Ozone Population Exposure Analysis for Selected Urban Areas

Draft Report

Office of Air Quality Planning and Standards
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

November 2005

DISCLAIMER

This draft document has been prepared by staff from the Health and Ecosystems Effects 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, WA 3-8) and Alion Science and Technology, Inc. (through Contract No. 68-D-00-206, WA 131). 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. This document is being circulated to obtain review and comment from the Clean Air Scientific Advisory Committee (CASAC) and the general public. Comments on this document should be addressed to John E. Langstaff, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, C539-01, Research Triangle Park, North Carolina 27711 (email: langstaff.john@epa.gov).

Table of Contents

List of Tables

List of Figures

Acknowledgements

The following people contributed to writing this document.

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

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

EPA
John Langstaff
Thomas McCurdy
Harvey Richmond

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 (Ozone Criteria Document, EPA, 2005a) was released for public comment and review by EPA's Clean Air Scientific Advisory Committee (CASAC). 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; and
- Assessment of the quality and limitations of the input data for supporting the goals of the ozone NAAQS exposure analysis.

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 draft Ozone Staff Paper (EPA, 2005b). Based on these criteria, EPA has selected the following 12 urban areas for inclusion in the exposure analysis:

- 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; and
- Washington, D.C.

1.2 Exposure Periods

The exposure periods modeled were April 1 through September 30 for the most recent year for which data are available, 2004.

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), active school-age children, 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 due to their increased dose rates.

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 APEX 4, described in the APEX User's Guide (EPA, 2005c).

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.

APEX can be thought of 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, the duration of 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 predicted distributions (e.g., human activity patterns). 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. Nevertheless, the largest portion of CHAD (about 40 percent) 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 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 (e.g., 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 yearly (e.g., age). This is generally only an issue for simulations with long timeframes.

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 up to six steps, each of which is described in the sections indicated below:

1. **Characterize the study area**. APEX selects tracts (e.g., census tracts) within a study area – and thus identifies the potentially exposed population – based on the user-defined center and radius of the study area 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 (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 mapping 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

