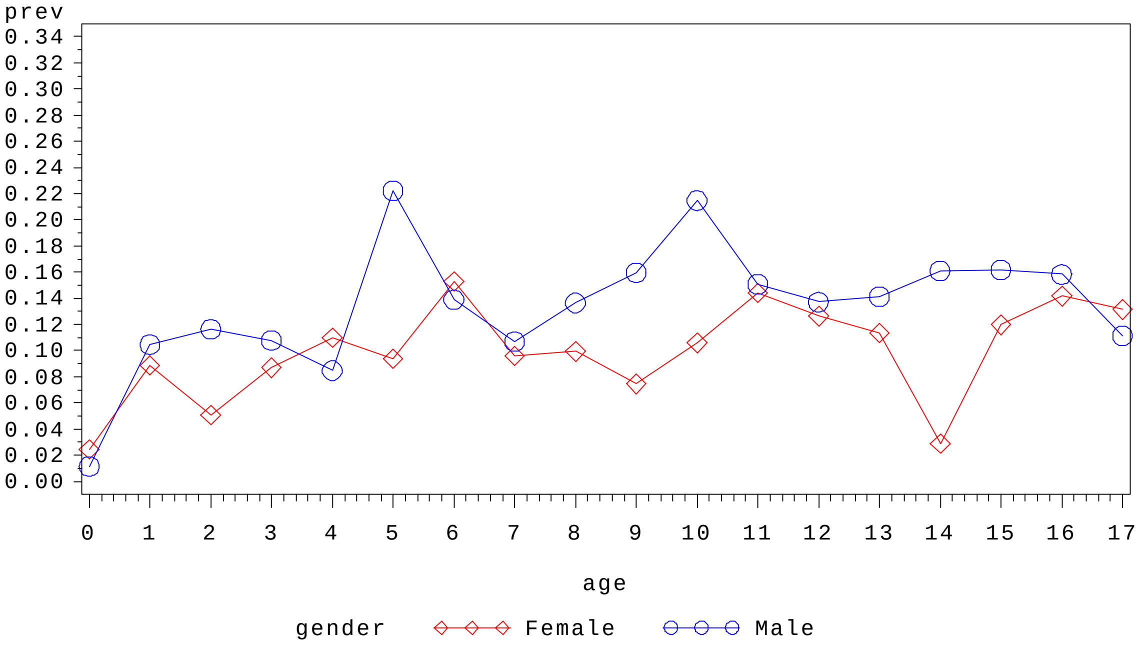


Figure 1. Raw asthma prevalence rates by age and gender for each region
region=West

