The HAPEM6 User's Guide Hazardous Air Pollutant Exposure Model, Version 6

January 2007

Prepared for:
Ted Palma
Office of Air Quality Planning and Standards
US Environmental Protection Agency
Research Triangle Park, North Carolina

Prepared by:
Arlene Rosenbaum
ICF International
4464 Hillview Way
Rohnert Park, CA 94111
(707) 586-2822

and

Michael Huang
ICF International
9300 Lee Highway
Fairfax, VA 22031
(703) 934-3982

This page intentionally left blank.

Contents

Figures

Tables

New Features in HAPEM6

The Hazardous Air Pollutant Exposure Model, version 6 (HAPEM6) includes a number of new features. These new features are designed to provide exposure estimates that better characterize the variability across the population. These new features are summarized here, and detailed in other portions of this User's Guide.

- Spatial variability for ambient concentrations within a tract can now be characterized for onroad mobile sources by accounting for concentration enhancements near major roadways. This is done by applying enhancement factors to outdoor concentrations in the vicinity of all indoor microenvironments for tract-specific fractions of tract residents.
- The fraction of the tract populations that commutes to work, and the mode of commuting (public or private transit) is now determined based on US Census data.
- Commuting time for each simulated individual who commutes to work is calculated according to the distance between the given home tract and the selected work tract, and the average commuting time for workers in the home tract based on US Census data.
- The number of demographic groups in the default files has been streamlined to 6 age groups (no gender stratification).
- The number of microenvironments in the default files has been streamlined to 14.

1. Introduction

The Hazardous Air Pollutant Exposure Model, version 6 (HAPEM6) User's Guide is designed to assist exposure analysts with running and interpreting results from HAPEM6. Throughout the User's Guide, the input file names and file types are in lowercase italics and program names are in all uppercase letters for easier identification. Likewise, model variables are presented in bold italics. When presented, input and output data and program source codes will be presented in a single lined box, indicating that the text inside the box is shown exactly as it exists in its electronic form. In addition, shaded text boxes appear throughout the document providing useful information and tips to users.

1.1. Organization of the User's Guide

The User's Guide is organized into six chapters and an appendix. Chapters 1 and 2 provide a general overview of the background functionality of HAPEM6, as well as basic instructions for running the model. The remaining chapters are designed to provide the user with more detailed information on the components of HAPEM6. These chapters are designed to be easily referenced without requiring the entire document to be read. We suggest, however, that the novice user read all of the chapters at least once to gain a better understanding of HAPEM6.

Chapter 1 Provides a brief introduction to HAPEM6 modeling fundamentals including a brief history of the development of HAPEM6.

Chapter 2 Provides an overview of the various components of HAPEM6 and basic information needed to run the model.

Chapter 3 Provides a description of the format, data, and options for each of HAPEM6 input files.

Chapter 4 Provides a description of the format and data associated with each of HAPEM6 output files.

Chapter 5 Provides a description of the purpose, operations, inputs, and outputs, including a brief description of the computer code, for each of HAPEM6 computer programs.

Chapter 6 References.

1.2. Background

The Hazardous Air Pollutant Exposure Model, version 6 (HAPEM6) is a screening-level exposure model appropriate for assessing average long-term inhalation exposures of the general population, or a specific sub-population, over spatial scales ranging from urban[1] to national. HAPEM6 provides a relatively transparent set of exposure assumptions and approximations, as is appropriate for a screening level model.

HAPEM6 uses the general approach of tracking representatives of specified demographic groups as they move among indoor and outdoor microenvironments and among geographic locations. The estimated pollutant concentrations in each microenvironment visited are combined into a time-weighted average concentration, which is assigned to members of the demographic group.

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

HAPEM6 uses four primary sources of information: population data from the US Census, population activity data, air quality data, and microenvironmental data. These data will be discussed briefly below, and in greater detail later in this User's Guide.

1.2.1. Population Data

The U.S. Census Bureau is the primary source of most population demographic data. The U.S. Census Bureau collects information on where people live, their demographic makeup (e.g., age, gender, ethnic group), and employment. The default population data for HAPEM6 are derived from 2000 US Census data reported at the spatial resolution of census tracts, which are small, relatively permanent statistical subdivisions of a county. Census tracts usually contain between 2,500 and 8,000 residents.

A second type population data used in HAPEM6 is an estimate of the fraction of the population of each Census tract that lives within certain distances of major roadways. These estimates were derived from the Environmental Sciences Research Center (ESRI) StreetMap US roadway geographic database and a geographic database of US Census block boundaries. (See Appendix B for more details.) They are used, in conjunction with the *PROX* factors described below, to implement a new feature in HAPEM6, which accounts for the enhanced outdoor concentrations of pollutant emitted from onroad vehicles at locations near major roadways, and the associated enhanced indoor concentrations.

1.2.2. Activity Data

HAPEM6 uses four types of population activity data: activity pattern data, commuting spatial pattern data, commuting time data, and commuting fraction data. Human activity pattern data are used to determine the frequency and duration of exposure for specific groups within various microenvironments. Activity pattern data are taken from demographic surveys of individuals'

[1] Urban refers to a scale that encompasses the size of a large city, and is generally on the order of tens of kilometers.

daily activities, the amount of time spent engaged in those activities, and the locations where the activities occur.