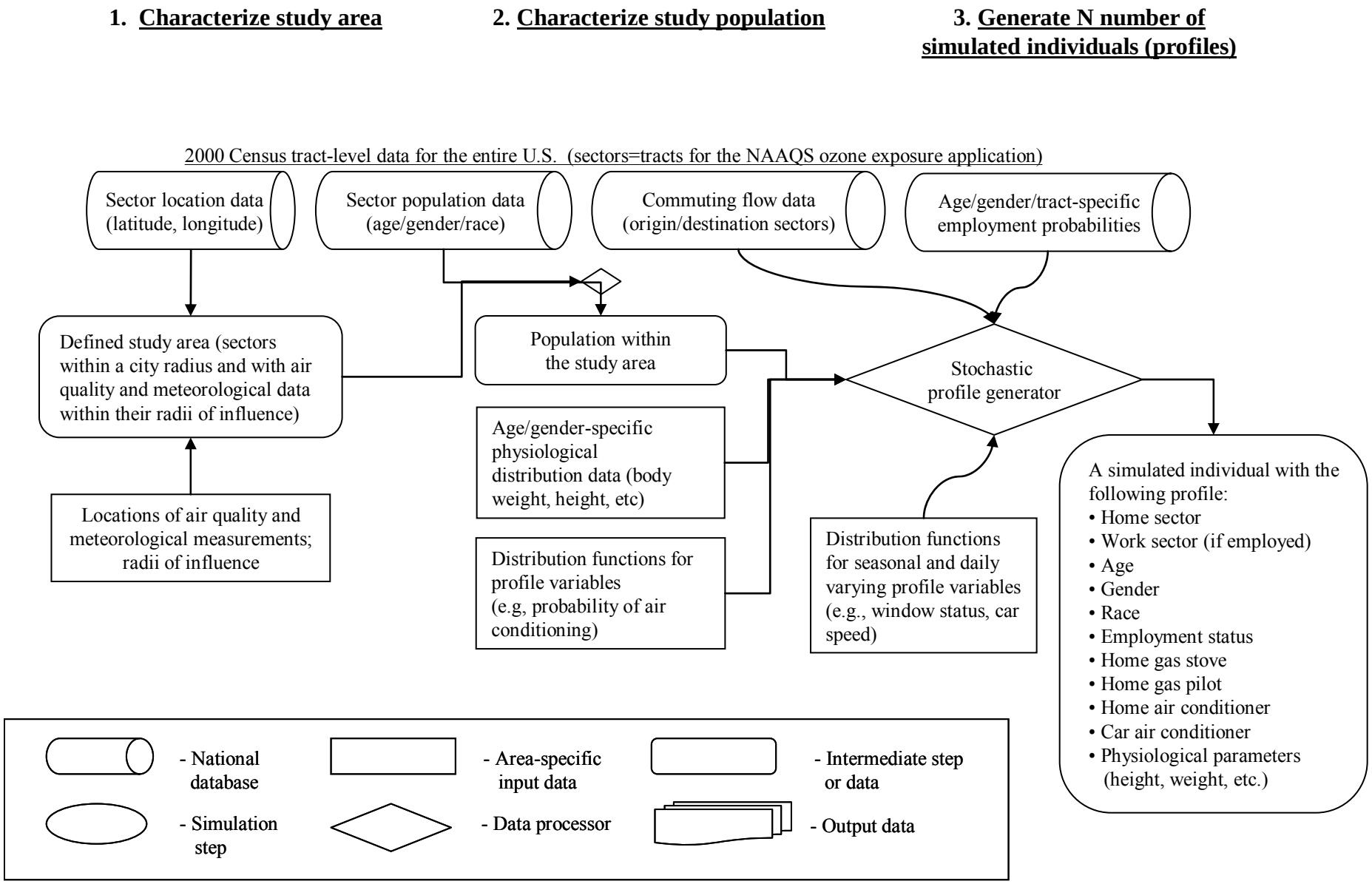


Figure 2-1. Overview of the APEX Model, continued

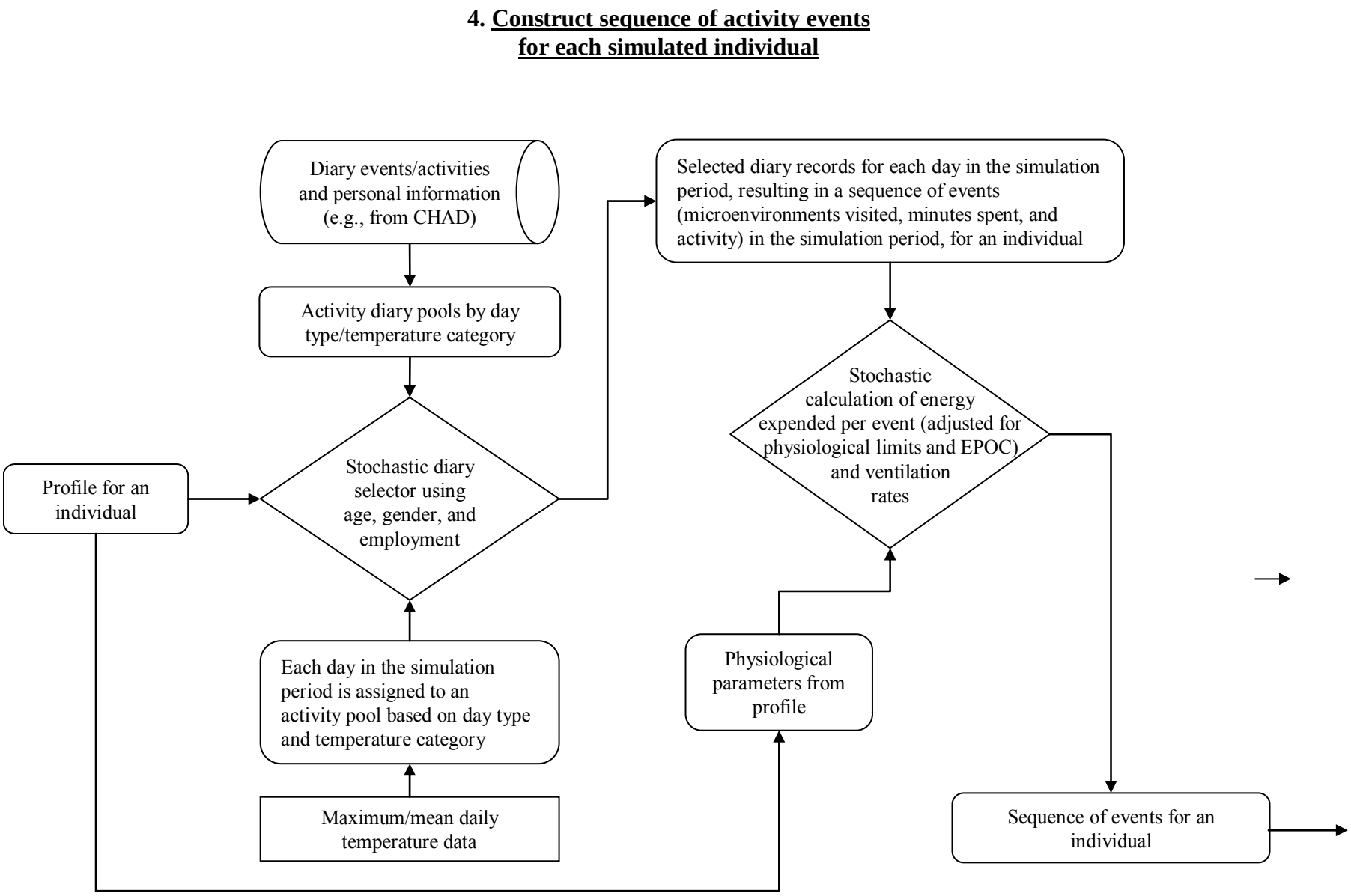
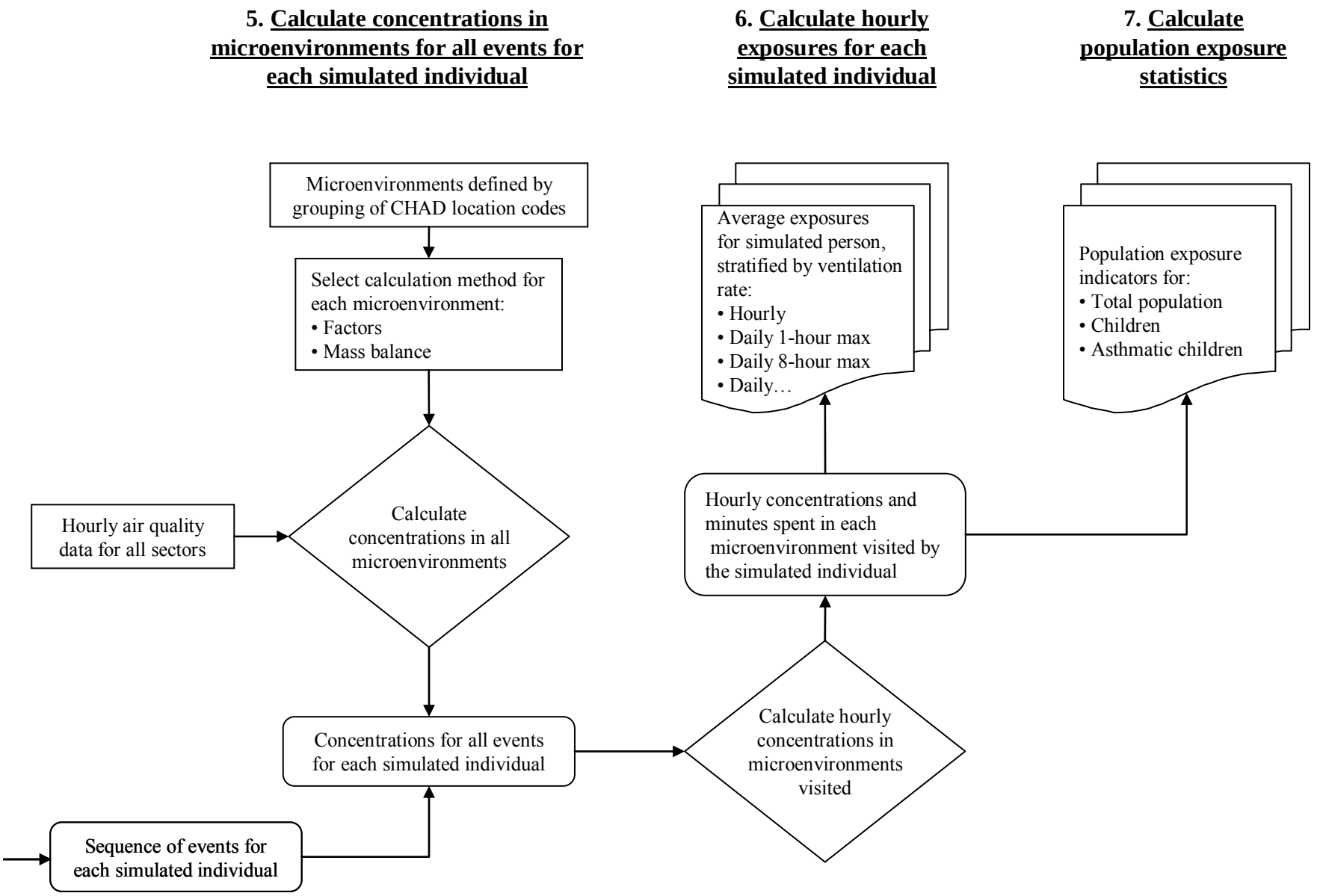


Figure 2-1. Overview of the APEX Model, concluded



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 such as consolidated metropolitan statistical areas (CMSAs). Even larger study areas are possible, 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 determines 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.

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. APEX randomly generates a value in the range of 0 to 1, assuming a uniform distribution. If this value is larger than 0.3, the simulated person will have an air conditioner. If the value is less than 0.3, the person will not have an air conditioner.

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 longitudinal diary for a simulated individual covering all days in the simulated period.

The method in APEX for creating longitudinal diaries which 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 determined to be D=0.22 and A=0.19. 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 and ventilation (respiration) rate 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. Ventilation (as discussed in Section 2.4.1 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.

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 of the subject air pollutant 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 2.4.2 and 2.4.3, 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 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:

$$V_{O2} = EE \; x \; ECF \tag{2-1}$$

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:

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

where:

METS = Metabolic equivalents 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. METS 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 METS, 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 2-1 by body mass (BM) and using equation 2-2, one obtains:

$$V_{O2} / BM = RMR \; x \; ECF \; x \; MET / BM \tag{2-3}$$

Graham and McCurdy (2004) describe an approach to estimate VE 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:

$$\ln(VE / BM)_i = b_0 + (b_1 * \ln(V_{O2} / BM_i)) + (b_2 * \ln(1 + age_i)) + (b_3 * gender_i) + eb_i + ew_i \tag{2-4}$$

where:

the V_{O2}/BM term is given in terms of the APEX variables by equation 2-3,
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 2-2. e_b and e_w have mean zero.

Table 2-2. 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

2.4.2 Excess Post-Exercise Oxygen Consumption

APEX has an algorithm for adjusting the METS values to account for excess post-exercise oxygen consumption (EPOC). This algorithm is described in Appendix B.

2.4.3 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 2-3 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 2-3. Mass Balance Model Parameters

Variable	Definition	Units	Value Range
$f_{proximity}$	Proximity factor	unitless	$f_{proximity} \geq 0$
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	$V > 0$

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

$$\frac{dC_{in}(t)}{dt} = C_{ambient}\ x\ f_{proximity}\ x\ f_{penetration}\ x\ R_{air\ exchange} = \Delta C_{in} \qquad (2\text{-}5)$$

where:

$dC_{in}(t)$, ΔC_{in} = Change in microenvironmental concentration due to influx of air (ppm/hour). Within the time period of an hour, ΔC_{in} is assumed to be constant.

t = Time

$C_{ambient}$ = Ambient hourly concentration (ppm)

$f_{proximity}$ = Proximity factor (unitless)

$f_{penetration}$ = Penetration factor (unitless)

$R_{air\ exchange}$ = Air exchange rate (1/hour)

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

$$\frac{dC_{out}(t)}{dt} = R_{air\ exchange}\ x\ C(t) \qquad (2\text{-}6)$$

where:

$dC_{out}(t)$ = Change in microenvironmental concentration due to outflux of air (ppm/hour)

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

$$\frac{dC_{removal}(t)}{dt} = (R_{deposition} + R_{filtration} + R_{chemical})C(t) = R_{removal}\, x\, C(t) \qquad (2\text{-}7)$$

where:

$dC_{removal}(t)$ = Change in microenvironmental concentration due to removal processes (ppm/hour)

$R_{deposition}$ = 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)

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

$$\Delta C_{in} = \frac{dC_{in}(t)}{dt} - \frac{dC_{out}(t)}{dt} - \frac{dC_{removal}(t)}{dt} \qquad (2\text{-}8)$$

where:

$C(t)$ = Concentration in a microenvironment at time t (ppm)

We are not modeling indoor emissions of ozone, so the optional term ΔC_{source}, which would represent an emission source inside the microenvironment, is not included.. Within the time period of an hour, *dCin* is assumed to be constant.