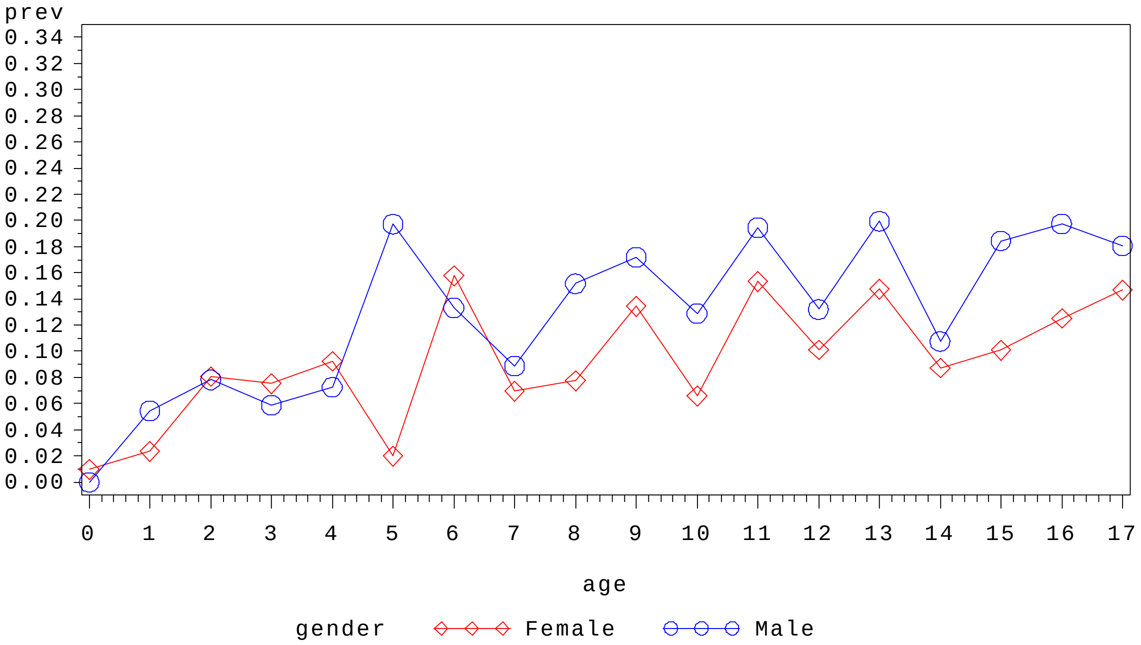


Figure 2. Smoothed asthma prevalence rates by age for each region and gender

region=Midwest

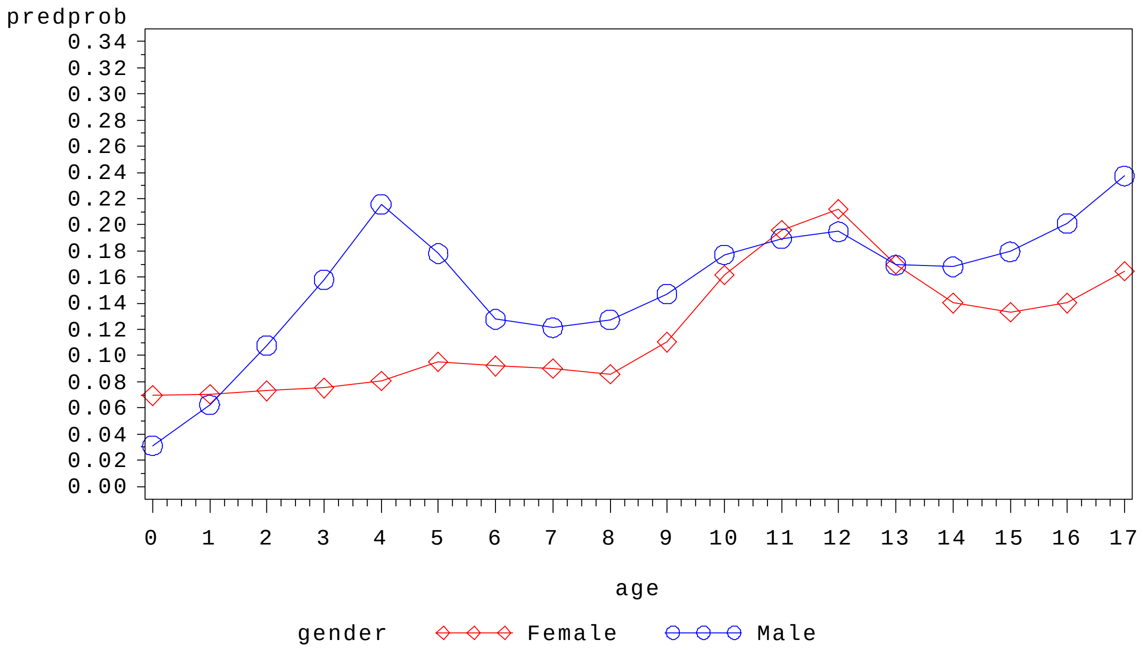


Figure 2. Smoothed asthma prevalence rates by age for each region and gender

region=Northeast

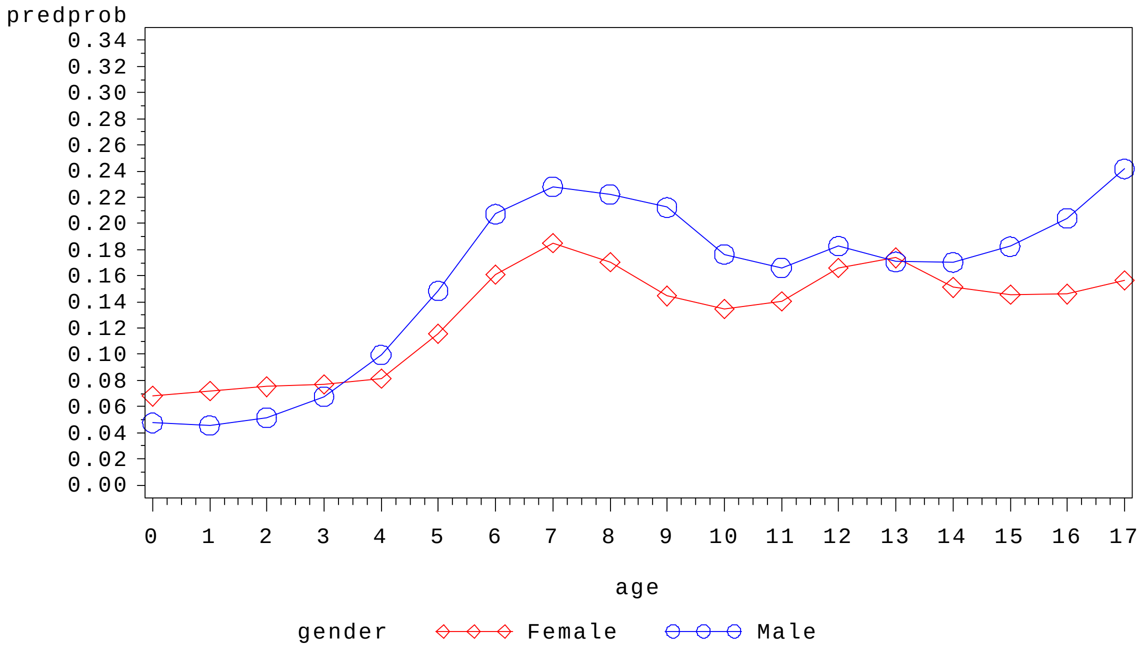


Figure 2. Smoothed asthma prevalence rates by age for each region and gender

region=South

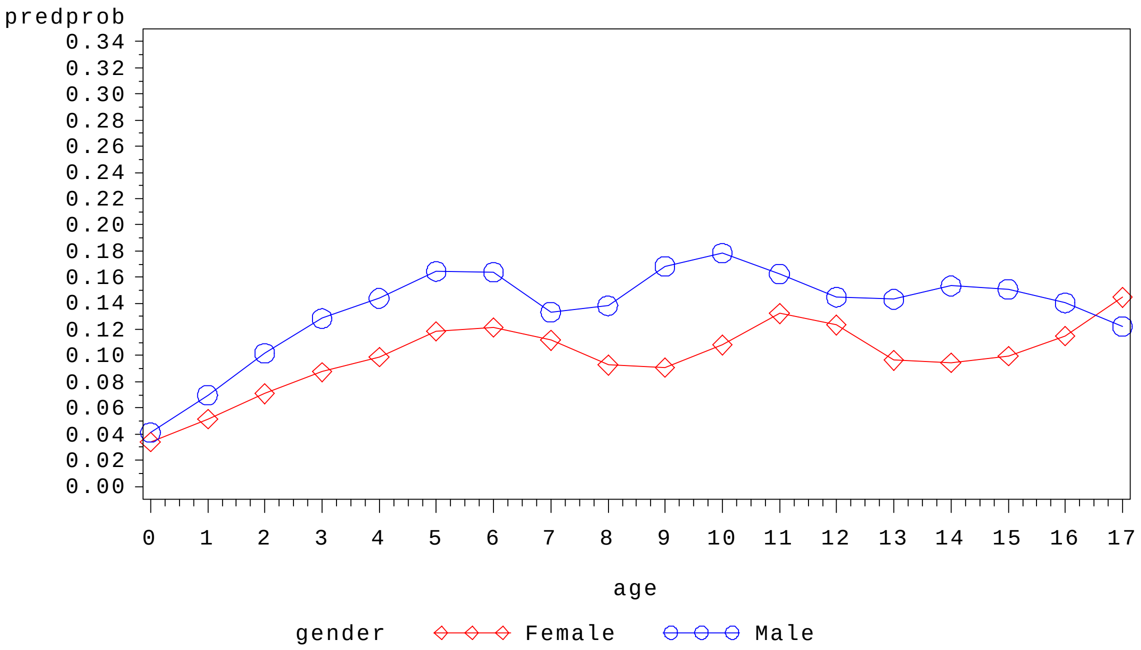


Figure 2. Smoothed asthma prevalence rates by age for each region and gender

region=West

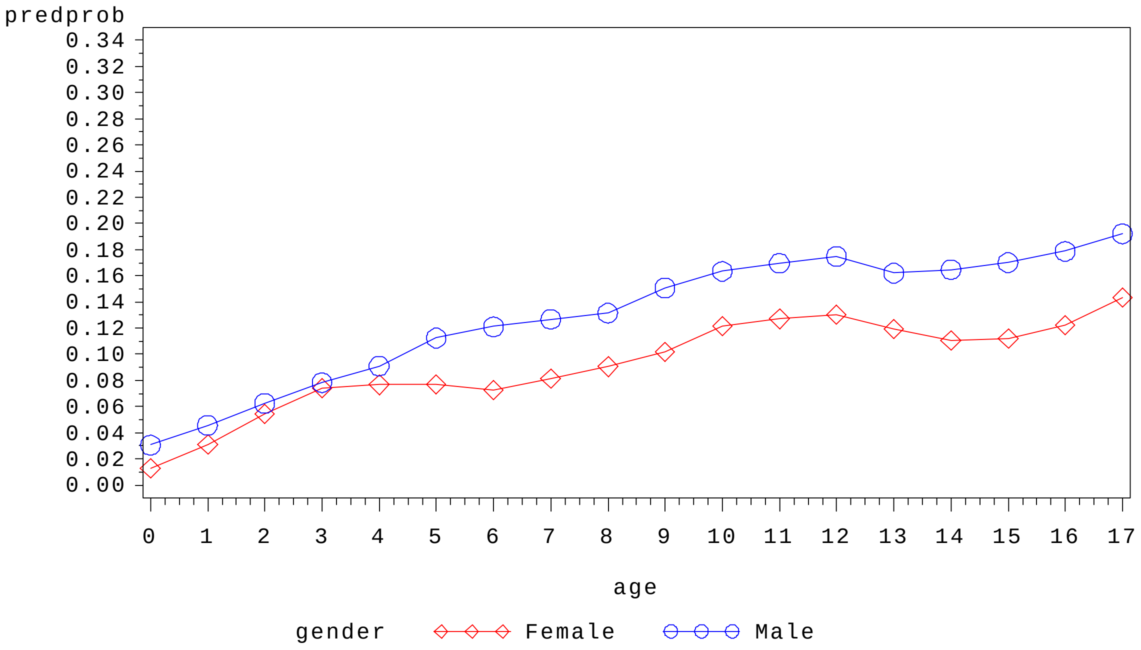


Figure 3. Raw asthma prevalence rates and confidence intervals
region=Midwest

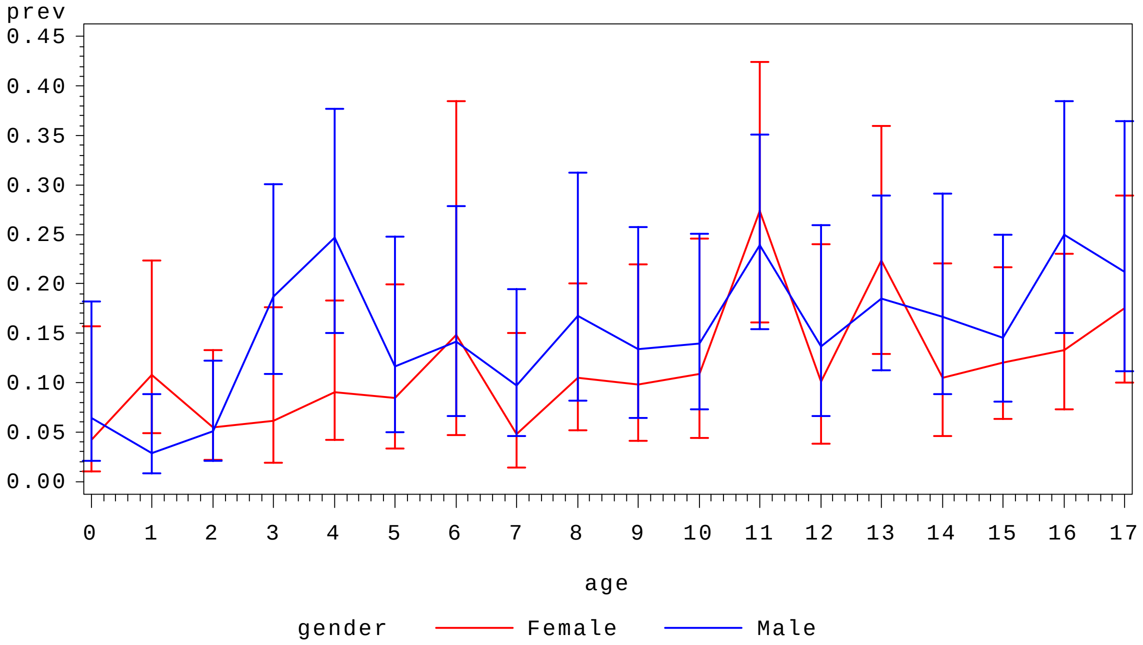


Figure 3. Raw asthma prevalence rates and confidence intervals
region=Northeast

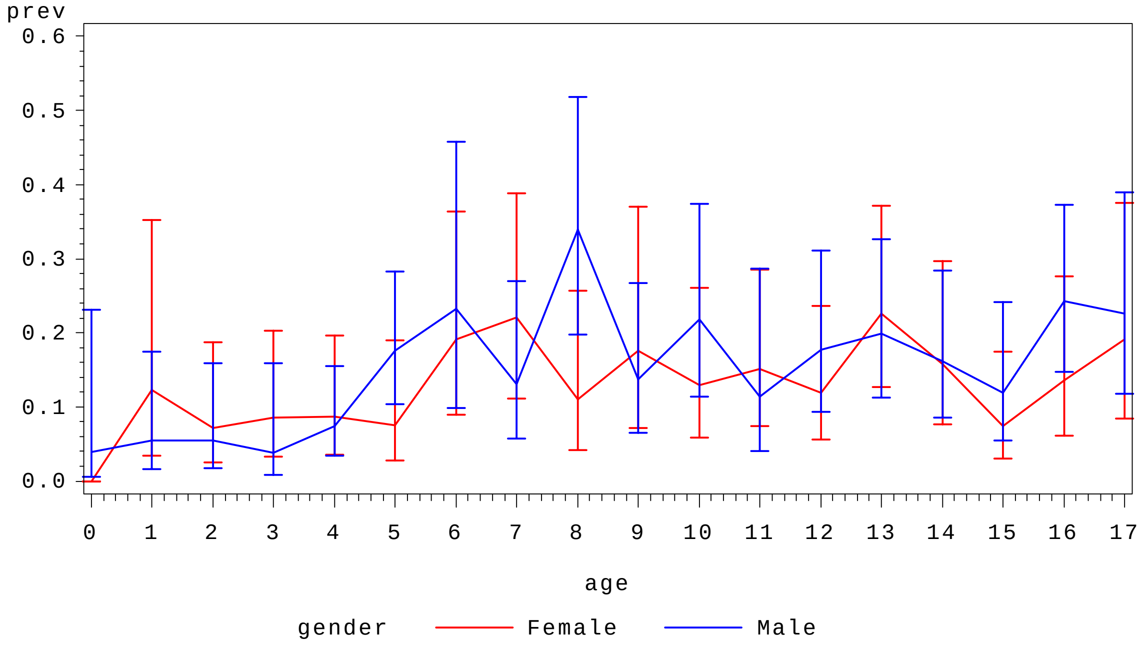


Figure 3. Raw asthma prevalence rates and confidence intervals

region=South

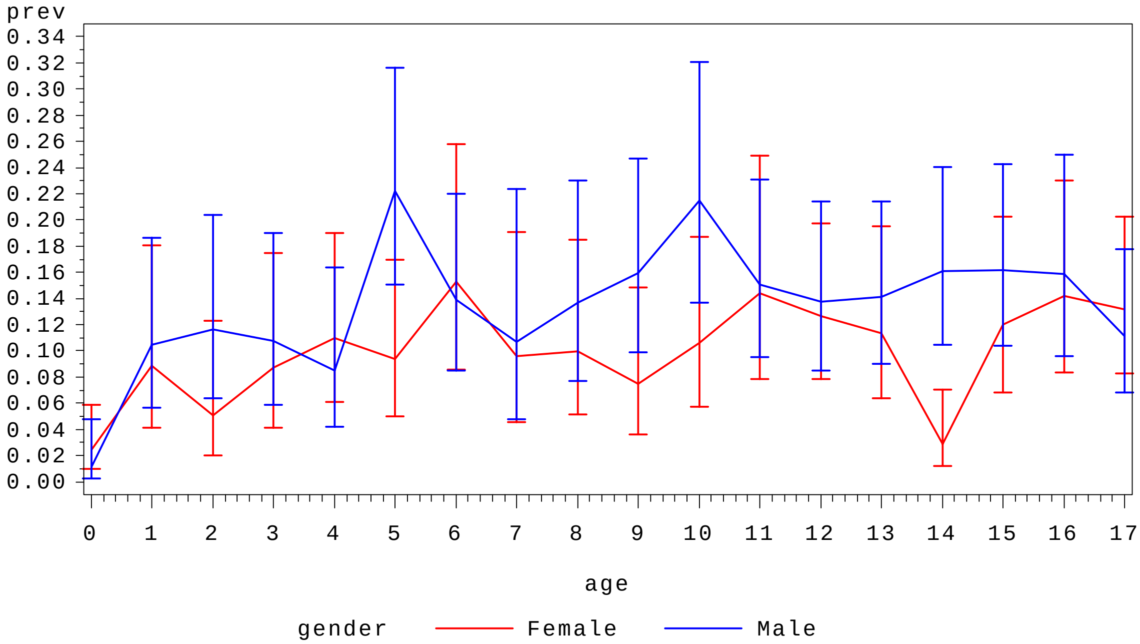


Figure 3. Raw asthma prevalence rates and confidence intervals

region=West

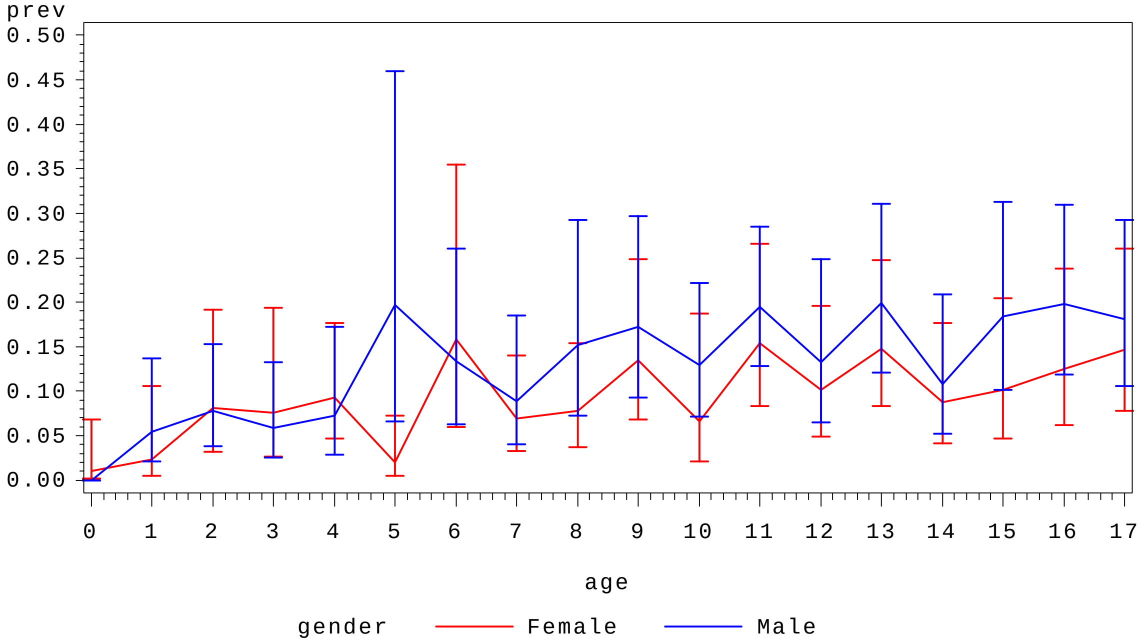


Figure 4. Smoothed asthma prevalence rates and confidence intervals

region=Midwest

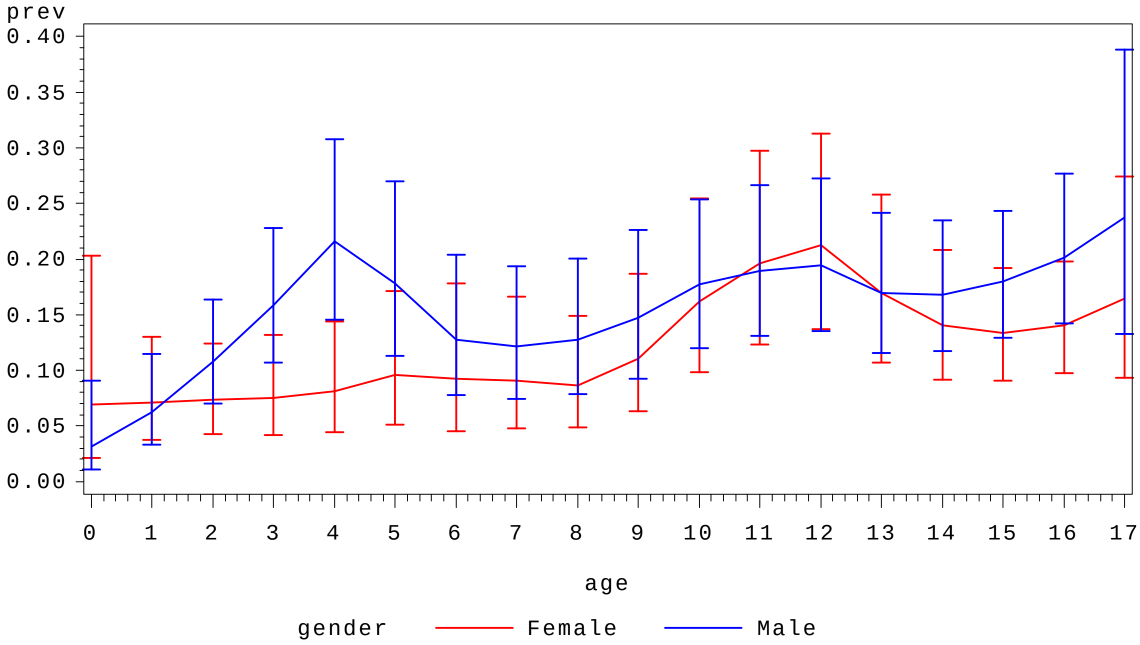


Figure 4. Smoothed asthma prevalence rates and confidence intervals

region=Northeast

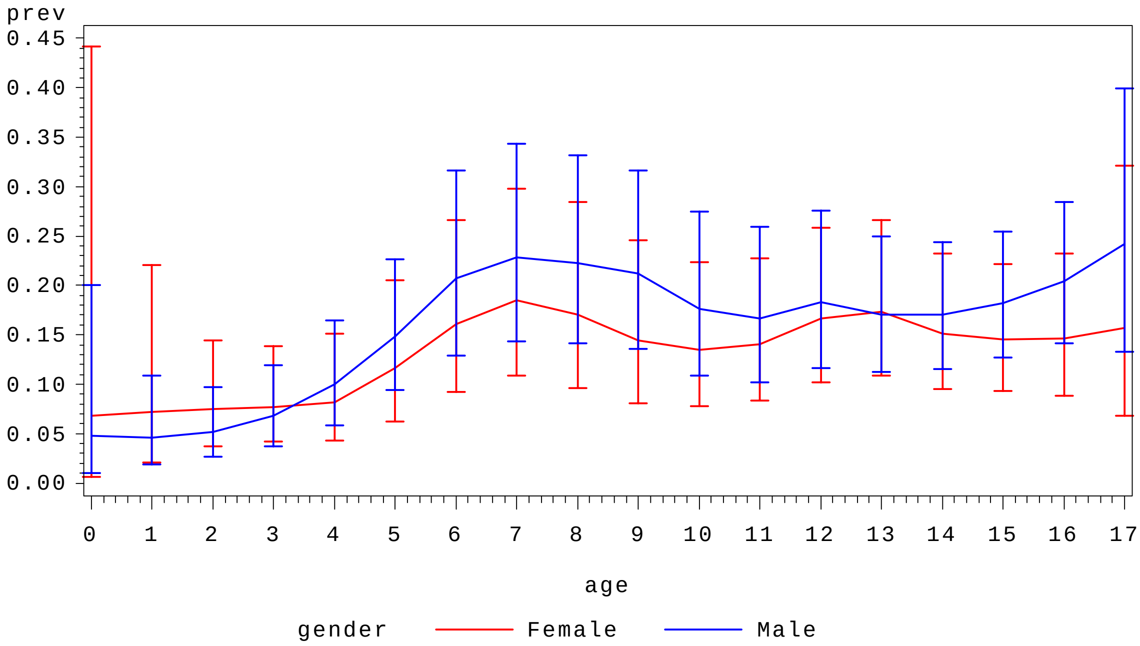


Figure 4. Smoothed asthma prevalence rates and confidence intervals

region=South

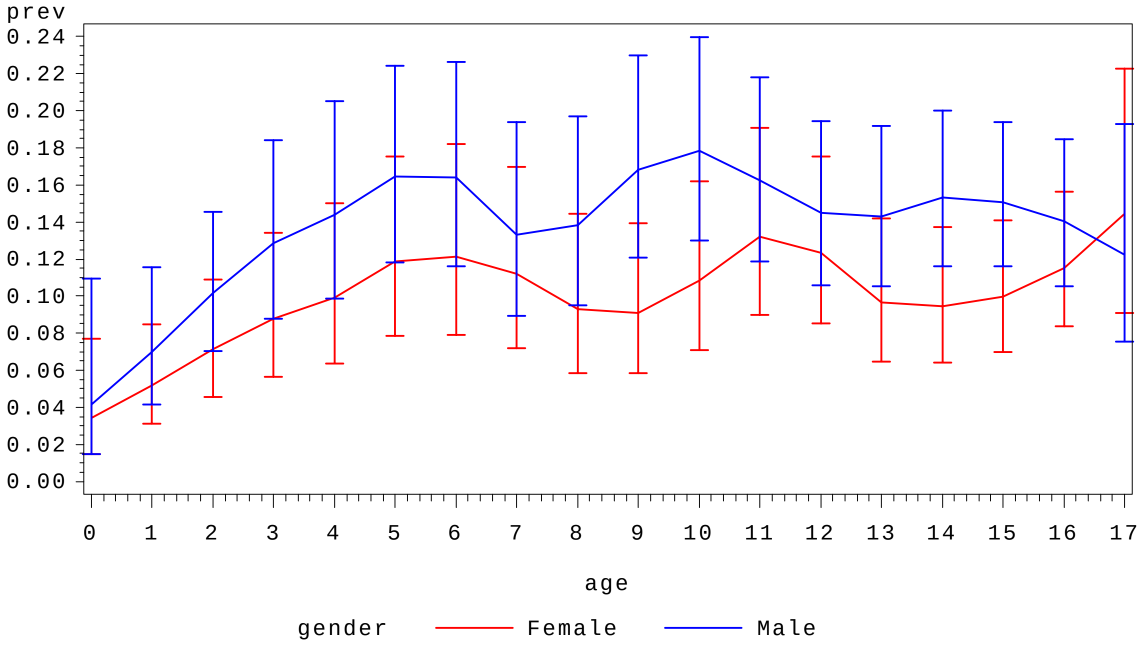


Figure 4. Smoothed asthma prevalence rates and confidence intervals

region=West

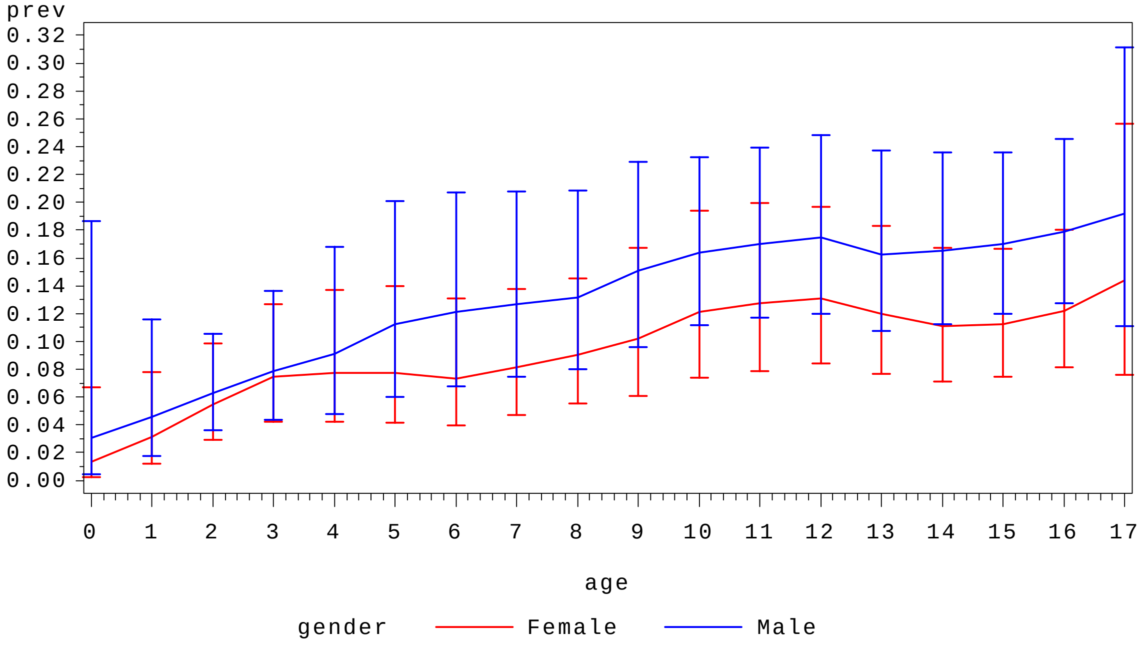