In addition to recording the duration and location of a person's activities, these surveys also collect important demographic information about the person. The demographic information usually includes the person's age, gender, and ethnic group. Most activity pattern studies also try to collect information on other attributes of a respondent, such as highest level of education completed, number of people in their household, whether the person or anyone in their household is a smoker, employment status, and the number of hours spent outdoors.

The default population activity file for HAPEM6 is derived from a database of activity pattern surveys called the Consolidated Human Activity Database (CHAD) (Glen et al. 1997). The CHAD is currently comprised of over 22,000 person-days of activity pattern data, including 140 activities and 114 locations, collected and organized from twelve human activity pattern surveys. The CHAD contains the sequential patterns of activities for each individual, and each activity has a corresponding location code so that the microenvironment of each activity is known. The microenvironment categories currently incorporated into the default population activity file for HAPEM6 are presented in Table 1-1.

Table 1-1. HAPEM6 microenvironments

MICRO – ENVIRONMENT No.	MICROENVIRONMENT	
	SPECIFIC	GENERAL
1	Residential	Indoors
2	Residential Garage	Indoors
3	School	Indoors
4	Hospital	Indoors
5	Office	Indoors
6	Public Access	Indoors
7	Bar/Restaurant	Indoors
8	Car/Truck	In Vehicle
9	Public Transit	In Vehicle
10	Air Travel	In Vehicle
11	Outdoors, Near Roadway	Outdoors
12	Outdoors, Service Station	Outdoors
13	Outdoors, Parking Garage	Outdoors
14	Outdoors, Other	Outdoors

Because available activity data are not adequate to estimate the exposure of each individual in a population, HAPEM6 groups activity patterns data together for people with similar demographic characteristics that are expected to influence exposure to air pollutants (e.g., age, commuting status), and makes exposure estimates for these groups. The activity profiles for each person in a demographic group have an equal chance of being selected from the activity database. (See discussion of stochastic elements below.) The result is that HAPEM6 provides a distribution of exposure concentrations for each demographic group in each census tract.

HAPEM6 divides the population into 6 demographic groups, based on age category. Activity pattern data are also separated into 3 day types (summer weekdays, other weekdays, and weekends), and commuting status (yes or no).

The commuting spatial pattern data contained in the HAPEM6 default file were derived by US EPA Office of Research and Development from the 2000 Census, data collected as part of the Census Transportation Planning Package (CTPP) and distributed by the US DOT Bureau of Transportation Statistics at their web site http://transtats.bts.gov/. The data files specify the number of residents of each tract that work in that tract and every other US Census tract (i.e., the population associated with each home tract/work tract pair) and the distance between the centroids of the two tracts. HAPEM6 uses this data in coordination with the activity pattern data to place an individual who commutes to work either in the home tract or the work tract at each time step.

The average commute time, stratified by public or private transit, for workers in each Census tract, as contained in the HAPEM6 default file, is derived from the 2000 Census (P32). These data are combined with data on the centroid-to-centroid distances between tracts, as described in Section 5.2.5 (HAPEM Processing), to estimate the commute time for each commuting replicate.

Data specifying the fraction of each demographic group in each Census tract that commute to work, as contained in the HAPEM6 default file, were derived from the 2000 Census (P31 and PCT35).

1.2.3. Air Quality Data

Some previous versions of HAPEM relied on measured outdoor air pollutant concentration data for the exposure calculations. This limited both the extent of the modeling domain and pollutants, because exposures could only be calculated for locations and pollutants with large monitoring networks. Typically, sufficient data were only available for large metropolitan areas and for the criteria pollutants[2].

HAPEM6 is able to estimate exposures over the entire U.S. at spatial scales as small as a US Census tract. In order to preserve any characteristic diurnal patterns in ambient concentrations that might be important in the estimation of population exposure, HAPEM6 can treat annual average concentration estimates that are stratified by time of day. For example, annual average concentrations may be stratified into (24) 1-hour time blocks, or (8) 3-hours. The stratified air quality data are then combined in HAPEM6 with similarly stratified activity data to estimate exposure concentrations. The default activity data are stratified into (8) 3-hour time blocks.

[2] Criteria pollutants are those for which a National Ambient Air Quality Standard (NAAQS) has been set. They are ground-level ozone, carbon monoxide, sulfur dioxide, nitrogen dioxide, lead, and particulate matter.

The air quality data can also be decomposed to reflect the contributions from various emission sources. The number of sources is a user-specified variable.

HAPEM6 is also able to incorporate spatial variability of air quality within each Census tract. That is, the air quality within a tract is not limited to a single point estimate (diurnally- and source-stratified). Spatial variability may be incorporated in two different ways. One method is to characterize the air quality in a Census tract by a set of up to 500 diurnally- and source-stratified values. How HAPEM6 handles this data set is explained below in Section 1.2.5 (Stochastic Elements).

When air quality is characterized by a single point estimate (diurnally- and source-stratified), a second method allows the user to specify a scalar factor to be applied to the tract air quality values, with the scalar dependant on the distance of the replicate's residence from a major roadway. This approach is discussed in Section 1.2.5 (Stochastic Elements).