Equation 2-8 combined with Equations 2-6 and 2-7 leads to:

$$\Delta C_{in} = \frac{dC_{in}(t)}{dt} - R_{air\ exchange}\ C(t) - R_{removal}\ C(t) \qquad (2\text{-}9)$$

Solving the differential equation in Equation 2-9 leads to:

$$C(t) = \frac{\Delta C_{in}}{R_{mean}} + (C(0) - \frac{\Delta C_{in}}{R_{mean}}) \exp(-R_{mean} t) \quad (2\text{-}10)$$

where:

C(0) = Concentration of a pollutant in a microenvironment at the beginning of a hour (ppm)

C(t) = Concentration of a pollutant in a microenvironment at time *t* within the time period of a hour (ppm)

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

Based on Equation 2-12, the following three hourly concentrations in a microenvironment are calculated:

$$C_{equil} = C(t \to \infty) = \frac{\Delta C_{in}}{R_{mean}} \quad (2\text{-}11)$$

$$C_{hourly\ end} = C_{equil} - (C(0) - C_{equil}) \exp(-R_{mean}) \quad (2\text{-}12)$$

$$C_{hourly\ mean} = \frac{\int_0^1 C(t)dt}{\int_0^1 dt} = C_{equil} + (C(0) - C_{equil}) \frac{1 - \exp(-R_{mean})}{R_{mean}} \quad (2\text{-}13)$$

where:

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

C(0) = Concentration in a microenvironment at the beginning of each hour (ppm)

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

$C_{hourly\ mean}$ = Hourly mean concentration in a microenvironment (ppm)

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

At each hour time step of the simulation period, APEX uses Equations 2-14, 2-15, and 2-16 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 $C_{hourly\ end}$ for the previous hour as *C(0)*.

2.4.4 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 2-4 lists the parameters required by the factors method to calculate concentrations in a microenvironment without emissions sources.

Table 2-4. Factors Model Parameters

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

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

$$C_{hourly} = C_{ambient} \; x \; f_{proximity} \; x \; f_{penetration} \tag{2-14}$$

where:

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

2.4.5 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:

$$BSA = e^{-2.2781} \; BM^{0.6821} \tag{2-15}$$

where:

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

2.4.6 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:

$$Ambient\ Concentrat\ ion = LeaverMult \times avg(t) + LeaverAdd \tag{2-16}$$

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 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:

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

where:

C_i	=	Hourly exposure concentration at clock hour *I* of the simulation period ($\mu g/m^3$ or ppm)
N	=	Number of events (i.e., microenvironments visited) in clock hour *I* of the simulation period.
$C_{hourly\ (j)}$	=	Hourly concentration in microenvironment *j* ($\mu g/m^3$ or 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.

2.6 Model Output

This section provides a brief overview of the APEX output files used in this analysis. Specific output generated for the purposes of this document are discussed in Section 3.1. All of the output files used by APEX are ASCII text files. Table 2-5 lists each of the output data types and provides descriptions of their content. 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.

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, 2005c).

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 35,000, an amount which initial 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 ensure that all the counties in each study area were included, and all counties were explicitly listed in the file.

The settings which 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 ensure a wide range of diaries were selected. The variable which controls the use of additional ages outside the target age window, was set to 0.1 to further enhance variability in diary selection.

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. For school-age children, we take the diversity statistic D to be 0.19 and the autocorrelation to be 0.22. 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. We use the school-age children values for all ages, and will discuss the implications of errors in these estimates as part of the uncertainty analysis.

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 year 2004. All of the sites in AIRS within the boundaries of the CSA were used in this analysis.

3.2.2 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 not replaced.

3.2.3 Spatial Interpolation

The hourly ozone concentrations at the AIRS sites in each CSA were interpolated to a 20 by 20 km rectangular grid covering the Census tracts in the CSA using a simple inverse squared-distance weighted average for each hour. Grid points further than 75 km from the closest ozone monitor were dropped. An analysis of the uncertainty of the ozone concentrations input to APEX is being conducted and will be discussed in the next draft of this Staff Paper.

3.3 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 temperature zone location file, contains the locations of meteorological data recordings, expressed in latitude and longitude coordinates. This file also contains start and end dates for data recording. 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.

3.4 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 does 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.

3.5 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.

3.6 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.

3.6.1 Origin of Data

The Consolidated Human Activity Database (CHAD) provides comprehensive 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 highly varied structures (Table 3-1). 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.

3.6.2 CHAD Data

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.

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 contain reasonable mapping, the model will not accurately simulate exposure related to daily activities.

Table 3-1. Description 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:	Los Angeles	09/1990-	13-17	19	43	Diary	Spier et al. (1992)

High School Adolescents		10/1990					
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)
University of Michigan children	National	02/1997-12/1997	0-13	2,887	5,616	Recall; Random	www.isr.umich/edu/frc/childevelopment/home.html
Valdez, AK	Valdez metropolitan area	11/1990-10/1991	11-71	401	401	Recall; Random	Goldstein et al. (1992)
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)

Personal Information file. CHAD 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.

3.7 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 (US EPA, 2005c).

3.8 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 3-2.

Table 3-2. 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	Constant = 1
	DecayRate (DE)	1/hr	Lognormal distribution
	AirExRate (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, penetration factors and concentration sources, 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 source, like concentration source, 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. The average removal rate, which is the sum of the decay rate and air exchange rate, can be explicitly modeled or left out of 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.

3.8.1 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 what the locations where actual people spend time on a daily basis.

As discussed previously, 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 3-3.

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

Microenvironment	Calculation Method	Parameter Types used
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

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 3.8.2 and Appendix A. The development of the decay rate distribution is described in Section 3.8.3.

3.8.2 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.

The first section of the excerpt specifies the air exchange rate distributions for the

```
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
```

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, exposure functions 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 functions are discussed in detail in Appendix A. The ozone decay rate is modeled as a lognormal distribution (as shown in the last section of the

excerpt). The development of the decay rate is discussed in Section 3.8.3. The file excerpt describing this microenvironment follows this paragraph.

In the code 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 model 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 follows this paragraph.

```
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
```

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 3.8.3.

Microenvironments 8 and 9: Outdoor microenvironments. The two outdoor microenvironments, which cover Near Road and Public Garage/Parking Lot environments, are different from the indoor microenvironments in that they use the simpler factors method to calculate pollutant exposure. Proximity factors were developed to estimate exposures in these microenvironments. Penetration factors were not used as air exchange in an outdoor environment is generally expected to result in sufficient atmospheric mixing. The optional concentration source variable is not relevant to ozone studies and was not included. 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-Other. The outdoors, other ozone concentrations should be well represented by the ambient monitors. Therefore the penetration factor and proximity factor for this microenvironment were set to the default value of 1, which eliminates their effect on microenvironment concentrations.

Microenvironments 11 and 12: In Vehicle- Cars and Trucks, and Mass Transit. Both of the In Vehicle 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 proximity factor distribution was developed using the inside-vehicle to outside-vehicle ratios from the Cincinnati ozone study previously mentioned (Johnson et al, 1995). Three penetration 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 3.9. Again, these distributions were developed based on the previously mentioned Cincinnati ozone study.

Micro number = 11 ! Cars & trucks
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	0.1	1.0	0.300	0.232

Micro number = 11
Parameter Type = PE
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

3.8.3 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.

3.8.4 Microenvironment Mapping

The purpose of the *Microenvironment Mapping* file is to match the APEX Microenvironments to CHAD Location codes. Table 3-4 gives the mapping used for the APEX simulations.