Table G-2. Raw and smoothed prevalence rates, with confidence intervals, by region, gender, and age.

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
1	Midwest	Female	0	No	0.04161	0.02965	0.01001	0.15717
2	Midwest	Female	0	Yes	0.06956	0.03574	0.02143	0.20330
3	Midwest	Female	1	No	0.10790	0.04254	0.04840	0.22336
4	Midwest	Female	1	Yes	0.07078	0.01995	0.03736	0.13008
5	Midwest	Female	2	No	0.05469	0.02578	0.02131	0.13325
6	Midwest	Female	2	Yes	0.07324	0.01778	0.04228	0.12395
7	Midwest	Female	3	No	0.06094	0.03474	0.01936	0.17579
8	Midwest	Female	3	Yes	0.07542	0.01944	0.04205	0.13163
9	Midwest	Female	4	No	0.09049	0.03407	0.04233	0.18298
10	Midwest	Female	4	Yes	0.08100	0.02163	0.04417	0.14393
11	Midwest	Female	5	No	0.08463	0.03917	0.03317	0.19942
12	Midwest	Female	5	Yes	0.09540	0.02613	0.05106	0.17131
13	Midwest	Female	6	No	0.14869	0.08250	0.04643	0.38520
14	Midwest	Female	6	Yes	0.09210	0.02854	0.04534	0.17808
15	Midwest	Female	7	No	0.04757	0.02927	0.01389	0.15051
16	Midwest	Female	7	Yes	0.09032	0.02563	0.04728	0.16571
17	Midwest	Female	8	No	0.10444	0.03638	0.05160	0.19997
18	Midwest	Female	8	Yes	0.08612	0.02181	0.04842	0.14857
19	Midwest	Female	9	No	0.09836	0.04283	0.04062	0.21943
20	Midwest	Female	9	Yes	0.11040	0.02709	0.06298	0.18643
21	Midwest	Female	10	No	0.10916	0.04859	0.04400	0.24600
22	Midwest	Female	10	Yes	0.16190	0.03486	0.09838	0.25484
23	Midwest	Female	11	No	0.27341	0.06817	0.16112	0.42437
24	Midwest	Female	11	Yes	0.19597	0.03920	0.12296	0.29763
25	Midwest	Female	12	No	0.10055	0.04780	0.03816	0.23952
26	Midwest	Female	12	Yes	0.21214	0.03957	0.13724	0.31309