1.2.4. Microenvironmental Data

In order to calculate the exposure concentration for each demographic group, an estimate is required of the concentration in each microenvironment (ME) specified by the activity pattern. In HAPEM6 these ME concentration estimates are derived from the outdoor concentration estimate for the geographic subdivision (e.g., US Census tract) and a set of 3 ME factors: *PEN*, *PROX*, and *ADD.* These account for penetration of outdoor air into the microenvironment, concentration enhancement due to proximity of the microenvironment to the emission source, and emission sources within the microenvironment.

The ME factors are entered into the model as data from input files that contain estimates of distributions for *PEN*, *PROX, and ADD* for three categories of pollutants: gases, particles, and semi-volatiles. The *PEN* distributions were obtained from an extensive review of literature and databases on indoor/outdoor ratios of hazardous air pollutants (HAPs). The *PROX* distributions for onroad mobile sources were derived from modeling studies of the concentration gradients of HAPs near major roadways[3]. How the distributions are utilized in HAPEM6 is discussed below in the section on stochastic elements.

As is the case with all other HAPEM6 input files, these data can be modified by the user. The ME factors should be updated as needed to reflect current knowledge, as available.

1.2.5. Stochastic Elements

Although it would be difficult to accurately represent the activities of an individual due to day-to-day variation, the general behavior of population groups can be well represented using stochastic processes. This makes it possible for estimates of population exposure to be characterized as distributions rather than point estimates. HAPEM6 incorporates six stochastic elements.

[3] The default *PROX* values for other emission source types are point values of 1.0 (i.e., no concentration enhancement due to proximity), and the default *ADD* values are point values of 0.0 (i.e., no indoor emission sources).

Commuting Status

The first stochastic element in the construction of a replicate is the determination of the commuting status (yes or no), according to the tract- and demographic-group-specific probabilities.

Activity Patterns

The second stochastic element is the selection of daily activity patterns to represent the demographic-group and commuting-status of the replicate. HAPEM6 estimates long-term average concentrations, but the available population activity data sequences are specified for 24-hour periods only. HAPEM6 contains a new approach for constructing long-term average activity sequences from short-term records. (See Appendix A for a detailed discussion, which is briefly summarized here.[4])

The general approach used by HAPEM6 is comprised of several steps. The first is to select three sets of 24-hour activity patterns, where each set is used to construct an average pattern for an individual for one of 3 specified day types: weekends, Summer weekdays, non-Summer weekdays. A set of patterns, rather than a single pattern, is selected for each day type to reflect the day-to-day variability of activity patterns for an individual. How the set of patterns is combined into an average pattern for the day-type is explained in the Implementation section below.

Next, the corresponding exposure concentration is calculated for each of the three day-type average activity patterns. Then a weighted average of the three exposure concentrations is calculated to represent the annual average concentration, where the weightings represent the number of days per year for each day type (i.e., 104 for weekends, 65 for Summer weekdays, 196 for non-Summer weekdays). This process is repeated for several replicates[5] for each census tract/demographic group combination, to create a set of annual exposure concentration estimates for each demographic group in each census tract.

Implementation

To implement this approach, first all the activity pattern data are grouped according to demographic-group, day type, and commuting status. Then for each demographic-group/commuting-status/day-type combination the activity patterns are stratified into from one to three categories, based on similarity of time spent in the various microenvironments, as determined by cluster analysis.

Transition probabilities between categories are derived from empirical data of sequenced diary records. Given that the first day of a 2-day sequence falls into category X, the transition probabilities specify the relative frequency of the second day falling into each possible category. For example, if half of the 2-day sequences with the first day in category X also have the second day in category X, the X-to-X transition probability would be 0.5.

The HAPEM6 algorithms construct an average activity pattern for each replicate by randomly selecting one activity pattern from each category and combining them with weighted averaging.

[4] Note that the results presented at the end of Appendix A were used for an earlier version of HAPEM. The default cluster and ClusTrans files for HAPEM6 were developed for different subsets of activity patterns.

[5] The number of replicates is a user-specified variable.

The weights represent the relative frequency of days from each category for the individual represented.

To determine the averaging weights to use, the algorithms perform a Markov process based on the category-to-category transition probabilities. For example, suppose the day type is summer weekday. Because there are 65 summer weekdays in a year, 65 random selections are made of categories. The category for the first day is selected randomly from the set of categories using the relative frequency of each category as the probability of selection. The category for the second day is selected according to the transition probabilities from the first day's category. The category for the third day is selected according to the transition probabilities from the second day's category. This is repeated until 65 category selections are made. The weight given each activity pattern in the averaging process is the number of times its category was selected in the Markov process.

Work Tract

Another stochastic process is applied in HAPEM6 for replicates that commute to work. For those groups a work tract is selected at random from the set of work tract specified for that home tract, using the proportion of workers commuting to each work tract for its selection probability.

Microenvironment Factors

Another stochastic feature of HAPEM6 is the ability to characterize microenvironment factors as variable, instead of uniform over the population. That is, three of the four microenvironment factors (*PEN*, *PROX*, and *ADD*) are represented by probability distributions rather than point estimates[6]. Several distribution types may be used, as discussed in Section 3.10. For each replicate a different set of microenvironment factors is randomly selected.