Table 3-4. 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
```

3.9 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 manually 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 Federal Highway Administration data. 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 of 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 aregeographical 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) which 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.

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 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 results at the station. Pollutant levels are highly

dependent on weather and wind, and other unknowns may effect how well the data represent pollutant concentrations in the area. In addition, because APEX only uses hourly average ozone concentrations, the model does not have a more temporally refined pollutant concentration record, which may affect 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. 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.

6. 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. February.

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. 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, 2000.

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 (May). 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.

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. 2005. Personal communication from Tom McCurdy to Dave McKee, April 6, 2005. Correction to McCurdy (2000). "The lower bound of moderately active PAI should be 1.76, not 1.75."

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 (2000).

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 (2005a). Review of National Ambient Air Quality Standards for Ozone: Assessment of Scientific and Technical Information - OAQPS Staff Paper (first draft). 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.

U.S. Environmental Protection Agency (2005b). Air Quality Criteria for Ozone and Other Related Photochemical Oxidants. Second External Review Draft. National Center for Environmental Assessment, U.S. Environmental Protection Agency, Research Triangle Park, NC. Available electronically on the internet at:
http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=137307

U.S. Environmental Protection Agency (2005c). Total Risk Integrated Methodology TRIM.Expo/Inhalation User's Document Volume I: Air Pollutants Exposure Model (APEX, version 4) User's Guide. Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC. Will be available electronically on the internet at: http://www.epa.gov/ttn/fera/human_apex.html.

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

APPENDIX A. ANALYSIS OF AIR EXCHANGE RATE DATA

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 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 1. Summary of Studies of Residential Air Exchange Rates

Study	Avol	RTP Panel	RIOPA	TEACH	Wilson 1984	Wilson 1991	Murray and Burmaster
Locations Studied	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 Studied	1994	2000; 2001	1999; 2000; 2001	1999; 2000	1984, 1985	1984	1982 – 1987
Months/ Seasons Studied	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 Studied with available AER Measurements	86	37	284	85	581	288	1,884
Total AER Measurements	161	854	524	151	1,362	316	2,844
Average Number of AER Measurements per Home	1.87	23.08	1.85	1.78	2.34	1.10	1.51
AER 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
AER 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
Max Temperature (C)	36.25	30.81	32.50	32.00	28.00	25.00	Not available

Study	Avol	RTP Panel	RIOPA	TEACH	Wilson 1984	Wilson 1991	Murray and Burmaster
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 Made	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 Made	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, 10^{th}, 20^{th} … 90^{th} 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 2.

Table 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 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 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 4.

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

Study	Variable	N	Mean	Std Dev	Min	25^{th} %ile	Median	75^{th} %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 5.

Table 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 6.

Table 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 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 5^{th} 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 7.

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

	FRACTION VMT BY ROAD TYPE

FEDERAL-AID URBANIZED AREA	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 which 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 8.

Table 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 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 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 METS VALUES IN HUMAN EXPOSURE MODELING FOR FATIGUE AND EPOC

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 METS 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 METS paradigm, which uses ratios of activity-specific to basal energy expenditure. However, the basic or "raw" METS 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 METS is lowered if events in the recent past have been at unusually high METS 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 METS levels above the 'raw' values pulled from the activity-based distributions.