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
27	Midwest	Female	13	No	0.22388	0.05905	0.12907	0.35959
28	Midwest	Female	13	Yes	0.16966	0.03371	0.10716	0.25807
29	Midwest	Female	14	No	0.10511	0.04233	0.04637	0.22104
30	Midwest	Female	14	Yes	0.14020	0.02603	0.09164	0.20857
31	Midwest	Female	15	No	0.12026	0.03805	0.06327	0.21670
32	Midwest	Female	15	Yes	0.13341	0.02266	0.09056	0.19226
33	Midwest	Female	16	No	0.13299	0.03933	0.07288	0.23037
34	Midwest	Female	16	Yes	0.14040	0.02235	0.09764	0.19777
35	Midwest	Female	17	No	0.17497	0.04786	0.09970	0.28884
36	Midwest	Female	17	Yes	0.16478	0.04037	0.09320	0.27468
37	Midwest	Male	0	No	0.06419	0.03612	0.02068	0.18227
38	Midwest	Male	0	Yes	0.03134	0.01537	0.01042	0.09046
39	Midwest	Male	1	No	0.02824	0.01694	0.00859	0.08879
40	Midwest	Male	1	Yes	0.06250	0.01751	0.03321	0.11457
41	Midwest	Male	2	No	0.05102	0.02343	0.02040	0.12189
42	Midwest	Male	2	Yes	0.10780	0.02078	0.06960	0.16328
43	Midwest	Male	3	No	0.18650	0.04864	0.10898	0.30057
44	Midwest	Male	3	Yes	0.15821	0.02705	0.10696	0.22775
45	Midwest	Male	4	No	0.24649	0.05823	0.15035	0.37686
46	Midwest	Male	4	Yes	0.21572	0.03661	0.14543	0.30774
47	Midwest	Male	5	No	0.11609	0.04818	0.04973	0.24793
48	Midwest	Male	5	Yes	0.17822	0.03525	0.11280	0.27003
49	Midwest	Male	6	No	0.14158	0.05280	0.06576	0.27873
50	Midwest	Male	6	Yes	0.12788	0.02799	0.07751	0.20375
51	Midwest	Male	7	No	0.09726	0.03614	0.04588	0.19448
52	Midwest	Male	7	Yes	0.12145	0.02642	0.07391	0.19317
53	Midwest	Male	8	No	0.16718	0.05814	0.08134	0.31276
54	Midwest	Male	8	Yes	0.12757	0.02700	0.07864	0.20031
55	Midwest	Male	9	No	0.13406	0.04783	0.06458	0.25769

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
56	Midwest	Male	9	Yes	0.14718	0.02976	0.09254	0.22603
57	Midwest	Male	10	No	0.13986	0.04422	0.07331	0.25050
58	Midwest	Male	10	Yes	0.17728	0.02996	0.12020	0.25366
59	Midwest	Male	11	No	0.23907	0.05031	0.15449	0.35075
60	Midwest	Male	11	Yes	0.18961	0.03044	0.13100	0.26639
61	Midwest	Male	12	No	0.13660	0.04784	0.06668	0.25946
62	Midwest	Male	12	Yes	0.19487	0.03078	0.13541	0.27221
63	Midwest	Male	13	No	0.18501	0.04498	0.11230	0.28945
64	Midwest	Male	13	Yes	0.16939	0.02841	0.11528	0.24195
65	Midwest	Male	14	No	0.16673	0.05094	0.08886	0.29104
66	Midwest	Male	14	Yes	0.16795	0.02631	0.11734	0.23459
67	Midwest	Male	15	No	0.14583	0.04241	0.08054	0.24967
68	Midwest	Male	15	Yes	0.17953	0.02561	0.12951	0.24347
69	Midwest	Male	16	No	0.24965	0.06037	0.15033	0.38489
70	Midwest	Male	16	Yes	0.20116	0.03048	0.14187	0.27721
71	Midwest	Male	17	No	0.21152	0.06481	0.11131	0.36490
72	Midwest	Male	17	Yes	0.23741	0.05816	0.13243	0.38835
73	Northeast	Female	0	No	0.00000	0.00000	0.00000	0.00000
74	Northeast	Female	0	Yes	0.06807	0.06565	0.00670	0.44174
75	Northeast	Female	1	No	0.12262	0.07443	0.03476	0.35164
76	Northeast	Female	1	Yes	0.07219	0.03765	0.02088	0.22109
77	Northeast	Female	2	No	0.07217	0.03707	0.02561	0.18713
78	Northeast	Female	2	Yes	0.07522	0.02212	0.03764	0.14468
79	Northeast	Female	3	No	0.08550	0.03991	0.03324	0.20269
80	Northeast	Female	3	Yes	0.07709	0.02021	0.04162	0.13840
81	Northeast	Female	4	No	0.08704	0.03804	0.03596	0.19592
82	Northeast	Female	4	Yes	0.08171	0.02252	0.04269	0.15080
83	Northeast	Female	5	No	0.07597	0.03754	0.02801	0.18998
84	Northeast	Female	5	Yes	0.11603	0.03012	0.06258	0.20515

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
85	Northeast	Female	6	No	0.19149	0.06960	0.08937	0.36372
86	Northeast	Female	6	Yes	0.16106	0.03737	0.09219	0.26629
87	Northeast	Female	7	No	0.22034	0.07076	0.11195	0.38783
88	Northeast	Female	7	Yes	0.18503	0.04087	0.10844	0.29764
89	Northeast	Female	8	No	0.11002	0.05128	0.04241	0.25654
90	Northeast	Female	8	Yes	0.17054	0.04039	0.09628	0.28407
91	Northeast	Female	9	No	0.17541	0.07488	0.07159	0.36981
92	Northeast	Female	9	Yes	0.14457	0.03538	0.08042	0.24618
93	Northeast	Female	10	No	0.12980	0.04964	0.05930	0.26087
94	Northeast	Female	10	Yes	0.13487	0.03098	0.07799	0.22319
95	Northeast	Female	11	No	0.15128	0.05287	0.07366	0.28547
96	Northeast	Female	11	Yes	0.14072	0.03068	0.08367	0.22704
97	Northeast	Female	12	No	0.11890	0.04426	0.05568	0.23597
98	Northeast	Female	12	Yes	0.16615	0.03375	0.10211	0.25877
99	Northeast	Female	13	No	0.22638	0.06285	0.12650	0.37158
100	Northeast	Female	13	Yes	0.17374	0.03402	0.10861	0.26626
101	Northeast	Female	14	No	0.15807	0.05513	0.07694	0.29719
102	Northeast	Female	14	Yes	0.15137	0.02946	0.09519	0.23220
103	Northeast	Female	15	No	0.07460	0.03409	0.02971	0.17506
104	Northeast	Female	15	Yes	0.14564	0.02761	0.09279	0.22127
105	Northeast	Female	16	No	0.13603	0.05328	0.06081	0.27686
106	Northeast	Female	16	Yes	0.14601	0.03095	0.08805	0.23241
107	Northeast	Female	17	No	0.19074	0.07382	0.08451	0.37568
108	Northeast	Female	17	Yes	0.15662	0.05374	0.06784	0.32151
109	Northeast	Male	0	No	0.03904	0.03829	0.00547	0.23095
110	Northeast	Male	0	Yes	0.04768	0.03299	0.00991	0.20023
111	Northeast	Male	1	No	0.05533	0.03425	0.01596	0.17461
112	Northeast	Male	1	Yes	0.04564	0.01831	0.01850	0.10821
113	Northeast	Male	2	No	0.05525	0.03119	0.01781	0.15872

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
114	Northeast	Male	2	Yes	0.05161	0.01505	0.02680	0.09709
115	Northeast	Male	3	No	0.03842	0.02923	0.00840	0.15853
116	Northeast	Male	3	Yes	0.06766	0.01784	0.03734	0.11955
117	Northeast	Male	4	No	0.07436	0.02906	0.03393	0.15522
118	Northeast	Male	4	Yes	0.09964	0.02330	0.05859	0.16441
119	Northeast	Male	5	No	0.17601	0.04519	0.10393	0.28234
120	Northeast	Male	5	Yes	0.14854	0.02948	0.09428	0.22623
121	Northeast	Male	6	No	0.23271	0.09319	0.09832	0.45756
122	Northeast	Male	6	Yes	0.20731	0.04235	0.12875	0.31640
123	Northeast	Male	7	No	0.13074	0.05195	0.05785	0.26922
124	Northeast	Male	7	Yes	0.22820	0.04524	0.14338	0.34311
125	Northeast	Male	8	No	0.33970	0.08456	0.19726	0.51855
126	Northeast	Male	8	Yes	0.22240	0.04298	0.14157	0.33157
127	Northeast	Male	9	No	0.13761	0.05024	0.06507	0.26785
128	Northeast	Male	9	Yes	0.21238	0.04071	0.13589	0.31617
129	Northeast	Male	10	No	0.21785	0.06659	0.11464	0.37465
130	Northeast	Male	10	Yes	0.17652	0.03731	0.10824	0.27460
131	Northeast	Male	11	No	0.11448	0.05849	0.04005	0.28601
132	Northeast	Male	11	Yes	0.16617	0.03516	0.10200	0.25907
133	Northeast	Male	12	No	0.17736	0.05489	0.09349	0.31067
134	Northeast	Male	12	Yes	0.18279	0.03589	0.11611	0.27581
135	Northeast	Male	13	No	0.19837	0.05450	0.11222	0.32635
136	Northeast	Male	13	Yes	0.17078	0.03078	0.11288	0.25000
137	Northeast	Male	14	No	0.16201	0.04973	0.08618	0.28386
138	Northeast	Male	14	Yes	0.17033	0.02889	0.11547	0.24408
139	Northeast	Male	15	No	0.11894	0.04584	0.05417	0.24139
140	Northeast	Male	15	Yes	0.18246	0.02858	0.12740	0.25438
141	Northeast	Male	16	No	0.24306	0.05798	0.14759	0.37326
142	Northeast	Male	16	Yes	0.20406	0.03216	0.14187	0.28447

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
143	Northeast	Male	17	No	0.22559	0.06980	0.11748	0.38930
144	Northeast	Male	17	Yes	0.24185	0.06066	0.13291	0.39898
145	South	Female	0	No	0.02459	0.01116	0.01002	0.05906
146	South	Female	0	Yes	0.03407	0.01282	0.01465	0.07723
147	South	Female	1	No	0.08869	0.03373	0.04118	0.18067
148	South	Female	1	Yes	0.05182	0.01167	0.03127	0.08472
149	South	Female	2	No	0.05097	0.02373	0.02012	0.12319
150	South	Female	2	Yes	0.07110	0.01386	0.04584	0.10869
151	South	Female	3	No	0.08717	0.03240	0.04122	0.17500
152	South	Female	3	Yes	0.08759	0.01718	0.05624	0.13394
153	South	Female	4	No	0.11010	0.03209	0.06113	0.19035
154	South	Female	4	Yes	0.09897	0.01914	0.06387	0.15025
155	South	Female	5	No	0.09409	0.02943	0.05015	0.16968
156	South	Female	5	Yes	0.11870	0.02157	0.07855	0.17548
157	South	Female	6	No	0.15318	0.04317	0.08611	0.25777
158	South	Female	6	Yes	0.12150	0.02282	0.07925	0.18182
159	South	Female	7	No	0.09608	0.03538	0.04565	0.19105
160	South	Female	7	Yes	0.11192	0.02171	0.07204	0.16985
161	South	Female	8	No	0.09955	0.03288	0.05111	0.18493
162	South	Female	8	Yes	0.09287	0.01897	0.05850	0.14436
163	South	Female	9	No	0.07477	0.02719	0.03606	0.14864
164	South	Female	9	Yes	0.09117	0.01786	0.05855	0.13929
165	South	Female	10	No	0.10602	0.03214	0.05750	0.18732
166	South	Female	10	Yes	0.10821	0.02026	0.07077	0.16201
167	South	Female	11	No	0.14411	0.04267	0.07875	0.24907
168	South	Female	11	Yes	0.13237	0.02251	0.08989	0.19071
169	South	Female	12	No	0.12646	0.02981	0.07860	0.19723
170	South	Female	12	Yes	0.12346	0.02004	0.08543	0.17519
171	South	Female	13	No	0.11376	0.03270	0.06365	0.19510