Air Quality - General

HAPEM6 has the ability to characterize outdoor air quality concentrations as spatially variable within a census tract. It can do this in two different ways. One approach is to characterize the air quality for each tract as a data set with up to 500 sets of value (i.e., diurnally- and source-stratified). Then for each replicate, a different set of ambient air quality concentrations is selected for the home (and work) tract to reflect the spatial variability in air quality within the tract.

Air Quality – Onroad Vehicle Related

When air quality is characterized by a single point estimate (diurnally- and source-stratified), another approach is used to account for enhanced onroad-vehicle-related HAP concentrations in the vicinity of major roadways. To implement this approach, the distance of the replicate's home (and workplace) from a major roadway is randomly selected based on tract- and demographic-group-specific probabilities. A *PROX* factor is then selected from a distribution and

[6] As noted above in practice the default *PROX* values for emission source types other than onroad vehicles are point values of 1.0 (i.e., no concentration enhancement due to proximity), and the default *ADD* values are point values of 0.0 (i.e., no indoor emission sources). However, HAPEM6 contains the structure to characterize these as distributions if appropriate data are available.

applied to the tract air quality values for onroad mobile sources, with the distribution dependent on the selected distance.

1.3. Strengths and Limitations of HAPEM6

All models have strengths and limitations. Therefore, for each application, it is important to carefully select the model that has the desired attributes. With this in mind, it is equally important to understand the strengths and weaknesses of the chosen model. The following sections provide a summary of the strengths and potential limitations of HAPEM6. However, this is not an exhaustive list and may not address features important for specific applications of an exposure model.

1.3.1. Strengths

HAPEM has undergone many enhancements in recent years. One is the ability to use air quality concentration estimates from the ASPEN modeling system. This capability allows exposure to population groups to be simulated at the census tract level, a much finer spatial resolution than was previously possible. It also means that estimation of population exposure no longer needs to rely solely on data from the limited (in both areal extent and pollutants measured) nationwide network of fixed-site monitors.

Another important feature of HAPEM6 is its versatility. The model is designed so that input data specific to different applications can be used without having to rewrite the computer source code. This flexibility is possible because most specifications are not “hard wired” into the model’s code. Instead, the necessary input data are entered through external databases and the modeling parameters are specified through an external file. This feature allows easier use of new data, or other information (e.g., microenvironmental factors) used by the model, as they become available.

Another strength of HAPEM6 is its ability to estimate the exposures of workers in the geographic area where they work, in addition to the geographic area where they live, since the pollutant concentrations in these locations may be very different.

Another important feature of HAPEM6 is the incorporation of stochastic processes for the selection activity patterns, work tracts, ambient air quality among locations within a tract, ME factors, so that more of the variability in the exposure estimates can be captured than simply the variability associated with residential tract.

Exposure assessment with HAPEM6 has also been facilitated by development of default input files derived from the databases discussed above: national US Census population and commuting information, CHAD activity data, and variable ME factors for gases, particles and semi-volatiles.

1.3.2. Limitations

HAPEM6 calculates long-term average exposure concentrations in order to address exposures to pollutants with carcinogenic and other long-term effects. Thus, HAPEM6 does not preserve the time-sequence of exposure events when sampling from the time/activity databases. The result is that information used to evaluate possible correlations in exposures to different pollutants due to activities that are related in time is not preserved.

HAPEM6 only estimates exposures experienced through inhalation. For certain HAPs, inhalation might not be the major route of exposure, and therefore, HAPEM6 may underestimate exposures in these instances. Also, although HAPEM6 is an inhalation exposure model, it does not include any measures of the ventilation rate associated with an activity, so there is no ability to calculate the potential dose received when engaging in various activities.

Uncertainty in the prediction distributions is not addressed. Some of the uncertainties are as follows.

- The population activity pattern data are limited. Only one of the 12 studies in CHAD was national in scope; therefore, the combined data set does not constitute a representative sample, at least with respect to geographic region.
- Commuting pattern data addresses only home-to-work travel. The population not employed outside the home is assumed to always remain in the residential Census tract. Further, although several of the HAPEM6 microenvironments account for time spent in travel, the travel is assumed to always occur either in the home or work tract. No provision is made for the possibility of passing through other tracts during travel.
- The ME *PEN* factor distributions incorporated into HAPEM6 were derived from reported measurement studies. The data available were quite limited. As a result most factors were not derived from a representative sample of measurements, and many were inferred on the basis of measurements of different pollutants and/or MEs that would be expected to be similar. In addition the derivation of the *PEN* factors was based on the assumption that measured I/O ratios of 1.0 or less indicate the absence of indoor emission sources. Because this assumption is unlikely to be uniformly valid, *PEN* factors are likely to overestimate penetration by some unknown amount.
- The ME *PROX* factor distributions incorporated into HAPEM6 for the on-road vehicle source category were derived from modeling studies for Portland OR. They are subject to the standard uncertainties of air dispersion modeling. They are also subject to the uncertainties of extrapolating from the traffic patterns of Portland OR to other locations.
- Air quality data from modeling studies are uncertain, due to simplifications incorporated into modeling algorithms and limitations of input data (e.g., emissions, meteorology). Air quality measurements are also uncertain due to limitations of measurement technology (e.g., minimum detection limits) and unknown representativeness of monitoring locations.