Historically, the logic for the downward adjustments (downward limitations on the maximum METS 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 METS 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 METS via the value of an oxygen deficit an individual has incurred. The method is more computationally efficient than previous METS-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 METS, 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 METS 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 METS levels corresponding to post-exercise diary events were based on maintaining an oxygen deficit below this maximum value. In addition, adjustments to METS 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 METS rather than total METS. The reserve is the amount over the basal rate (METS=1). Furthermore, we defined M as the normalized reserve, so that M=0 at METS=1, and M=1 at maximum METS:

$$M = \frac{METS - 1}{METS_{max} - 1} \tag{1}$$

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

Nomenclature

METS	Metabolic equivalent (unitless)
$METS_{max}$	Maximum achievable metabolic equivalent for an individual (unitless)
M	Normalized METS 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)

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:

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

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 $METS_{max}$, the METS value at which they exercised, and the corresponding normalized reserve METS (M). (Note that the METS and $METS_{max}$ 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 $METS_{max}$. 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 METS 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 the Appendix (A2). 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 METS:

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

where $METStoO_2$ is the conversion factor[2] for mlO_2 to METS-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 METS/min) to the maximum (3.7 METS/min) slope values, and so values were selected from a uniform distribution having these bounds. Converting units and normalizing to M, one obtains:

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

The data for all studies are given in the Appendix (A3).

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 METS 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:

$$EPOC_{fast} = 0.5\frac{(\Delta M)^2}{S_{fast} t_e} \quad (15)$$

$EPOC_{fast}$ will thus have units of M (normalized reserve METS). The M level for the post-exercise events will be incremented by $EPOC_{fast.}$

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 METS for the slow recovery process must be

$$EPOC_{slow} = \frac{D_{max}}{t_r} \quad (16)$$

Every diary event with the full rate of slow recovery will have its M value adjusted upward by $EPOC_{slow}$. 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

$$M_{adj} = M + EPOC_{fast} + EPOC_{slow} \quad (17)$$

and the new METS value for the event is

$$METS_{adj} = M_{adj}(METS_{max} - 1) + 1 \quad (18)$$

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 METS* (Kristin Isaacs and Graham Glen, February 5, 2005). Though that report utilized an earlier version of METS-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 METS (and thus PAI) a bit more than the earlier method. However, it is expected that this decrease will be fairly uniform across cohorts.

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

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

EPOC

The adjustments to single-event METS for EPOC when the algorithm was applied to CHAD were very small compared to the adjustments made for fatigue. In general, the increases in METS for EPOC were small, usually less than one MET. In a few cases the adjustments were bigger (on the order of 4-6 METS), due to the fact that the adjustments were applied to a very short event. An example of a METS 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 METS 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 METS 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 METS 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 (ml/kg)	Time to Reach Exhaustion (Hours)	METSMax	METS	M
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

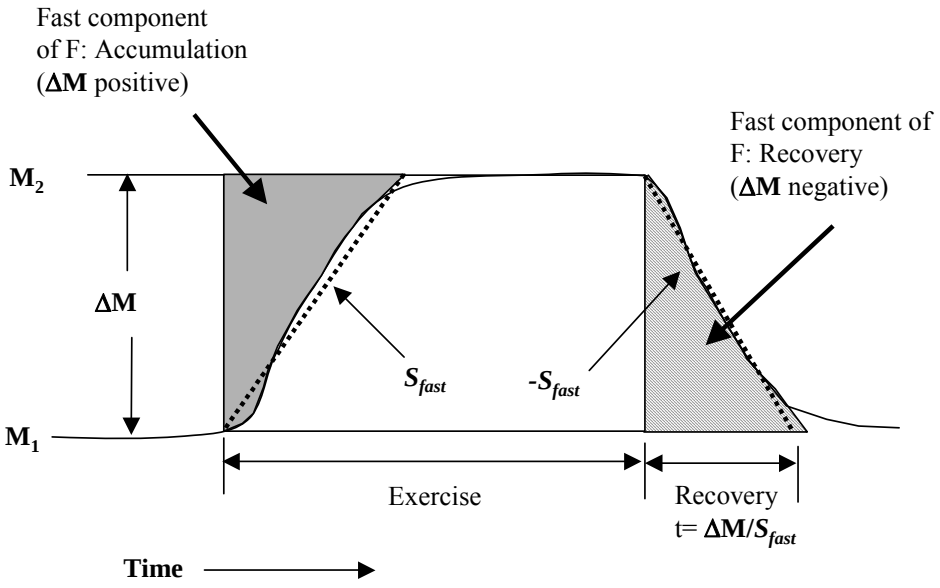


Figure 1. Fast components of oxygen deficit and recovery.

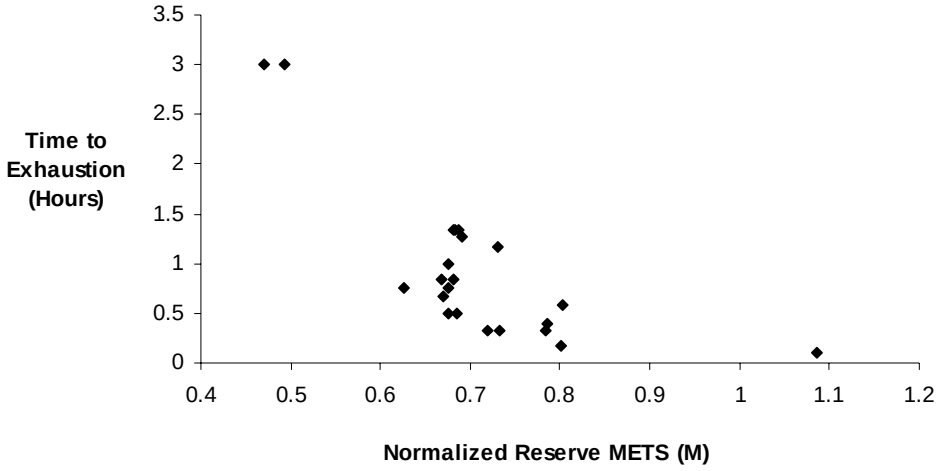


Figure 2. Exercise level in normalized reserve METS (M) versus time to reach exhaustion. Note nonlinear relationship.

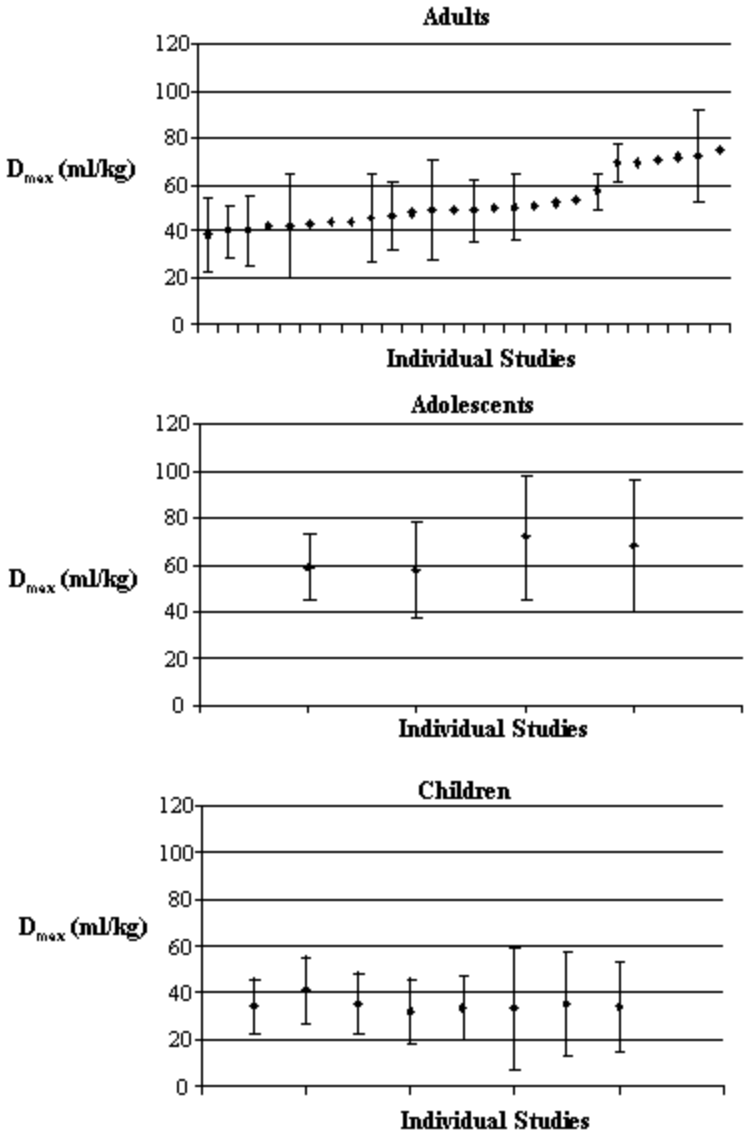


Figure 3. Values of the maximum accumulated oxygen deficit at exhaustion in adults[14-29], adolescents[30], and children[31-32]. The diamonds represent means from different exercise protocols. Bars are ± one standard deviation.

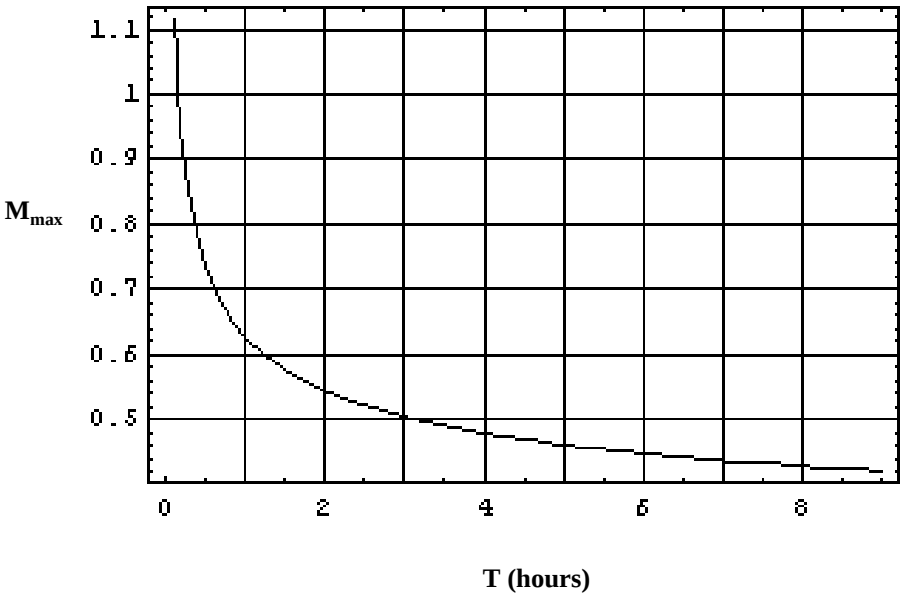


Figure 4. Maximum M (METS reserve) value that can be sustained during constant–intensity exercise of duration T.

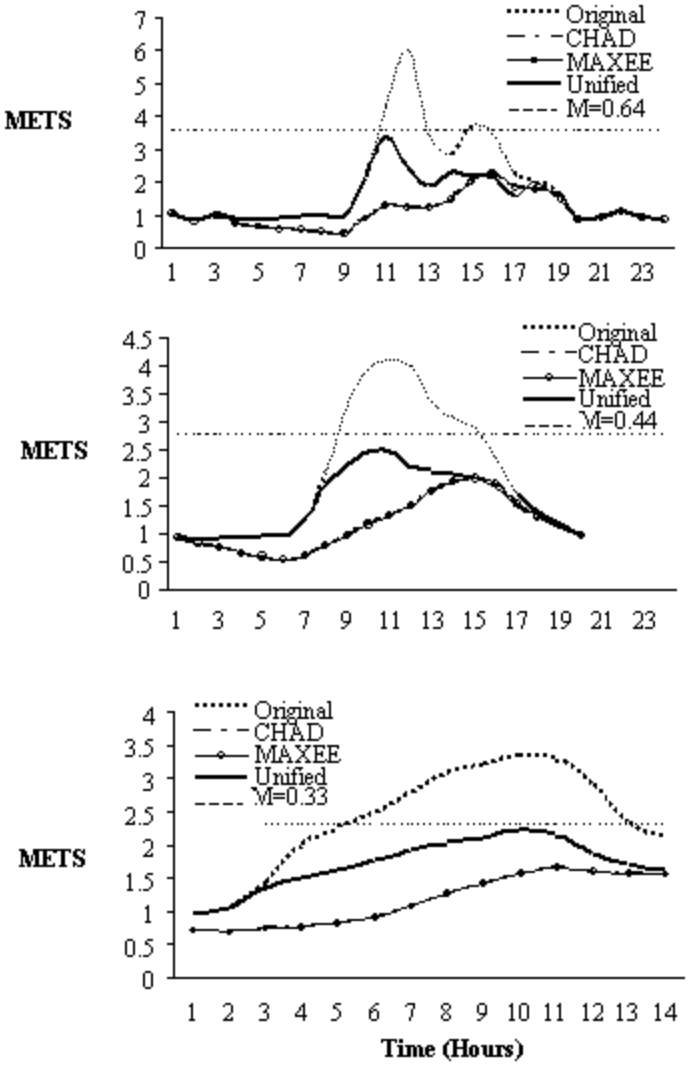


Figure 5. From the top, one, four, and nine hour running averages of METS for different METS-adjustment algorithms (for a male, age= 3 years, METSmax=5). Straight lines are the limits imposed by the Bink[43] equation.

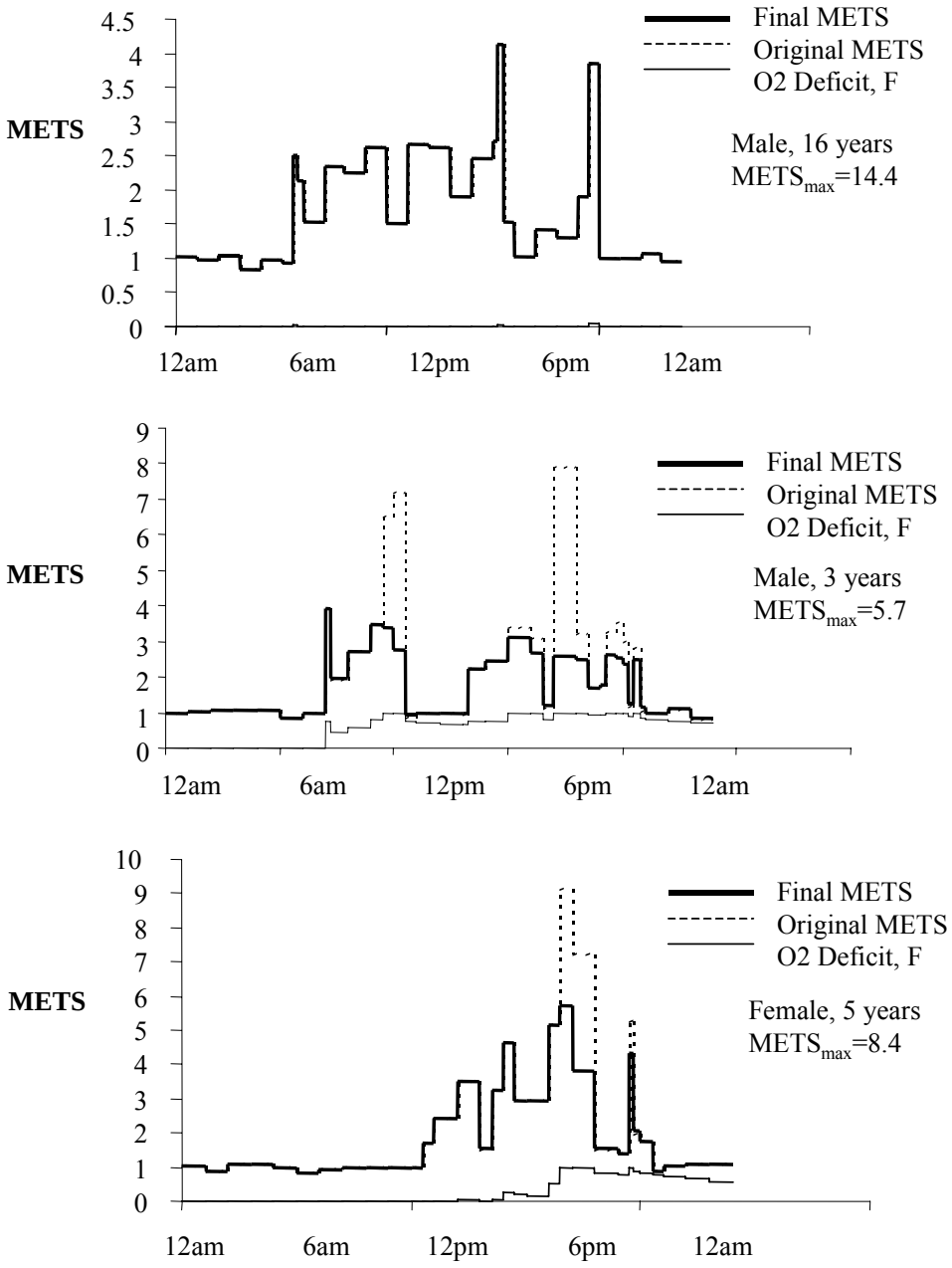


Figure 6. Three examples of original and corrected METS event time series. In the example in the top panel, the METS values pulled from the CHAD diary required no adjustment. The METS series for the individuals represented in the bottom two panels were adjusted for fatigue.

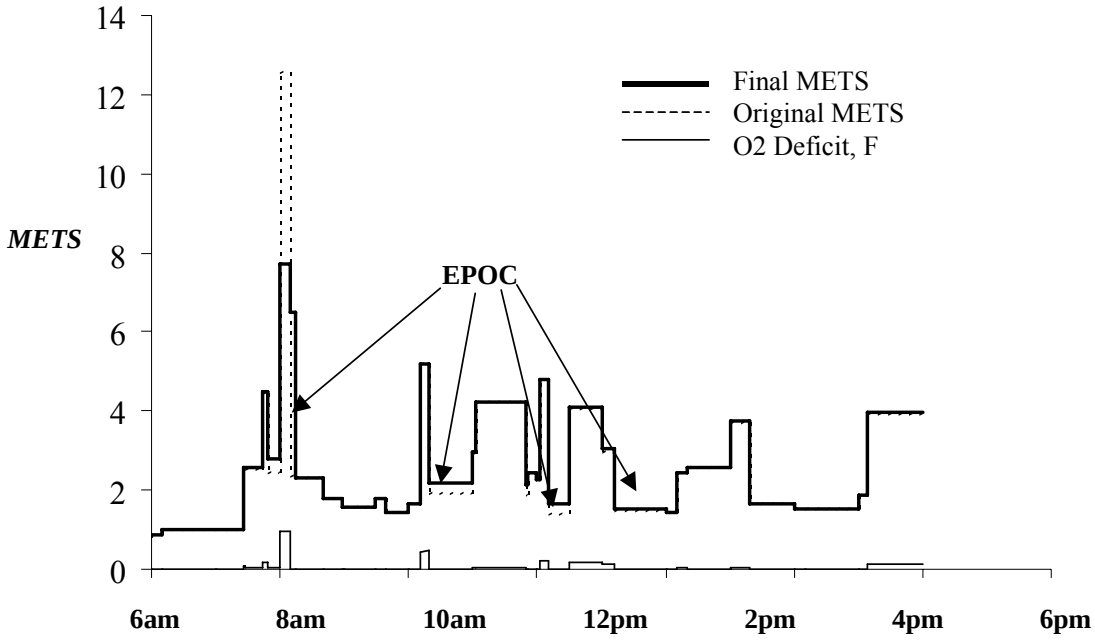


Figure 7. Event series for a 13-year-old female (METSmax=14), showing small upwards adjustments in METS for EPOC.

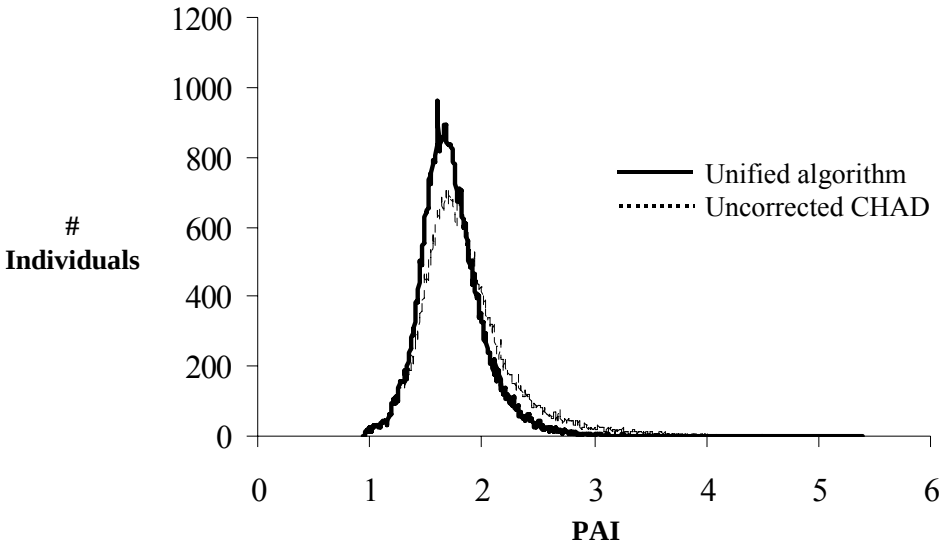


Figure 8. PAI distributions calculated using the uncorrected CHAD METS values compared with the values resulting from the unified algorithm.

Appendix.

A1. Algorithm FORTRAN Code

The APEX ventilation routine code is given below. The highlighting marks the statements associated with the unified algorithm.