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
172	South	Female	13	Yes	0.09653	0.01717	0.06458	0.14190
173	South	Female	14	No	0.02915	0.01339	0.01174	0.07054
174	South	Female	14	Yes	0.09469	0.01619	0.06436	0.13721
175	South	Female	15	No	0.11985	0.03357	0.06801	0.20259
176	South	Female	15	Yes	0.09988	0.01586	0.06978	0.14099
177	South	Female	16	No	0.14183	0.03685	0.08366	0.23028
178	South	Female	16	Yes	0.11501	0.01620	0.08365	0.15612
179	South	Female	17	No	0.13141	0.03007	0.08280	0.20226
180	South	Female	17	Yes	0.14466	0.02946	0.09067	0.22291
181	South	Male	0	No	0.01164	0.00852	0.00275	0.04790
182	South	Male	0	Yes	0.04132	0.01867	0.01487	0.10956
183	South	Male	1	No	0.10465	0.03216	0.05629	0.18635
184	South	Male	1	Yes	0.06981	0.01623	0.04125	0.11576
185	South	Male	2	No	0.11644	0.03486	0.06353	0.20382
186	South	Male	2	Yes	0.10189	0.01672	0.07024	0.14557
187	South	Male	3	No	0.10794	0.03253	0.05874	0.19005
188	South	Male	3	Yes	0.12852	0.02139	0.08793	0.18405
189	South	Male	4	No	0.08480	0.02973	0.04190	0.16410
190	South	Male	4	Yes	0.14393	0.02379	0.09861	0.20534
191	South	Male	5	No	0.22243	0.04227	0.15052	0.31592
192	South	Male	5	Yes	0.16450	0.02373	0.11821	0.22430
193	South	Male	6	No	0.13908	0.03392	0.08485	0.21964
194	South	Male	6	Yes	0.16386	0.02460	0.11613	0.22617
195	South	Male	7	No	0.10695	0.04272	0.04747	0.22347
196	South	Male	7	Yes	0.13329	0.02322	0.08951	0.19392
197	South	Male	8	No	0.13660	0.03841	0.07712	0.23049
198	South	Male	8	Yes	0.13818	0.02276	0.09484	0.19702
199	South	Male	9	No	0.15978	0.03742	0.09920	0.24720
200	South	Male	9	Yes	0.16839	0.02450	0.12062	0.23012

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
201	South	Male	10	No	0.21482	0.04702	0.13676	0.32086
202	South	Male	10	Yes	0.17848	0.02453	0.13021	0.23972
203	South	Male	11	No	0.15078	0.03440	0.09492	0.23112
204	South	Male	11	Yes	0.16247	0.02224	0.11881	0.21820
205	South	Male	12	No	0.13727	0.03260	0.08489	0.21438
206	South	Male	12	Yes	0.14480	0.01976	0.10610	0.19453
207	South	Male	13	No	0.14136	0.03119	0.09049	0.21409
208	South	Male	13	Yes	0.14318	0.01928	0.10537	0.19165
209	South	Male	14	No	0.16110	0.03444	0.10438	0.24037
210	South	Male	14	Yes	0.15339	0.01875	0.11612	0.19992
211	South	Male	15	No	0.16172	0.03519	0.10394	0.24291
212	South	Male	15	Yes	0.15088	0.01746	0.11598	0.19398
213	South	Male	16	No	0.15836	0.03879	0.09614	0.24974
214	South	Male	16	Yes	0.14038	0.01773	0.10533	0.18467
215	South	Male	17	No	0.11156	0.02737	0.06810	0.17746
216	South	Male	17	Yes	0.12247	0.02596	0.07537	0.19286
217	West	Female	0	No	0.00983	0.00990	0.00135	0.06802
218	West	Female	0	Yes	0.01318	0.00987	0.00248	0.06700
219	West	Female	1	No	0.02367	0.01862	0.00497	0.10522
220	West	Female	1	Yes	0.03105	0.01312	0.01204	0.07769
221	West	Female	2	No	0.08097	0.03759	0.03170	0.19166
222	West	Female	2	Yes	0.05440	0.01482	0.02948	0.09825
223	West	Female	3	No	0.07528	0.03851	0.02679	0.19404
224	West	Female	3	Yes	0.07444	0.01842	0.04257	0.12701
225	West	Female	4	No	0.09263	0.03196	0.04621	0.17703
226	West	Female	4	Yes	0.07696	0.02064	0.04194	0.13701
227	West	Female	5	No	0.01976	0.01347	0.00513	0.07302
228	West	Female	5	Yes	0.07737	0.02123	0.04157	0.13949
229	West	Female	6	No	0.15792	0.07301	0.06009	0.35487

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
230	West	Female	6	Yes	0.07298	0.01985	0.03947	0.13107
231	West	Female	7	No	0.06955	0.02567	0.03321	0.13989
232	West	Female	7	Yes	0.08146	0.01987	0.04691	0.13776
233	West	Female	8	No	0.07753	0.02825	0.03731	0.15417
234	West	Female	8	Yes	0.09062	0.01994	0.05507	0.14558
235	West	Female	9	No	0.13440	0.04481	0.06802	0.24832
236	West	Female	9	Yes	0.10215	0.02347	0.06061	0.16709
237	West	Female	10	No	0.06573	0.03719	0.02102	0.18736
238	West	Female	10	Yes	0.12152	0.02660	0.07376	0.19374
239	West	Female	11	No	0.15354	0.04584	0.08329	0.26584
240	West	Female	11	Yes	0.12719	0.02688	0.07852	0.19950
241	West	Female	12	No	0.10120	0.03594	0.04934	0.19631
242	West	Female	12	Yes	0.13054	0.02498	0.08440	0.19650
243	West	Female	13	No	0.14759	0.04125	0.08346	0.24769
244	West	Female	13	Yes	0.11968	0.02369	0.07629	0.18284
245	West	Female	14	No	0.08748	0.03284	0.04105	0.17675
246	West	Female	14	Yes	0.11063	0.02132	0.07145	0.16744
247	West	Female	15	No	0.10099	0.03841	0.04674	0.20471
248	West	Female	15	Yes	0.11236	0.02051	0.07428	0.16645
249	West	Female	16	No	0.12538	0.04343	0.06188	0.23755
250	West	Female	16	Yes	0.12224	0.02210	0.08108	0.18021
251	West	Female	17	No	0.14672	0.04582	0.07743	0.26052
252	West	Female	17	Yes	0.14371	0.03992	0.07558	0.25621
253	West	Male	0	No	0.00000	0.00000	0.00000	0.00000
254	West	Male	0	Yes	0.03075	0.02534	0.00437	0.18642
255	West	Male	1	No	0.05457	0.02662	0.02056	0.13695
256	West	Male	1	Yes	0.04584	0.01889	0.01729	0.11595
257	West	Male	2	No	0.07833	0.02789	0.03833	0.15342
258	West	Male	2	Yes	0.06254	0.01442	0.03627	0.10573

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
259	West	Male	3	No	0.05897	0.02530	0.02500	0.13281
260	West	Male	3	Yes	0.07844	0.01913	0.04398	0.13607
261	West	Male	4	No	0.07267	0.03354	0.02870	0.17208
262	West	Male	4	Yes	0.09122	0.02482	0.04765	0.16763
263	West	Male	5	No	0.19732	0.10033	0.06632	0.45969
264	West	Male	5	Yes	0.11262	0.02937	0.06021	0.20092
265	West	Male	6	No	0.13335	0.04859	0.06322	0.25970
266	West	Male	6	Yes	0.12119	0.02916	0.06799	0.20680
267	West	Male	7	No	0.08881	0.03493	0.04015	0.18508
268	West	Male	7	Yes	0.12691	0.02806	0.07464	0.20758
269	West	Male	8	No	0.15183	0.05484	0.07210	0.29200
270	West	Male	8	Yes	0.13161	0.02705	0.08037	0.20811
271	West	Male	9	No	0.17199	0.05164	0.09260	0.29715
272	West	Male	9	Yes	0.15079	0.02837	0.09590	0.22915
273	West	Male	10	No	0.12897	0.03747	0.07151	0.22159
274	West	Male	10	Yes	0.16356	0.02584	0.11192	0.23279
275	West	Male	11	No	0.19469	0.04002	0.12785	0.28505
276	West	Male	11	Yes	0.16965	0.02623	0.11699	0.23956
277	West	Male	12	No	0.13214	0.04542	0.06547	0.24865
278	West	Male	12	Yes	0.17494	0.02738	0.12002	0.24792
279	West	Male	13	No	0.19947	0.04814	0.12127	0.31029
280	West	Male	13	Yes	0.16217	0.02773	0.10747	0.23732
281	West	Male	14	No	0.10759	0.03838	0.05220	0.20880
282	West	Male	14	Yes	0.16487	0.02644	0.11214	0.23582
283	West	Male	15	No	0.18459	0.05348	0.10138	0.31235
284	West	Male	15	Yes	0.17018	0.02480	0.11996	0.23578
285	West	Male	16	No	0.19757	0.04862	0.11892	0.30993
286	West	Male	16	Yes	0.17888	0.02540	0.12718	0.24569

Obs	Region	Gender	Age	Smoothed	Prevalence	Std Error	95 % Conf Interval – Lower Bound	95 % Conf Interval – Upper Bound
287	West	Male	17	No	0.18078	0.04735	0.10548	0.29227
288	West	Male	17	Yes	0.19218	0.04291	0.11118	0.31153

APPENDIX H.
ALTERNATIVE LONGITUDINAL ACTIVITY PATTERN ALGORITHM

This page intentionally left blank.

ICF
CONSULTING

TECHNICAL MEMORANDUM

To: Ted Palma, US EPA

From: Arlene Rosenbaum and Jonathan Cohen

Date: November 4, 2004

Re: Evaluation of a multi-day activity pattern algorithm for creating longitudinal activity patterns.

BACKGROUND

In previous work ICF reviewed the HAPEM4 modeling approach for developing annual average activity patterns from the CHAD database and recommended an approach to improve the model's pattern selection process to better represent the variability among individuals. This section summarizes the recommended approach. (For details see the memorandum of July 23, 2002 from ICF Consulting to Ted Palma.)

Using cluster analysis, first the CHAD daily activity patterns are grouped into either two or three categories of similar patterns for each of the 30 combinations of day type (summer weekday, non-summer weekday, and weekend) and demographic group (males or females; age groups: 0-4, 5-11, 12-17, 18-64, 65+). Next, for each combination of day type and demographic group, category-to-category transition probabilities are defined by the relative frequencies of each second-day category associated with each given first-day category, where the same individual was observed for two consecutive days. (Consecutive day activity pattern records for a single individual constitute a small subset of the CHAD data.)

To implement the proposed algorithm, for each day type and demographic group, one daily activity pattern per category is randomly selected from the corresponding CHAD data to represent that category. That is, if there are 3 cluster categories for each of 3 day types, 9 unique activity patterns are selected to be averaged together to create an annual average activity pattern to represent an individual in a given demographic group and census tract.

The weighting for each of the 9 activity patterns used in the averaging process is determined by the product of two factors. The first is the relative frequency of its day type, i.e., 0.18 for summer weekdays, 0.54 for non-summer weekdays, and 0.28 for weekends.

The second factor in the weighting for the selected activity pattern is determined by simulating a sequence of category-types as a one-stage Markov chain process using the transition probabilities. The category for the first day is selected according to the relative frequencies of each category. The category for the second day is selected according to the category-to-category transition probabilities for the category selected for the first day. The category for the third day is selected according to the transition probabilities for the category selected for the second day. This is repeated for all days in the day type (65 for summer weekdays, 195 for non-summer weekdays, 104 for weekends), producing a sequence of daily categories. The relative frequency of the category-type in the sequence associated with the selected activity pattern is the second factor in the weighting.

PROPOSED ALGORITHM STEPS

The proposed algorithm is summarized in Figure 1. Each step is explained in this section.

Data Preparation

Step 1: Each daily activity pattern in the CHAD data base is summarized by the total minutes in each of five micro-environments: indoors – residence; indoors – other building; outdoors – near road; outdoors – away from road; in vehicle. These five numbers are assumed to represent the most important features of the activity pattern for their exposure impact.

Step 2: All CHAD activity patterns for a given day-type and demographic group are subjected to cluster analysis, resulting in 2 or 3 cluster categories. Each daily activity pattern is tagged with a cluster category.

Step 3: For each day-type and demographic group, the relative frequency of each day-type in the CHAD data base is determined.

Step 4: All CHAD activity patterns for a given day-type and demographic group that are consecutive days for a single individual, are analyzed to determine the category-to-category transition frequencies in the CHAD data base. These transition frequencies are used to calculate category-to-category transition probabilities.