1.4. Applicability

HAPEM6 is a screening-level exposure model appropriate for assessing **average long-term** inhalation exposures of the general population, or a specific sub-population, over spatial scales ranging from urban to national. Due to its design features, HAPEM6 is not appropriate for modeling short-term (*e.g.*, hourly or daily) exposure events, nor should the model be used to assess the exposure of individuals.

The model is designed to look at the “typical” inhalation exposures of different groups, including their variance across the population. However, it should not be used to quantify episodic “high-end" inhalation exposure that results from highly localized pollutant concentrations and/or activities that, by their nature, could result in potentially high exposures (e.g., occupational exposures). Furthermore, HAPEM6 cannot address cumulative exposure from multiple pollutants nor pollutant mixtures.

1.5. Brief History of the Hazardous Air Pollutant Exposure Model

In 1985, the EPA's Office of Mobile Sources (OMS)[7] developed a model for estimating human exposure to nonreactive pollutants emitted by mobile sources. This model was similar to the probabilistic NAAQS Exposure Model (pNEM) in that both simulated the movements of population groups between home and work locations and through various microenvironments. They differed, however, in several respects. The pNEM provided minute-by-minute exposure estimates, which could be averaged over longer time periods, whereas HAPEM provided annual average exposure estimates. The pNEM included stochastic processes for estimating uncertainty and variability, while HAPEM provided only point estimates. HAPEM also included the ability to estimate cancer incidence through the use of risk factors developed by EPA, a capability not available to pNEM.

The OMS extended the modeling methodology in 1991 to estimate annual average carbon monoxide (CO) exposures in urban and rural areas under specified control scenarios. The model was renamed the Hazardous Air Pollutant Exposure Model for Mobile Sources (HAPEM-MS). HAPEM-MS used the estimated annual average CO exposures to estimate annual average exposures to various HAPs associated with mobile sources. This was achieved by assuming the annual average exposure to each HAP was linearly proportional to the annual average CO exposure. The model was limited by the fact that it could only be run for specified urban areas with ambient fixed-site CO monitors.

Shortly after, EPA's Office of Research and Development (ORD) developed an enhanced version of HAPEM-MS, called HAPEM-MS2. HAPEM-MS2 sub-divided the annual exposures by calendar quarter (*i.e.*, 3-month periods) to more accurately estimate exposures to mobile sources as a function of outdoor air temperature. HAPEM-MS2 also increased the number of microenvironments from 5 to 37, increased the number of demographic groups from 11 to 23, and increased the size of the activity pattern database.

In 1996, ORD further enhanced HAPEM by creating another generation of the model called HAPEM-MS3. These enhancements included adding the ability to customize the demographic groups, updating the census data using the 1990 U.S. census, and developing an algorithm for estimating ambient impacts in residences with attached garages.

Until the spring of 1998, HAPEM-MS3 could only be run on an EPA mainframe computer. During early model development, use of the mainframe was necessary, because the model required the storage of large data files and the calculation of large internal arrays. After 1998, with advances in computing technology, it became possible for HAPEM-MS3 to be executed on a "workstation." To this end, in the spring of 1998, HAPEM-MS3 was migrated (*i.e.*, transferred) to the UNIX operating system on a workstation. During the migration, further enhancements to the model were made, including a new time-activity database derived from CHAD, a new air quality program that

[7] The EPA changed this name to the Office of Transportation and Air Quality in 1999.

automatically selects air pollutant monitoring sites, and a more efficient implementation of the commuting algorithm.

Immediately after the release of the UNIX-version of HAPEM-MS3, ORD, in association with the EPA's Office of Air Quality Planning and Standards (OAQPS), again made substantial improvements to the model. The newer model had two distinct improvements over the 1998 UNIX-version. First, the flexibility of the model was expanded to allow the use of modeled air quality data as well as measured data. This added functionality allowed the second improvement, expanding the areal extent of the model to include the entire contiguous U.S. at the census tract levels. With these improvement, the model was able to **directly** estimate exposures to HAPs, and hence the model was again renamed by dropping the mobile source (-MS) acronym.

An earlier version of the model, HAPEM4, had other enhancements as well. These included broader flexibility in defining the study area (this can range from a single Census tract up to the entire contiguous U.S.), population and commuting data for all census tracts in the country, a database of (non-variable) ME factors for more than 30 HAPs, stochastic selection of activity data, and the ability to allow the user to change internal modeling parameters such as the number of microenvironments.

The EPA used HAPEM4 in its National Air Toxics Assessment (NATA) national-scale assessment for 1996. This program was designed to address the air toxics problem in the U.S. and is an important part of EPA's Integrated Urban Air Toxics Strategy.

The previous version of the model, HAPEM5, incorporated additional enhancements. These included the use of variable ME factors and air quality data that are spatially variable within Census tracts. It also contained an more refined approach for extrapolating short-term (24-hour) activity patterns into annual activity patterns, to better reflect the day-to-day variability in an individual's activities. HAPEM5 was applied as part of the NATA national-scale assessment for 1999.