```
SUBROUTINE Ventilation(P)                                          ! Called from main APEX
program
   INTEGER (KIND=IK), INTENT(IN) :: P                              ! Input variable (profile
number)
   INTEGER (KIND=IK), PARAMETER  :: MAXEVENTS = 60000              ! Max # of events per person
   INTEGER (KIND=IK) :: H, I, J, K                                 ! Local integer variables
   REAL    (KIND=RK) :: RecoveryT, OxdFrac, DeltaOxdFrac           ! Local real variables
   REAL    (KIND=RK) :: Over, MaxOver,  origmets                   ! Local real variables
   REAL    (KIND=RK) :: METSFactor, METSMax, PAI, PAIold           ! Various limits on METS
   REAL    (KIND=RK) :: LogTerm, LogVEBM, VO2Factor, METStoO2      ! More local real variables
   REAL    (KIND=RK) :: Z(MAXEVENTS), Random, Weightkg             ! More local real variables
   REAL    (KIND=RK) :: Sfast, Dmax, Duration, M, F, Mlast, Morig  ! More local real variables
   REAL    (KIND=RK) :: Flast, DelM, EPOCfast, EPOCslow, METS, Del2 ! More local real variables
   REAL    (KIND=RK) :: DeltaMETS, METSlast, DelFfast, DelFslow    ! More local real variables
   REAL    (KIND=RK) :: NewMax,A, B, ChangeDur, Rand(2), Fraction  ! More local real variables
   ! Global constants: IK, LogU, LV_MS2, RK, SL_FM2, ST_MS2, YES
   ! Global vars used: DebugLevel, DoDose, NumDays, NumEvents, NumHours, Phys
   ! Global vars set : EventSeq, HourSeq, ISdPrs, PersonDay, ProcLevel, ProcName
   ! Module vars set : PAIArray
   ! Intrinsic procs : EXP, LOG, MAX, MAXVAL, MIN, MOD, RANDOM_SEED, REAL, SUM, WRITE
   ! APEX procedures : RnNor
   ProcLevel = ProcLevel+1                                         ! Increment procedure depth
   ProcName(ProcLevel) = "Ventilation"                             ! Procedure name for
messages
   IF (DebugLevel>1) WRITE(LogU,SL_FM2) ST_MS2, ProcName(ProcLevel)  ! Message for debugging
   IF (P==1) THEN                                                  ! If start of simulation
      ALLOCATE(PAIArray(NumDays),STAT=AllocErr)                    ! Buffer for PAI values
      CALL CheckAllocation(AllocErr)                               ! CheckAllocation
   ENDIF                                                           ! End start simulation logic
   EventSeq(:)%VA  = 0.                                            ! Initialize VA value per
event
   EventSeq(:)%VE  = 0.                                            ! Initialize VE value per
event
   HourSeq(:)%VE   = 0.                                            ! Initialize VE value per
hour
   HourSeq(:)%VA   = 0.                                            ! Initialize VA value per
hour
   HourSeq(:)%EE   = 0.                                            ! Initialize EE value per
hour
   CALL RANDOM_SEED(Put=ISdPrs(1:KSeed))                           ! Initialize physiology seed
   IF (NumEvents>MAXEVENTS) Call Fatal(1)                          ! Array bounds of Z exceeded
   CALL RnNor(NumEvents,Z,-4.,4.)                                  ! Pick NumEvents normal
numbers
   CALL RANDOM_SEED(Get=ISdPrs(1:KSeed))                           ! Store the physiology
random seed
   CALL RnNor(2,Rand,-2.,2.)                                       ! Generate normals
   CALL RANDOM_NUMBER(Random)                                      ! Generate normals
   RecoveryT       = Phys(P)%RecTime                               ! Recovery time (hours)
   Weightkg        = Phys(P)%Weight*.4536                          ! Weight in Kg
   METSMax         = Phys(P)%MetsMax                               ! METS limit for profile P
   METSFactor      = Phys(P)%ECF*Phys(P)%RMR*19630.                ! METS to VA conversion
factor
   METStoO2        = 3.5*Weightkg*60                               ! METS to O2 factor
([mlO2/hr]/MET)
   IF (Person(P)%Age >  17) Dmax = 52.33+15.11*Rand(1)             ! Dmax for adults (ml/kg)
   IF (Person(P)%Age <= 17) Dmax = 63.95+31.05*Rand(1)             ! Dmax for adolescents
(ml/kg)
   IF (Person(P)%Age <  12) Dmax = 34.74+12.70*Rand(1)             ! Dmax for children (ml/kg)
   Dmax         = (Dmax*Weightkg)/(METStoO2*(METSMax-1))           ! Dmax (M-hr)
   Sfast        = (0.6+2.3*Random)*60                              ! Slope of fast recovery
(METS/hr)
```