For example, if there are 2 categories, A and B, then

PAA = the probability that a type A pattern is followed by a type A pattern,
PAB = the probability that a type A pattern is followed by a type B pattern (PAB = 1 – PAA),
PBB = the probability that a type B pattern is followed by a type B pattern, and
P_{BA} = the probability that a type B pattern is followed by a type A pattern (P_{BA} = 1 – P_{BB}).

Activity Pattern Selection

For each day-type and demographic group in each census tract

Step 5: One activity pattern is randomly selected from each cluster category group (i.e., 2 to 3 activity patterns)

Creating Weights for Day-type Averaging

For each day-type and demographic group in each census tract

Step 6: A cluster category is selected for the first day of the day-type sequence, according to the relative frequency of the cluster category days in the CHAD data set.

Step 7: A cluster category is selected for each subsequent day in the day-type sequence day by day using the category-to-category transition probabilities.

Step 8: The relative frequency of each cluster category in the day-type sequence is determined.

Step 9: The activity patterns selected for each cluster category (Step 5) are averaged together using the cluster category frequencies (Step 8) as weights, to create a day-type average activity pattern.

Creating Annual Average Activity Patterns

For each demographic group in each census tract

Step 10: The day-type average activity patterns are averaged together using the relative frequency of day-types as weights, to create an annual average activity pattern.

Creating Replicates

For each demographic group in each census tract

Step 11: Steps 5 through 10 are repeated 29 times to create 30 annual average activity patterns.

EVALUATING THE ALGORITHM

The purpose of this study is to evaluate how well the proposed one-stage Markov chain algorithm can reproduce observed multi-day activity patterns with respect to demographic group means and inter-individual variability, while using one-day selection.

In order to accomplish this we propose to apply the algorithm to observed multi-day activity patterns provided by the WAM, and compare the means and variances of the predicted multi-day patterns with the observed patterns.

Current APEX Algorithm

Because the algorithm is being considered for incorporation into APEX, we would like the evaluation to be consistent with the approach taken in APEX for selection of activity patterns for creating multi-day sequences. The APEX approach for creating multi-day activity sequences is as follows.

Step1: A profile for a simulated individual is generated by selection of gender, race (not implemented?), age group, and home sector from a given set of distributions consistent with the population of the study area.

Step 2: A specific age within the age group is selected from a uniform distribution.

Step 3: The employment status is simulated as a function of the age.

Step 4: For each simulated day, the user defines an initial pool of possible diary days based on a user-specified function of the day type (e.g., weekday/weekend) and temperature.

Step 5: The pool is further restricted to match the target gender and employment status exactly and the age within 2A years for some parameter A. The diary days within the pool are assigned a weight of 1 if the age is within A years of the target age and a weight of w (user-defined parameter) if the age difference is between A and 2A years. For each simulated day, the probability of selecting a given diary day is equal to the age weight divided by the total of the age weights for all diary days in the pool for that day.

Approach to Incorporation of Day-to-Day Dependence into APEX Algorithm

If we were going to incorporate day-to-day dependence of activity patterns into the APEX model, we would propose preparing the data with cluster analysis and transition probabilities as described in Steps 1-4 for the proposed HAPEM 5 algorithm, with the following modifications.

- For Step 2 the activity patterns would be divided into groups based on day-type (weekday, weekend), temperature, gender, employment status, and age, with cluster analysis applied to each group. However, because the day-to-day transitions in the APEX activity selection algorithm can cross temperature bins, we would propose to use broad temperature bins for the clustering and transition probability calculations so that the cluster definitions would be fairly uniform across temperature bins. Thus we would probably define the bins according to season (e.g., summer, non-summer).

- In contrast to HAPEM, the sequence of activity patterns may be important in APEX. Therefore, for Step 4 transition probabilities would be specified for transitions between days with the same day-type and season, as in HAPEM, and also between days with different day-types and/or seasons. For example, transition probabilities would be specified for transitions between summer weekdays of each category and summer weekends of each category.

Another issue for dividing the CHAD activity records for the purposes of clustering and calculating transition probabilities is that the diary pools specified for the APEX activity selection algorithm use varying and overlapping age ranges. One way to address this problem would be to simply not include consideration of age in the clustering process, under the assumption that cluster categories are similar across age groups, even if the frequency of each cluster category varies by age group. This assumption could be tested by examination of the cluster categories stratified by age group that were developed for HAPEM5. If the assumption is found to be valid, then the cluster categories could be pre-determined for input to APEX, while the transition probabilities could be calculated within APEX during the simulation for each age range specified for dairy pools.

If the assumption is found to be invalid, then an alternative approach could be implemented that would create overlapping age groups for purposes of clustering as follows. APEX age group ranges and age window percentages would be constrained to some maximum values. Then a set of overlapping age ranges that would be at least as large as the largest possible dairy pool age ranges would be defined for the purposes of cluster analysis and transition probability calculation. The resulting sets of cluster categories and transition probabilities would be pre-determined for input into APEX and the appropriate set used by APEX for each diary pool used during the simulation.

The actual activity pattern sequence selection would be implemented as follows. The activity pattern for first day in the year would be selected exactly as is currently done in APEX, as described above. For the selecting the second day's activity pattern, each age weight would be multiplied by the transition probability P_{AB} where A is the cluster for the first day's activity pattern and B is the cluster for a given activity pattern in the available pool of diary days for day 2. (Note that day 2 may be a different day-type and/or season than day 1.) The probability of selecting a given diary day on day 2 is equal to the age weight times P_{AB} divided by the total of the products of age weight and P_{AB} for all diary days in the pool for day 2. Similarly, for the transitions from day 2 to day 3, day 3 to day 4, etc.

Testing the Approach with the Multi-day Data set

We tested this approach using the available multi-day data set. For purposes of clustering we characterized the activity pattern records according to time spent in each of 5 microenvironments: indoors-home, indoors-school, indoors-other, outdoors (aggregate of the 3 outdoor microenvironments), and in-transit.

For purposes of defining diary pools and for clustering and calculating transition probabilities we divided the activity pattern records by day type (i.e., weekday, weekend), season (i.e., summer or ozone season, non-summer or non-ozone season), age (6-10 and 11-12), and gender. Since all the subjects are 6-12 years of age and all are presumably unemployed, we need not account for differences in employment status. For each day type, season, age, and gender, we found that the activity patterns appeared to group in three clusters.

In this case, we simulated week-long sequences (Wednesday through Tuesday) for each of 100 people in each age/gender group for each season, using the transition probabilities. To evaluate the algorithm we calculated the following statistics for the predicted multi-day activity patterns for comparison with the actual multi-day diary data.

- For each age/gender group for each season, the average time in each microenvironment
- For each age/gender group, season, and microenvironment, the average of the within-person variance across all simulated persons (We defined the within-person variance as the variance of the total time per day spent in the microenvironment across the week.)
- For each age/gender group, season, and microenvironment, the variance across persons of the mean time spent in that microenvironment

In each case we compared the predicted statistic for the stratum to the statistic for the corresponding stratum in the actual diary data.[14]

We also calculated the mean normalized bias for the statistic, which is a common performance measure used in dispersion model performance and which is calculated as follows.

$$NBIAS = \frac{100}{N}\sum_{1}^{N}\frac{(predicted - observed)}{observed}\ \%$$

RESULTS

Comparisons of simulated and observed data for time in each of the 5 microenvironments are presented in Tables 1 – 3 and Figures 2-5.

Average Time in Microenvironment

Table 1 and Figure 2 show the comparisons for the average time spent in each of the 5 microenvironments for each age/gender group and season. Figure 3 shows the comparison for all the microenvironments except indoor, home in order to highlight the lower values.

Table 1 and the figures show that the predicted time-in-microenvironment averages match well with the observed values. For combinations of microenvironment/age/gender/season the normalized bias ranges from –35% to +41%. Sixty percent of the predicted averages have bias between –9% and +9%, and the mean bias across any microenvironment ranges from -9% to +4%. Fourteen predictions have positive bias and 23 have negative bias. A Wilcoxon signed rank test that the median bias across the 40 combinations = 0 % was not significant (p-value = 0.40) supporting the conclusion of no overall bias.

Variance Across Persons

Table 2 and Figure 4 show the comparisons for the variance across persons for the average time spent in each microenvironment. In this case the bias ranges from –40% to +120% for any microenvironment/age/gender/season. Sixty-five percent of the predicted variances have bias between –22% and +24%. The mean normalized bias across any microenvironment ranges from –10% to +28%. Eighteen predictions have positive bias and 20 have negative bias. Figure 4 suggests a reasonably good match of predicted to observed variance in spite of 2 or 3 outliers. A Wilcoxon signed rank test that the median bias across the 40 combinations = 0 % was not significant (p-value = 0.93) supporting the conclusion of no overall bias.

Within-Person Variance for Persons

Table 3 and Figure 5 show the comparisons for the within-person variance for time spent in each microenvironment. In this case the bias ranges from –47% to +150% for any

[14] For the diary data, because the number of days per person varies, the average of the within-person variances was calculated as a weighted average, where the weight is the degrees of freedom, i.e., one less than the number of days simulated. Similarly, the variance across persons of the mean time was appropriately adjusted for the different degrees of freedom using analysis of variance.

microenvironment/age/gender/season. Seventy percent of the predicted variances have bias between –25% and +30%. The mean normalized bias across any microenvironment ranges from –11% to +47%. Twenty-eight predictions have positive bias and 12 have negative bias, suggesting some tendency for overprediction of this variance measure. And indeed a Wilcoxon signed rank test that the median bias across the 40 combinations = 0 % was very significant (p-value = 0.01) showing that the within-person variance was significantly overpredicted. Still, Figure 4 suggests a reasonably good match of predicted to observed variance in most cases, with a few overpredicting outliers at the higher end of the distribution. So although the positive bias is significant in a statistical sense (i.e., the variance is more likely to be overpredicted than underpredicted), it is not clear whether the bias is large enough to be important.

CONCLUSIONS

The proposed algorithm appears to be able to replicate the observed data reasonably well, although the within-person variance is somewhat overpredicted.

It would be informative to compare this algorithm with the earlier alternative approaches in order to gain perspective on the degree of improvement, if any, afforded by this approach. Two earlier approaches were:

1. Select a single activity pattern for each day-type/season combination from the appropriate set, and use that pattern for every day in the multi-day sequence that corresponds to that day-type and season.
2. Re-select an activity pattern for each day in the multi-day sequence from the appropriate set for the corresponding day-type and season.

Goodness-of-fit statistics could be developed to compare the three approaches and find which model best fits the data for a given stratum.

Table 1. Average time spent in each microenvironment: comparison of predicted and observed.

Microenvironment	Demographic Group	Season	Observed (hours/day)	Predicted (hours/day)	Normalized Bias
Indoor, home	Girls, 6-10	Summer	15.5	16.5	6%
		Not Summer	15.8	15.5	-2%
	Boys, 6-10	Summer	15.7	15.2	-3%
		Not Summer	15.8	16.4	4%
	Girls, 11-12	Summer	16.2	15.3	-5%
		Not Summer	16.5	16.5	0%
	Boys, 11-12	Summer	16.0	15.6	-3%
		Not Summer	16.2	16.1	-1%
	MEAN				-1%
Indoor, school	Girls, 6-10	Summer	0.7	0.7	-9%
		Not Summer	2.3	2.5	7%
	Boys, 6-10	Summer	0.8	0.5	-34%
		Not Summer	2.2	2.2	0%
	Girls, 11-12	Summer	0.7	0.7	6%
		Not Summer	2.1	2.4	13%
	Boys, 11-12	Summer	0.6	0.9	38%
		Not Summer	2.4	2.7	11%
	MEAN				4%
Indoor, other	Girls, 6-10	Summer	2.9	2.4	-14%
		Not Summer	2.4	2.7	13%
	Boys, 6-10	Summer	2.2	2.7	21%
		Not Summer	1.9	1.8	-3%
	Girls, 11-12	Summer	2.2	1.6	-25%
		Not Summer	2.2	2.1	-2%
	Boys, 11-12	Summer	2.3	2.2	-5%
		Not Summer	1.9	2.0	4%
	MEAN				-2%
Outdoors	Girls, 6-10	Summer	3.7	3.5	-6%
		Not Summer	2.5	2.5	0%
	Boys, 6-10	Summer	4.1	4.3	4%
		Not Summer	3.1	2.7	-12%
	Girls, 11-12	Summer	3.7	5.2	41%
		Not Summer	2.3	2.1	-5%
	Boys, 11-12	Summer	3.9	4.3	9%
		Not Summer	2.6	2.4	-7%
	MEAN				3%

In-vehicle	Girls, 6-10	Summer	1.1	0.9	-20%
		Not Summer	1.0	0.9	-13%
	Boys, 6-10	Summer	1.1	1.3	13%
		Not Summer	1.0	0.9	-16%
	Girls, 11-12	Summer	1.2	1.1	-12%
		Not Summer	0.9	0.8	-15%
	Boys, 11-12	Summer	1.1	1.0	-5%
		Not Summer	0.9	0.8	-7%
	MEAN				-9%

Table 2. Variance across persons for time spent in each microenvironment: comparison of predicted and observed.