The most recent improvements to HAPEM for the current version include the ability to account for enhanced onroad-vehicle-related HAP concentrations in the vicinity of major roadways, a more accurate characterization of the fraction of the population of each Census tract that commutes to work, and a more accurate estimates of the duration of commuting to work.

NOTE: The current version of HAPEM6 also contains enhanced algorithms for estimating exposure concentrations from indoor emission sources. However, the algorithms have undergone only limited testing, and the data bases required to implement these algorithms are currently under development. Therefore, we do not recommend the use of these algorithms at the present time.

This page intentionally left blank.

2. Getting Started—An Overview of HAPEM6

This chapter provides the user the basic information needed to run HAPEM6. The topics addressed in this chapter include the functions of the programs that comprise the HAPEM6, the contents of the various input and output files, and the meanings of parameter values. The chapter has been separated into the following sections.

Section 2.1	Describes the general structure of HAPEM6, the input and output files, and the parameter settings.
Section 2.2	Discusses considerations for changing parameter settings.
Section 2.3	Provides instructions for setting up and running HAPEM6.

Figure 2-1 presents a graphical overview of the HAPEM6 model, including the types of data needed and the types of output produced by the model. The user should refer back to the figure while reading this chapter to understand how all the pieces of the model fit together.

2.1. Model Structure

Five programs comprise HAPEM6. These are:

1. DURAV
2. INDEXPOP
3. COMMUTE
4. AIRQUAL
5. HAPEM

Because several output files of these programs are used as input to other programs of the set, it is important to execute them in the order presented. Program 3, COMMUTE, is omitted if commuting is not included in the exposure assessment.

For a given modeling domain (e.g., a state, a set of states, entire US) the first three of these programs need to be executed only once, even if several different air quality scenarios/pollutants are evaluated. Programs 4 and 5 need to be executed one time each for each air quality scenario/pollutant. The modeling domain for running programs 4 and 5 must be included in the modeling domain used for running Programs 1- 3, but may be smaller. For example, if programs 1- 3 are run for the entire US, the output files from these runs may then be used by programs 4 and 5 for evaluating a single state or set of states.

The HAPEM6 programs use twelve user-supplied input data files, and two or more *parameter* files. All are in ASCII format. A *parameter* file identifies the user-supplied input files, the output files available to the user, and specifies the parameter settings for a model run.

2.1.1. Parameter Files

The required *parameter* file information for running each of the 5 HAPEM programs is presented in Table 2-1 as user defined files and user defined parameters. The contents of each of the user defined files is described below. With one exception, noted below, any information in the *parameter* file in addition to that required will be ignored by the program. This allows wide flexibility in the use of *parameter* files. For example, one approach would be to construct and use a separate *parameter* file for each program in the HAPEM set, with each parameter file including only the information required by its corresponding program. An alternative approach is to use the same parameter file for running more than one program by aggregating all the information needed for each program into the file. We recommend using one *parameter* file for running programs 1 - 3, and a separate *parameter* file for each set of program 4 - 5 runs, i.e., each air quality scenario. This configuration provides a balance between avoiding errors in duplicating information used by more than one program, and keeping track of the input files used for each air quality scenario. In order to avoid using the wrong parameter file, a checking feature has been included in the first 3 programs so that they will stop if the keyword ***nreplic*** (required by the AIRQUAL and HAPEM programs) is encountered in the parameter file.

We recommend that the user prepare a separate parameter file for each air quality scenario/pollutant evaluation. Using distinct files, rather than re-using the same file repeatedly (*i.e.*, by editing it between runs), will assist the user in keeping track of the differences between various model runs, because the parameter file serves as a record of the job settings.

The name of the *parameter* file is specified on the command line just after the name of the executable file to be run.

In order for a record in the *parameter* file to be processed by the program it must contain an equal sign, i.e., “=”. Other records in the file are ignored by the program. The left side of the equal sign contains a user supplied key word or phrase for each user defined file and parameter, as indicated in Table 2-1. Note that the word “file” is part of the file key phrase, e.g., “activity file”. On the right side of the equal sign a full file path name (all files except the final exposure output files and the indoor source files), a path name (the final exposure output files and the indoor source files[8]), or a parameter value is specified. As currently configured HAPEM6 creates an exposure output file for each state/pollutant combination. The names of these files are constructed by the program based on the pollutant SAROAD code and the state FIPS code, so that the user need not supply names for these files in the *parameter* file. However, the user must supply the SAROAD code for the pollutant in the *parameter* file of the HAPEM program as the value for the parameter ***sarod***.

The names of the other user defined input and output files should consist of two parts, separated by a dot (“.”). The part of the name preceding the dot, including the path, is the root and the part following the dot is the extension. Note that the maximum record length in the parameter file that will be processed by the program is 120 characters, including the key word/phrase, the equal sign, and the file name/path or parameter value. The number of spaces between the keywords and the “=” signs and between the “=” signs and the file names are not fixed, and therefore can be any reasonable number. Figure 2-2a and 2-2b present example

[8] Indoor source algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword ***CAS*** to 99999.

parameter files that can be used to run all the first 3 and the last two HAPEM6 programs, respectively. Note that the input and output file names must be listed before the parameter settings.