```
   Sfast        = (Sfast)/(METSMax-1)                                  ! Slope of fast recovery
(M/hr)
   VO2Factor    = Phys(P)%ECF*Phys(P)%RMR/Phys(P)%BM                   ! VO2/BM at rest (l-
O2/min/kg)
   PersonDay%EndGn = Phys(P)%ENDGN1                                    ! Normal EndGn rate (all
days)
   K               = Phys(P)%Start                                     ! Phase of menstrual cycle
   IF (K>0) THEN                                                       ! If menstrual cycle exists
      DO J=1,NumDays                                                   ! Loop over days in
simulation
         IF (1+MOD(J+K,28)>14) PersonDay(J)%EndGn = Phys(P)%ENDGN2    ! Set 14 days per 28 to
ENDGN2
      ENDDO                                                            ! Continue with next day
   ENDIF                                                               ! End of ENDGN logic
   METSFactor      = Phys(P)%ECF*Phys(P)%RMR*19630.                    ! METS to VA conversion
factor
   CALL RnNor(NumEvents,Z,-4.,4.)                                      ! Pick NumEvents normal
numbers
   CALL RnNor(2,Rand,-2.,2.)                                           ! Generate normals
   CALL RANDOM_NUMBER(Random)                                          ! Generate normals
   CALL RANDOM_SEED(Get=ISdPrs(1:KSeed))                               ! Store the physiology
random seed
   Flast        = 0.                                                   ! Initialize O2 debt
   Mlast        = 0.                                                   ! Initialize prev METS value
   IF (Person(P)%Age >  17) Dmax = 54.95+14.46*Rand(1)                 ! Max O2 debt for adults
(ml/kg)
   IF (Person(P)%Age <= 17) Dmax = 63.95+21.12*Rand(1)                 ! Max debt for adolescents
(ml/kg)
   IF (Person(P)%Age <  12) Dmax = 34.74+13.10*Rand(1)                 ! Max O2 debt for children
(ml/kg)
   RecoveryT    = Phys(P)%RecTime                                      ! Recovery time (hours)
   MetsMax      = Phys(P)%MetsMax                                      ! METS limit for profile P
   A            = 5.20-(1.54/RecoveryT)+(3.92/RecoveryT**2)            ! O2 deficit regression
coeff A
   B            = 3.93-(3.57/RecoveryT)+(3.66/RecoveryT**2)            ! O2 deficit regression
coeff B
   Dmax         = Dmax/(3.5*60.*(METSMax-1.))                          ! Convert Dmax units to (M-
hr)
   Sfast        = (0.6+3.1*Random)*60./(METSMax-1.)                    ! Slope of fast recovery
(M/hr)
   DO I=1,NumEvents                                                    ! Loop over events
      Origmets  = EventSeq(I)%METS                                     ! Current unadjusted METS
      Mets      = MIN(MAX(Origmets,1.001),MetsMax)                     ! METS, bounded by 1 and
METSmax
      Duration  = EventSeq(I)%Duration/60.                             ! Duration in hours
      M         = (Mets-1.)/(MetsMax-1.)                               ! Unadjusted value for M
      DO WHILE (M>0)                                                   ! Loop until fatigue is
acceptable
         DelM      = M-Mlast                                           ! Change in M from prior
event
         DelFfast  = 0.5*((DelM)*(abs(DelM)))/(Sfast*Dmax)             ! DeltaF fast
         Del2      = ABS(DelM)-Sfast*Duration                          ! Determine if Fast is
truncated
         IF (Del2 > 0.) DelFfast = DelFfast*(1.-Del2**2/DelM**2)     ! Cut DelF(fast) if
truncated
         DelFslow  = (A*(M**B)-1./RecoveryT)*Duration                 ! Calculate DelF(slow)
         IF (Flast+DelFfast+DelFslow < 0.) THEN                       ! If more than full recovery
            fraction = Flast/ABS(DelFfast+DelFslow)                   ! Determine excess recovery
            DelFFast = DelFfast*fraction                              ! Limit fast recovery
            DelFSlow = DelFslow*fraction                              ! Limit slow recovery
         ENDIF                                                        ! End of excess recovery
correction
         EventSeq(I)%Deficit = Flast + DelFfast + DelFslow            ! New deficit (F) value
         IF (EventSeq(I)%Deficit < 1.) EXIT                           ! If F value is ok, exit
         M = M - 0.01                                                 ! Adjust M downwards and try
again
      ENDDO                                                           ! End fatigue adjustment
logic
      EPOCfast = 0.                                                   ! Default is no fast EPOC
      IF (DelFfast<0.) EPOCfast = ABS(DelFfast)*Dmax/duration         ! Determine fast EPOC
      EPOCslow = 0.                                                   ! Default is no slow EPOC
```