Microenvironment	Demographic Group	Season	Observed (hours/day)^2	Predicted (hours/day)^2	Normalized Bias
Indoor, home	Girls, 6-10	Summer	70	42	-40%
		Not Summer	67	60	-9%
	Boys, 6-10	Summer	54	49	-9%
		Not Summer	35	30	-12%
	Girls, 11-12	Summer	56	47	-17%
		Not Summer	42	38	-10%
	Boys, 11-12	Summer	57	63	12%
		Not Summer	39	42	8%
	MEAN				-10%
Indoor, school	Girls, 6-10	Summer	6.0	5.2	-13%
		Not Summer	9.5	5.9	-38%
	Boys, 6-10	Summer	5.6	3.8	-32%
		Not Summer	5.3	8.2	53%
	Girls, 11-12	Summer	4.9	5.5	11%
		Not Summer	5.4	5.3	-1%
	Boys, 11-12	Summer	5.6	6.0	6%
		Not Summer	9.2	11	23%
	MEAN				1%
Indoor, other	Girls, 6-10	Summer	46	32	-30%
		Not Summer	44	46.	6%
	Boys, 6-10	Summer	34	33	-4%
		Not Summer	23	16	-27%
	Girls, 11-12	Summer	21	18	-15%
		Not Summer	28	22	-22%
	Boys, 11-12	Summer	33	31	-6%
		Not Summer	30	30	0%
	MEAN				-12%
Outdoors	Girls, 6-10	Summer	17	23	37%
		Not Summer	9.3	6.8	-27%
	Boys, 6-10	Summer	17	18	3%
		Not Summer	8.3	7.6	-8%
	Girls, 11-12	Summer	22	22	0%
		Not Summer	9.0	9.1	1%
	Boys, 11-12	Summer	13	29	120%
		Not Summer	10	11	8%
	MEAN				17%

In-vehicle	Girls, 6-10	Summer	1.9	2.3	24%
		Not Summer	1.8	1.6	-11%
	Boys, 6-10	Summer	2.5	4.7	93%
		Not Summer	1.5	1.6	9%
	Girls, 11-12	Summer	3.5	4.7	34%
		Not Summer	2.8	2.0	-28%
	Boys, 11-12	Summer	3.2	5.4	69%
		Not Summer	1.3	1.7	35%
	MEAN				28%

Table 3. Average within person variance for time spent in each microenvironment: comparison of predicted and observed.

Microenvironment	Demographic Group	Season	Observed (hours/day)^2	Predicted (hours/day)^2	Normalized Bias
Indoor, home	Girls, 6-10	Summer	20	29	49%
		Not Summer	18	23	25%
	Boys, 6-10	Summer	17	30	75%
		Not Summer	15	24	64%
	Girls, 11-12	Summer	22	42	93%
		Not Summer	22	25	13%
	Boys, 11-12	Summer	21	24	16%
		Not Summer	17	24	38%
	MEAN				47%
Indoor, school	Girls, 6-10	Summer	2.3	2.4	5%
		Not Summer	7.3	6.4	-12%
	Boys, 6-10	Summer	2.0	1.5	-25%
		Not Summer	6.7	5.8	-14%
	Girls, 11-12	Summer	1.7	2.1	29%
		Not Summer	7.4	7.6	3%
	Boys, 11-12	Summer	1.4	2.9	101%
		Not Summer	7.3	7.8	6%
	MEAN				12%
Indoor, other	Girls, 6-10	Summer	14	14	-4%
		Not Summer	14	18	30%
	Boys, 6-10	Summer	12	17	42%
		Not Summer	10	13	26%
	Girls, 11-12	Summer	10	10	1%
		Not Summer	14	15	7%
	Boys, 11-12	Summer	11	14	26%
		Not Summer	12	13	7%
	MEAN				17%
Outdoors	Girls, 6-10	Summer	8.4	9.5	13%
		Not Summer	3.4	3.2	-3%
	Boys, 8-10	Summer	6.7	9.5	42%
		Not Summer	3.4	4.4	28%
	Girls, 11-12	Summer	10	25	150%
		Not Summer	4.0	4.5	11%
	Boys, 11-12	Summer	9.2	7.4	-20%
		Not Summer	4.3	3.7	-15%
	MEAN				26%

In-vehicle	Girls, 6-10	Summer	1.0	0.90	-13%
		Not Summer	0.90	0.48	-47%
	Boys, 6-10	Summer	1.1	1.4	31%
		Not Summer	0.81	0.71	-12%
	Girls, 11-12	Summer	1.3	1.3	4%
		Not Summer	1.3	1.1	-16%
	Boys, 11-12	Summer	2.4	1.6	-34%
		Not Summer	0.85	0.85	1%
	MEAN				-11%

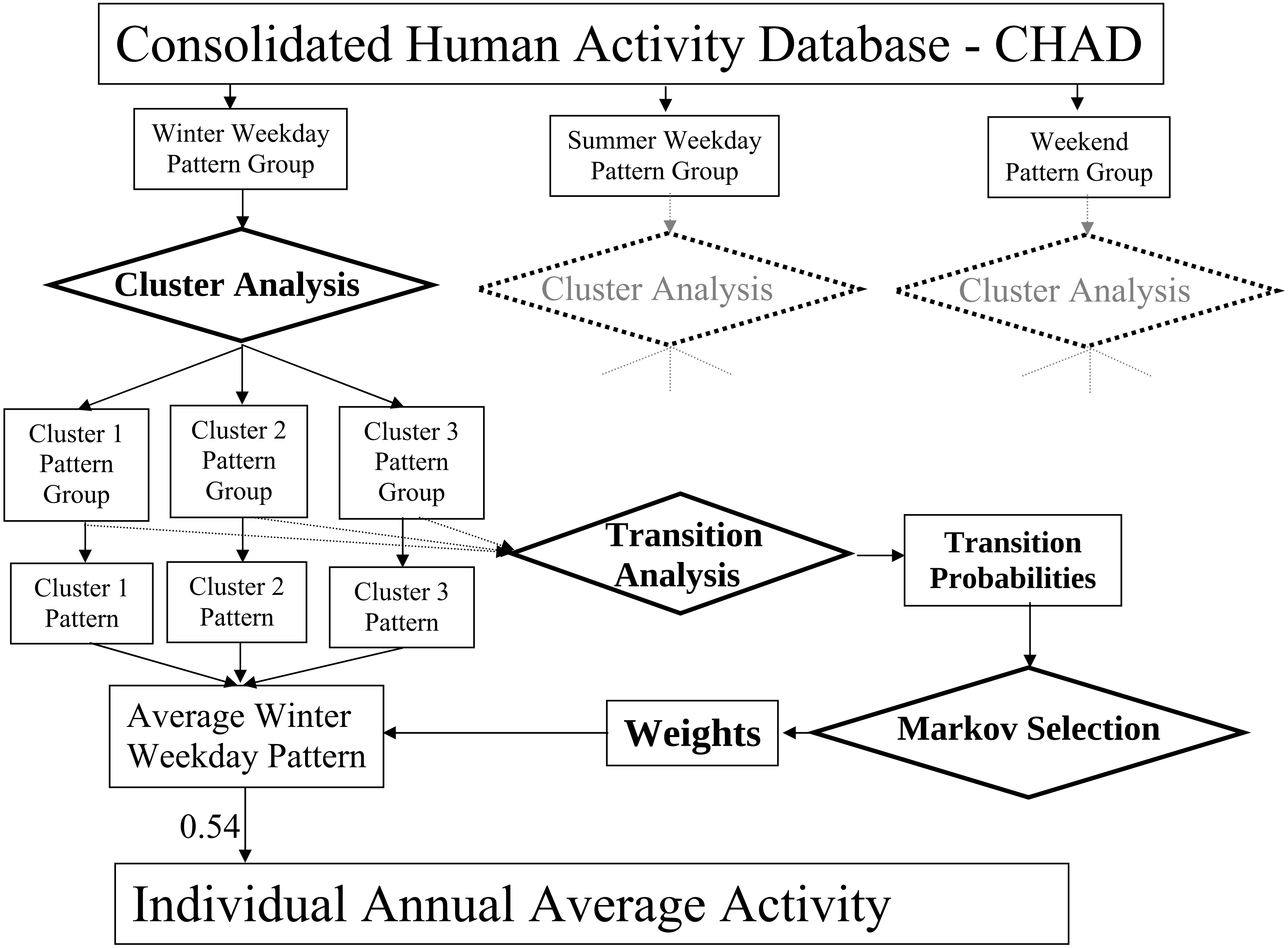


Figure 1. Flow diagram of proposed algorithm for creating annual average activity patterns for HAPEM5.

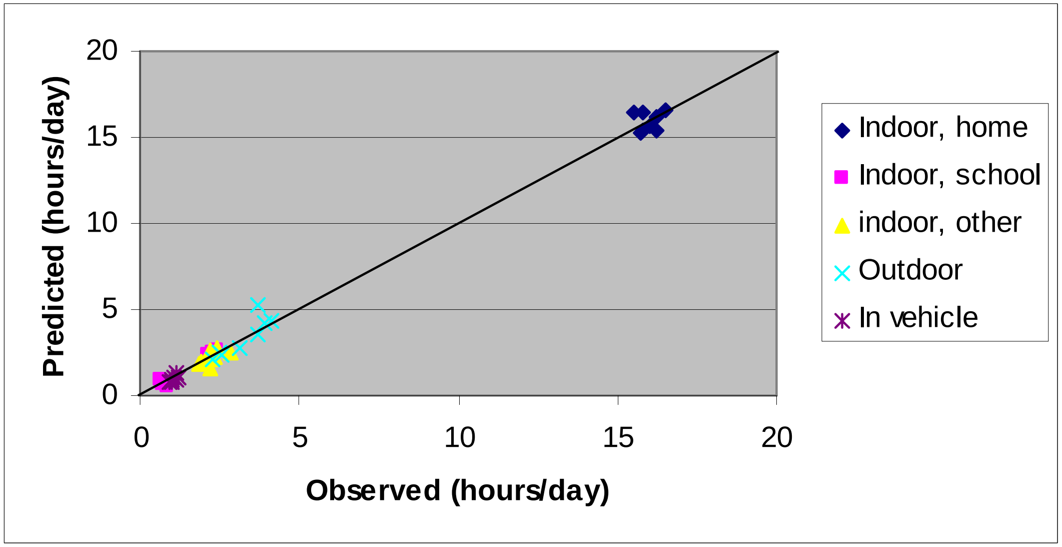


Figure 2. Comparison of predicted and observed average time in each of 5 microenvironments for age/gender groups and seasons.

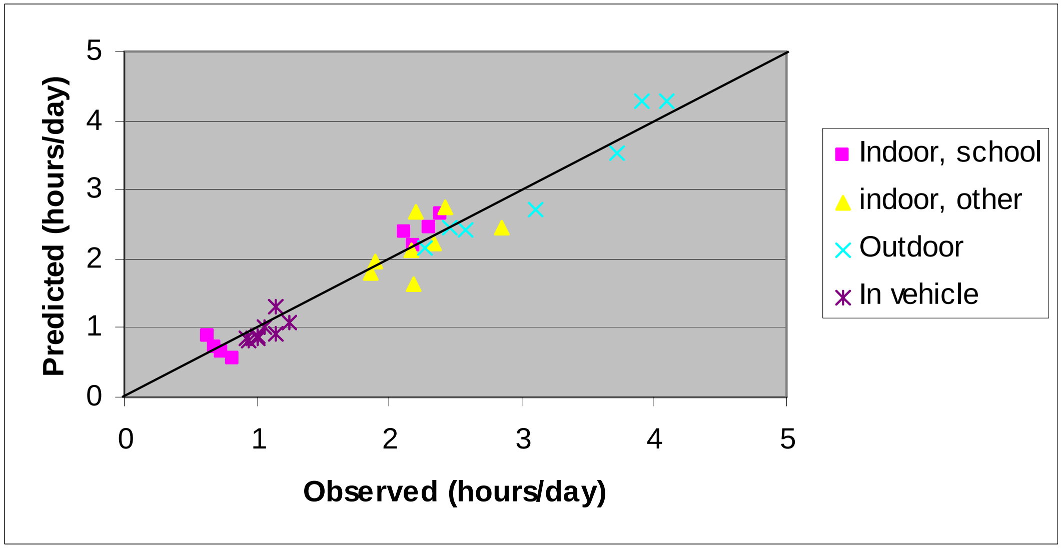


Figure 3. Comparison of predicted and observed average time in each of 4 microenvironments for age/gender groups and seasons.