Figure 2-1.
Overview of HAPEM6

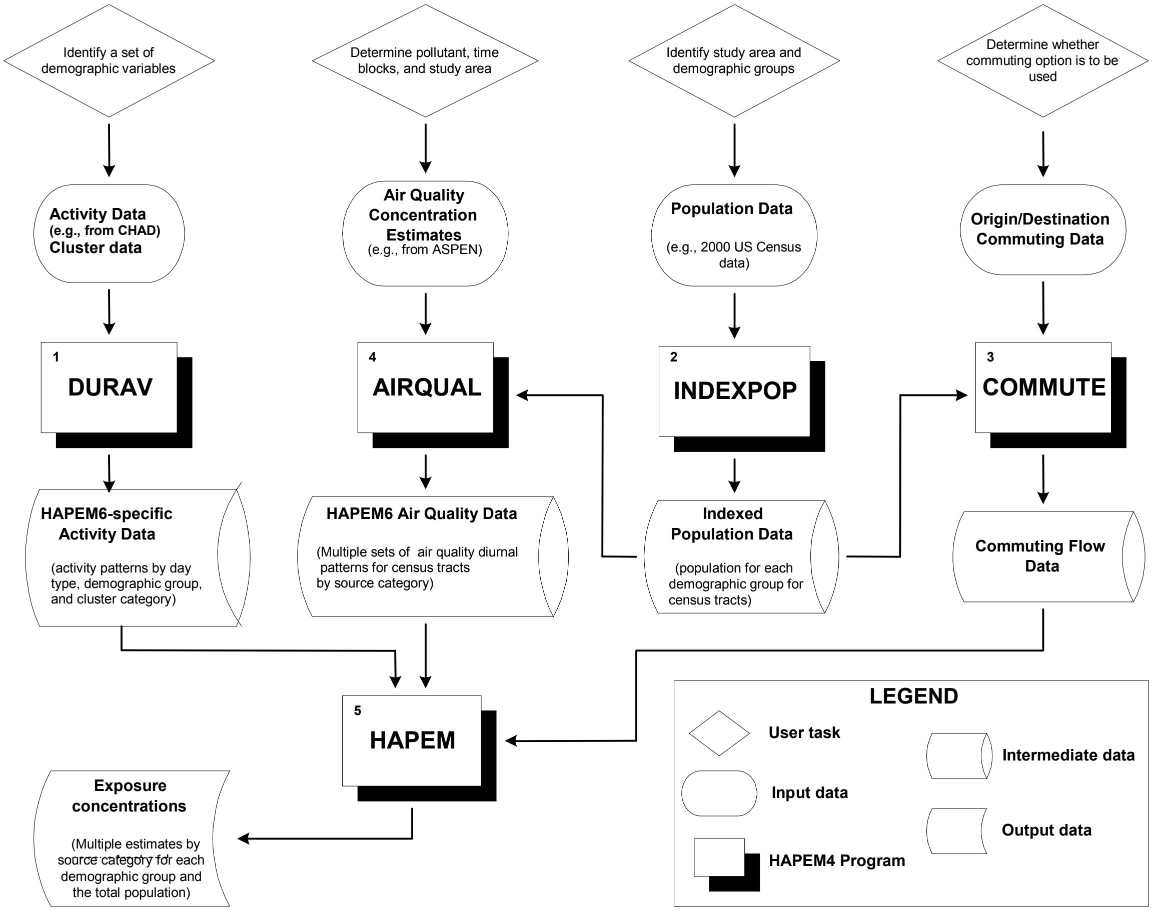
Identify a set of demographic variables
Determine pollutant, time blocks, and study area
Identify study area and demographic groups
Determine whether commuting option is to be used
Activity Data
(e.g., from CHAD)
Cluster data
Air Quality Concentration Estimates
(e.g., from ASPEN)
Population Data
(e.g., 2000 US Census data)
Origin/Destination Commuting Data
1
DURAV
4
AIRQUAL
2
INDEXPOP
3
COMMUTE
HAPEM6-specific Activity Data
(activity patterns by day type, demographic group, and cluster category)
HAPEM6 Air Quality Data
(Multiple sets of air quality diurnal patterns for census tracts by source category)
Indexed Population Data
(population for each demographic group for census tracts)
Commuting Flow Data
5
HAPEM
Exposure concentrations
(Multiple estimates by source category for each demographic group and the total population)
LEGEND
User task
Input data
HAPEM4 Program
Intermediate data
Output data

Table 2-1.
Keywords for parameter files and example file names

User/Model Defined	Inputs	Outputs
	DURAV.f90	
User defined files	*activity* file (e.g., *durhw.txt*) *cluster* file (e.g., durhw_cluster.txt)	*log* file *counter* file
User defined parameters	nmicro nblock hblock ntype ngroup	
HAPEM6 defined files		*durhw.wrong_chad* *durhw.da* *durhw.nonzero*
	INDEXPOP.f90	
User defined files	*population* file (e.g.,*census2000.txt*) *DistToRoad* file (e.g., *distprob.txt*) *CommutTime* file (e.g., *commtime.txt*) *CommutFrac* file (e.g., *commfrac.txt*) *statefip* file	*log* file *counter* file
User defined parameters	region1 region2 ngroup	
HAPEM6 defined files		*census.da* *census_direct.ind* *census.county_tract_pop_range* *census.state_county_pop_range* *distprob.STIDX* *distprob.dat* *commtime.STIDX* *commtime.dat* *commfrac.STIDX* *commfrac.dat*