```
      IF (DelFslow<0.) EPOCslow = ABS(DelFslow)*Dmax/duration        ! Determine slow EPOC
      M = M + EPOCfast + EPOCslow                                    ! Adjust M for EPOC
      DeltaMets = 1.+M*(MetsMax-1.)-Mets                             ! Change in absolute METS
      IF (OrigMets<1.)  Mets = OrigMets                              ! Return METS<1 to orig vals
      EventSeq(I)%Mets = Mets+DeltaMets                              ! Update event METS w/EPOC
      if (EventSeq(I)%Deficit<0) EventSeq(I)%Deficit=0               ! Fix any roundoff error in
F
      Flast = EventSeq(I)%Deficit                                    ! Update last F
      Mlast = M                                                      ! Update last M
   ENDDO                                                             ! End loop over events
      DO I=1,NumEvents                                               ! Loop over diary events
      EventSeq(I)%VA = EventSeq(I)%METS * METSFactor                 ! VA is proportional to METS
      LogTerm        = LOG(EventSeq(I)%METS*VO2Factor)               ! Log(VO2/BM) term
      IF (P==37118) Write(LogU,'(I6)') I                             ! jel
      LogVEBM        = Phys(P)%VEinter + Phys(P)%VEslope*LogTerm    &
                        +Phys(P)%VEresid*Z(I)                        ! Regression for Log(VE/BM)
      EventSeq(I)%VE = Phys(P)%BM * EXP(LogVEBM) * 1000.             ! Solve for VE in (ml/min)
   ENDDO                                                             ! Continue with next diary
event
   DO I=1,NumHours                                                   ! Loop for update hour
variables
      J = HourSeq(I)%FirstEvent                                      ! First event for given hour
      K = HourSeq(I)%LastEvent                                       ! Last event for given hour
      HourSeq(I)%Mets =                                             &
         SUM(EventSeq(J:K)%Mets*EventSeq(J:K)%Duration)/60.          ! Mean METS value for given
hour
      HourSeq(I)%EE = Phys(P)%RMR*HourSeq(I)%Mets                    ! Energy expenditure kcal/hr
      HourSeq(I)%VE =                                               &
         SUM(EventSeq(J:K)%VE*EventSeq(J:K)%Duration)/60.            ! Mean VE value for given
hour
      HourSeq(I)%VA =                                               &
         SUM(EventSeq(J:K)%VA*EventSeq(J:K)%Duration)/60.            ! Mean VA value for given
hour
      HourSeq(I)%EVR = HourSeq(I)%VE/(1000*(Phys(P)%BSA))            ! Mean EVR for hour (l/min-
m^2)
      HourSeq(I)%PAI =                                              &
         SUM(EventSeq(J:K)%Mets*EventSeq(J:K)%Duration)/60.          ! Mean PAI value for given
hour
      IF (I<=8) THEN                                                 ! First 8 hours use special
logic
         HourSeq(I)%Run8EVR  = SUM(HourSeq(1:I)%EVR)/I               ! Running 8-hour average for
EVR
      ELSE                                                           ! If beyond first 8 hours
         HourSeq(I)%Run8EVR  = HourSeq(I-1)%EVR  +                  &
            (HourSeq(I)%EVR-HourSeq(I-8)%EVR)/8.                     ! Add new hour and drop off
old
      ENDIF
   ENDDO                                                             ! End hour sequence update
   DO I=1,NumDays                                                    ! Loop over days in
simulation
      PersonDay(I)%PAI = SUM(HourSeq(24*I-23:24*I)%PAI)/24           ! Daily average PAI
   ENDDO                                                             ! End day loop
   PAIArray=REAL(PersonDay%PAI,LONG)                                 ! Fill the temp PAI array
   CALL Percentiles(PAIArray,                                       &
      NumDays,50.0,1,Person(P)%PAI)                                  ! Get median PAI over days
   IF (DebugLevel>1) WRITE(LogU,SL_FM2) LV_MS2, ProcName(ProcLevel)  ! Message for debugging
   ProcLevel = ProcLevel-1                                           ! Decrement procedure depth
END SUBROUTINE Ventilation                                           ! Return to main APEX
program
```

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

Abbreviations

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

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

A3. Data for Calculation of the Slope of the Fast EPOC Component

Study	Peak VO2 (ml/min)	Baseline (ml/min)	VO2 post-EPOCfast ml/min	Duration of EPOCfast min	Slope ml/min/min	Slope 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

APPENDIX C. A NEW METHOD OF LONGITUDINAL DIARY ASSEMBLY

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.

<u>The statistics D and A</u>

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

$$\sigma_w^2 = (1/P) \sum_{i=1}^{P} \sigma_i^2 = 1/(JP) \sum_{i=1}^{P} \sum_{j=1}^{J} (t_{ij} - \mu_i)^2 = V^2 - (1/P) \sum_{i=1}^{P} \mu_i^2 \quad (1\text{-}6)$$

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

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

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

$$\sigma_w^2 + \sigma_b^2 = V^2 - \mu^2 = \sigma^2 \quad (1\text{-}8)$$

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

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

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

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

So this statistic reflects the distribution of the personal means μ_i. It does not reflect any patterns or ordering of the t_{ij} within a diary. Note that it is possible to interchange two or more days (interchange t_{ij} and t_{ik} 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_{ij} 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_{ij} are the X-scores. There are two constraints to be met. The set of selected values t_{ij} should produce a sample D statistic close to the requested value, and the set of t_{ij} (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

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

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.

$$\sigma_w^2 = \int p(T_i)\,s_i^2\,d\,T_i$$

$$= ((J-1)/J) \int p(T_i)\,T_i\,(1 - T_i\,)\,/\,(1 + S_i\,)\,d\,T_i\,. \qquad (2\text{-}11)$$

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} moment of μ_i 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} moment of μ_i about 1/2 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

$$a = S\,T_i = \frac{2\,T_i}{(1 - 3\,w^2 / (1 + 2\,\alpha))} . \quad (2\text{-}23)$$

Therefore

$$b = S\text{-}a = \frac{2\,(1 - T_i)}{(1 - 3\,w^2 / (1 + 2\,\alpha))} . \quad (2\text{-}24)$$

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

$$D^{\#} = 3\,w^2 / (1 + 2\,\alpha) \quad (2\text{-}25)$$

and rewrite equation (2-17) in terms of this new parameter:

$$\sigma_b^2 = [1/(J + J\,(2/(1-D^{\#})))]\,[1/4 - D^{\#}/12] + D^{\#}/12. \quad (2\text{-}26)$$

which can be solved for $D^{\#}$ in terms of σ_b^2

$$D^{\#} = (12\,J\,\sigma_b^2 - 1) / (J - 1) . \quad (2\text{-}27)$$

Also, equation (1-9) together with (2-20) give

$$D = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2) = 12\,\sigma_b^2 . \quad (2\text{-}28)$$

Hence

$$D^{\#} = (J\,D - 1) / (J - 1) \quad (2\text{-}29)$$

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 $D^{\#}$

$$R = (D^{\#})^{1/2} . \quad (2\text{-}30)$$

It is found empirically that the following relationship between α and R gives a nearly uniform distribution of X-scores:

$$\alpha = 1 - (4/5)\,[4\,R\,(1 - R)]^3 \quad (2\text{-}31)$$

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

$$\alpha = 1 \qquad (2\text{-}32)$$

whereupon equation (2-25) reduces to

$$w^2 = D^{\#} . \qquad (2\text{-}33)$$

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:

$$R_j = \text{ceil}(K\,x), \quad \text{and} \quad R_{j+1} = \text{ceil}(K\,y) \tag{3-3}$$

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:

$$E\,[(R_j - \rho)\,(R_{j+1} - \rho)\,] = \Sigma\ \Sigma\ (R_j - \rho)\,(R_{j+1} - \rho)\,p(R_j)\,p(R_{j+1} \mid R_j) \tag{3-5}$$

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 $\int p(y|x)\,dy = 1$. Also note that for the integers from 1 to K, the mean is $\rho = (K+1)/2$, so $(1/2-\rho) = -K/2$. 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

$$b = S-a = S/2 + S\,w/2 - S\,w\,x. \tag{3-9}$$

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

$$E\left[(R_j - \rho)\,(R_{j+1} - \rho)\right] = \Sigma\,(R_j - K/2 - 1/2)\,w\,(x-1/2). \qquad (3\text{-}11)$$

Replacing x by $(R_j - 1/2)/K$ and noting that the sums evaluate to

$$\Sigma\,R_j^2 = K\,(K+1)\,(2K+1)/6, \quad \Sigma\,R_j = K\,(K+1)/2, \quad \Sigma\,1 = K, \qquad (3\text{-}12)$$

then equation (3-11) can be expanded and evaluated to give

$$E\left[(R_j - \rho)\,(R_{j+1} - \rho)\right] = w\,(K^2 - 1)\,/12. \qquad (3\text{-}13)$$

With this choice of the beta distribution, equation (3-2) reduces to the very simple form

$$w = a_i. \qquad (3\text{-}14)$$

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 k^{th} moment about zero of a uniform distribution from zero to one is

$$m_k = \int x^k\,p(x)\,dx = 1/(k+1) \qquad (3\text{-}15)$$

since p(x) =1 for a uniform. The moments of the 'y' values are

$$E(y^k) = \int\int y^k\,p(y|x)\,p(x)\,dy\,dx$$

$$= \int p(x)\,dx \int y^k\,p(y|x)\,dy\,. \qquad (3\text{-}16)$$

The second integral is the k^{th} 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

$$M_2 = a\,(a+1)\,/\,(S\,(S+1))$$

$$= (S/2 - S\,w/2 + S\,w\,x)\,(1+S/2 - S\,w/2 + S\,w\,x)\,/\,[S\,(S+1)]$$

$$= [\,(1-w)(2+S-S\,w)/4 + (w + S\,w - s\,w^2)\,x + S\,w^2\,x^2\,]\,/\,(S+1). \qquad (3\text{-}17)$$

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

$$E(y^2) = [(1-w)(2+S-S\,w)/4 + (w+S\,w - S\,w^2)/2 + S\,w^2/3\,] / (S+1)$$

$$= (6 + 3\,S + S\,w^2) / (12 + 12\,S) \qquad (3\text{-}19)$$

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

$$S = 2 / (1- w^2) \qquad (3\text{-}20)$$

Matching the third moments $m_3 = E(y^3)$ results in the same relationship $S = 2 / (1- w^2)$. Moments higher than this generally will not match.

To summarize, the parameter values should be

$$w = a_i$$

$$S = 2 / (1- w^2). \qquad (3\text{-}21)$$

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

$$R_{j+1} = \text{ceil}(K\, y) \qquad (4\text{-}1)$$

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

$$d_{j+1} = \text{ceil}(D_{j+1}\ x_{j+1}). \qquad (4\text{-}2)$$

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.