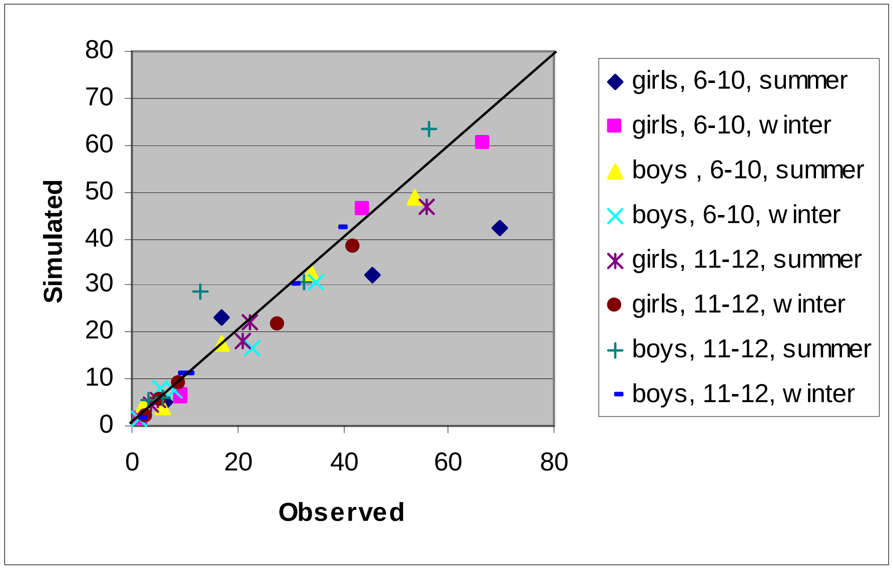


Figure 4. Comparison of predicted and observed variance across persons for time spent in each of 5 microenvironments for age/gender groups and seasons.

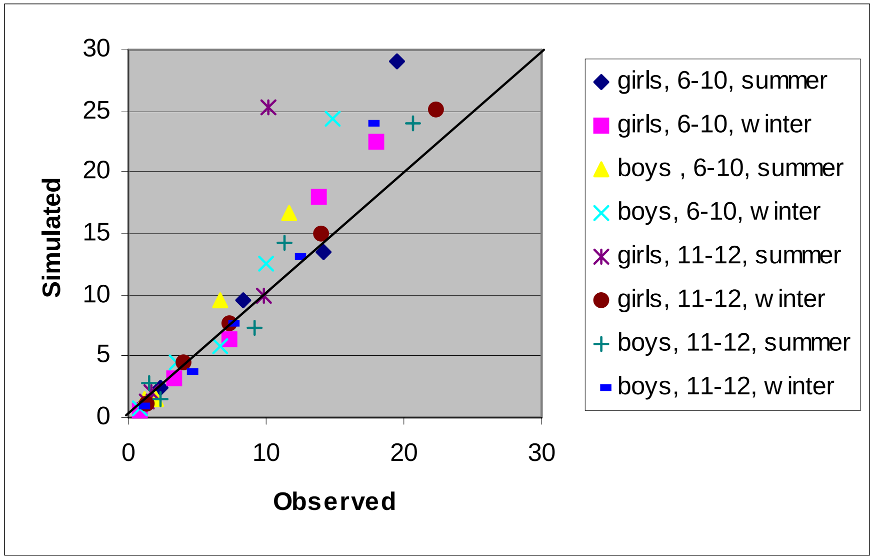


Figure 5. Comparison of predicted and observed the average within-person variance for time spent in each of 5 microenvironments by age/gender groups and seasons.

APPENDIX I. ESTIMATING NEAR ROADWAY POPULATIONS

This page intentionally left blank.

MEMORANDUM

To: Chad Bailey

From: Arlene Rosenbaum and Kevin Wright

Date: December 28, 2005

Re: Estimating near roadway populations and areas for HAPEM6

PURPOSE AND BACKGROUND

In its 2001 regulation of mobile source air toxics (the “MSAT Rule”) EPA’s Office of Transportation and Air Quality (OTAQ) committed to further study of the range of concentrations to which people are exposed for consideration in future rulemaking. As part of the Technical Analysis Plan outlined in that research, OTAQ undertook research activity looking at the air quality in immediate proximity of busy roadways and highways. Concentrations of pollutants directly emitted by motor vehicles show statistically significant elevation in concentrations with increased proximity to busy roadways.

The Hazardous Air Pollutant Exposure Model (HAPEM) is a screening-level exposure model appropriate for assessing average long-term inhalation exposures of the general population, or a specific sub-population, over spatial scales ranging from urban to national. HAPEM uses the general approach of tracking representatives of specified demographic groups as they move among indoor and outdoor microenvironments and among geographic locations. The estimated pollutant concentrations in each microenvironment visited are combined into a time-weighted average concentration, which is assigned to members of the demographic group.

Indoor microenvironment concentrations are estimated by applying scalar factors to outdoor tract concentrations, which are some of the required inputs. These scalar factors are derived from published studies of concurrent concentration measurements indoors and outdoors.

In the previous version, HAPEM5, if only a single outdoor concentration is provided for each Census tract, as is typical, this concentration is assumed to uniformly apply to the entire Census tract. For this version, HAPEM6, we refined the model to account for the spatial variability of outdoor concentrations within a tract due to enhanced outdoor concentrations of onroad mobile source pollutants at locations near major roadways. The term “major roadway” is used to describe a “Limited Access Highway”, “Highway”, “Major Road” or “Ramp”, as defined by the Census Feature Class Codes (CFCC). The new version of HAPEM more accurately reflects the average and variability of exposure concentrations within each Census tract by accounting for some of the spatial variability in the outdoor concentrations within the tract, and by extension some of the spatial variability in indoor concentrations within the tract.

Accomplishing this refinement to HAPEM required several activities, including the development and implementation of an approach for creating a database of the fraction of people within each US Census tract living near major roadways. This memorandum describes that activity.

OVERVIEW AND SPECIFICATIONS

The objective of this task was to estimate the fraction of people in each of 6 demographic groups in each US Census tract living near major roadways.

The basic analysis was conducted at the US Census block level for populations stratified by age, gender, and race/ethnicity. The block level data was then aggregated up to the tract level for populations stratified by age only for use in HAPEM6.

The data bases used for this task were:

- The Environmental Sciences Research Center (ESRI) StreetMap US roadway geographic database (which includes NavTech, GDT and TeleAtlas rectified street data)
- A geographic database of US Census block boundaries, extracted using the PCensus 2000 Census data extraction tool for Census file SF1
- A geographic data for US Census block boundaries in Puerto Rico and the US Virgin Islands obtained from Proximity

Although the block file is an intermediate product for this project, it will be retained to facilitate the re-specification of demographic groups for possible future analyses. Therefore, this file contains the most resolved age-gender groups available at the block level from the US Census STF1. The age groups for the block level data are as follows:

- 19 single-year age groups from 0-19 (P14)
- 2 single-year age groups 20-21 (P12)
- 16 age groups (P12)
    - 22 to 24 years
    - 25 to 29 years
    - 30 to 34 years
    - 35 to 39 years
    - 40 to 44 years
    - 45 to 49 years
    - 50 to 54 years
    - 55 to 59 years
    - 60 and 61 years
    - 62 to 64 years
    - 65 and 66 years
    - 67 to 69 years
    - 70 to 74 years
    - 75 to 79 years
    - 80 to 84 years
    - 85 years and over.

The aggregated age groups for the tract level data are:

- 0-1
- 2-4
- 5-15
- 16-17
- 18-64
- 65+

The race/ethnic groups (block level only) are:

- non-Hispanic White (alone or in combination - P010003)
- non-Hispanic Black (alone or in combination - P010004)
- non-Hispanic American Indian /Alaskan Native (alone or in combination - P010005)
- non-Hispanic Asian (alone or in combination - P010006)
- non-Hispanic Native Hawaiian/ Pacific Isalander (P010007)
- non-Hispanic other (alone or in combination - P010008)
- Hispanic (alone or in combination - P010009)

The spatial stratifications of the populations (block and tract level) are:

- Those residing within 75 meters of a major roadway
- Those residing from 75 to 200 meters from a major roadway
- Those residing at greater than 200 meters from a roadway.

In addition, the fraction of the area of each Census block and tract that is located within the same distance ranges from a major roadway was determined.

PROCEDURES

For all the spatial modeling and geoprocessing operations in this study ICF utilized ArcInfo software. ArcInfo is the most extensive version of ArcGIS 9.1, the industry's standard for Geographic Information Systems, produced by ESRI of Redlands, CA.

Due to the size of the roadway and block geography files, most of the processing was conducted on a county-by-county basis. The files for some counties, however, still exceeded ArcInfo's capacity and were processed tract-by-tract. A few counties in Arizona needed special handling because even at the tract level they exceeded ArcInfo's capacity and were disaggregated into smaller pieces for processing.

1. Because populations are not generally evenly distributed within blocks, it was assumed that the block populations all reside within 150 meters of *any* road within the block of designation "local" or greater as defined by the Census Feature Class Codes (CFCC). Thus, the first step was to create a 150-meter buffer around all roadways within the block. This buffer served as a "clipped" block boundary defining the

portion of the block containing residential populations. The block population was assumed to be uniformly distributed within the "clipped" block boundary.

2. Next a 75-meter buffer and a 200-meter buffer were created around all major roadways within the block. These buffers were overlaid on the "clipped" block boundary, and the fraction of the "clipped" block area that that fell within each buffer was calculated. This area fraction was assumed to equal the population fraction that fell within each buffer, and the fractions were applied to each population stratification.

3. The 75-meter buffer and the 200-meter buffer were also overlaid on the unclipped block boundary to determine the fraction of the total block area that fell with each of the buffers.

4. The block level fractions for area and populations were then aggregated up to the tract level, and the population stratifications were aggregated up to the 6 tract age groups only.

RESULTS

The resulting database consists of 2 files types: (1) a block file for each state, and (2) a nation-wide tract file.

The block files contains the following 249 fields for each block:

- block FIPS code
- total population
- total area
- area within 75 meters of a major roadway
- area from 75 to 200 meters from a major roadway
- for each of 74 age-gender groups:
    - population residing within 75 meters of a major roadway
    - population residing between 75 and 200 meters from a major roadway
    - population residing more than 200 meters from a major roadway
- sum of race/ethnic populations (note; this may differ slightly from the total population due to some double-counting of persons with more than 1 race/ethnicity)
- for each of 7 race/ethnic groups:
    - population residing within 75 meters of a major roadway
    - population residing between 75 and 200 meters from a major roadway
    - population residing more than 200 meters from a major roadway

Note that because of the limitations of the US Census data the block level populations could not be stratified by age, gender, and race together, The tract file contains the following 22 fields for each tract

- tract FIPS code
- fraction of area within 75 meters of a major roadway
- fraction of area between 75 and 200 meters from a major roadway
- fraction of area more than 200 meters from a major roadway
- for each of 6 age groups:
    - fraction of population residing within 75 meters of a major roadway
    - fraction of population residing between 75 and 200 meters from a major roadway
    - fraction of population residing more than 200 meters from a major roadway

To date only a subset of states have been completely processed. For this subset state summaries of the fraction of population living within various distances of major roadways are presented in Table 1.

Table 1. Fraction of population residing at various distances from major roadways for selected states.

STATE	Distance from major roadways		
	< 75 meters	75 – 200 meters	> 200 meters
Colorado	**0.22**	**0.33**	**0.45**
Georgia	**0.17**	**0.24**	**0.59**
New York	**0.31**	**0.36**	**0.33**

United States
Environmental Protection
Agency

Office of Air Quality Planning and Standards
Health and Environmental Impacts Division
Research Triangle Park, NC

Publication No. EPA 452/R-07-004
January 2007