(continued)

Table 2-1.
Keywords for parameter files and example file names

User/Model Defined	Inputs	Outputs
COMMUTE.f90		
User defined files	*commuting* file (e.g.,*comm2000.txt*) *population* file (*e.g.,census2000.txt*) *DistToRoad* file (e.g., *distprob.txt*) *CommutTime* file (e.g., *commtime.txt*) *CommutFrac* file (e.g., *commfrac.txt*) *statefip* file	*log* file *counter* file *mistract* file
User defined parameters	region1 region2 keep	
HAPEM6 defined files	*census_direct.ind* *census.county_tract_pop_range* *census.state_county_pop_range* *distprob.STIDX* *distprob.dat* *commtime.STIDX* *commtime.dat* *commfrac.STIDX* *commfrac.dat*	*comm.da*; *comm.ind* *comm.st_comm1_ fip_ range*
AIRQUAL.f90		
User defined files	*air quality* file (*e.g.,benzene.txt*) *population* file (*e.g.,census.txt*) *DistToRoad* file (e.g., *distprob.txt*) *statefip* file	*log* file *counter* file *mistract* file
User defined parameters	hblock nsource ngroup region1 region2 nreplic	

(continued)

Table 2-1.
Keywords for parameter files and example file names

User/Model Defined	Inputs	Outputs
HAPEM6 generated files	*census.da* *census_direct.ind* *distprob.STIDX* *distprob.dat*	*benzene.da* *benzene.air_da* *benzene.pop_air_da* *benzene.state_air_fip_range* *benzene.state_air1_fip_range* *benzene.state_air2_fip_range*
	HAPEM.f90	
User defined files	*factors* file (e.g., *factors.txt*) *mobiles* file (e.g., *mobiles.txt*) *population* file (e.g.,*census2000.txt*) *air quality* file (*e.g.,benzene.txt*) *commuting* file (e.g.,*comm2000.txt*) *activity* file (e.g., *durhw.txt*) *ClusTrans file (e.g., clustertransa.txt)* *Product* files[9] (specify path only) *AutoPduct file*[10]	*log* file *counter* file *mistract* file *afile* file (specify path only)

(continued)

[9] A path to DistToRoad one or more indoor emission source inputs for the HAPEM6 indoor source algorithms is specified in this statement. These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword ***CAS*** to 99999. Since no indoor source files will then actually be utilized by the HAPEM program, any existing path may be specified.

[10] The full path name of an existing file must be specified as the *AutoPduct* file in HAPEM6, although its only function is as input to the indoor source algorithms. These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword ***CAS*** to 99999. Since the *AutoPduct* file will then not actually be utilized by the HAPEM program, any existing file name may be specified, other than those otherwise specified for input or output for the HAPEM program.

Table 2-1.
Keywords for parameter files and example file names

User/Model Defined	Inputs	Outputs
User defined parameters □	pollutant CAS[11] unit EPA nmobiles nemicro nbmicro nvehicles npublict year backg sarod nmicro hblock ntype ngroup nsource nreplic region1 region2 Rseed1 Rseed2 Rseed3 B_00 B_02 B_05 B_16 B_18 B_65	

(continued)

[11] The Chemical Abstract Service (CAS) registry number is used to identify files for inputs to the HAPEM indoor source algorithms. These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword ***CAS*** to 99999.

Table 2-1.
Keywords for parameter files and example file names

User/Model Defined	Inputs	Outputs
HAPEM6 generated files	*durhw.da* *durhw.nonzero* *benzene.da* *benzene.air_da* *benzene.pop_air_da* *benzene.state_air_fip_range* *benzene.state_air1_fip_range* *comm.da* *comm.ind* *comm.st_comm1_fip_ range*	

The HAPEM programs also create several intermediate output files that are used as input to other programs in the HAPEM set, but are not directly useful for the user. The HAPEM6 programs generate the names of the intermediate output files by changing the filename extensions (*i.e.*, the text after the dot) of the input file names. An example set of filenames, including the intermediate files generated by the programs, is shown in Table 2-1, with example user defined filenames in parentheses. In the COMMUTE program, two of these intermediate files (*census.county_tract_pop_range* and *census.state_county_pop_range)* will be deleted at the end of the program unless the keyword variable *keep* is set to *yes*.

Besides the input and output files, the HAPEM6 programs create a set of user defined diagnostic output files. The main one is a *log* file, which records information about the execution of the programs, including some error messages. Another is a *counter* file that keeps track of the numbers of elements in various processed files, some of which are used by subsequent programs. A third diagnostic file is the *mistract* file. This file keeps track of tracts in the *population* file that are not matched by tracts in the *commuting* file, tracts in the *population* file that are not matched by tracts in the *air quality* file, and of tracts in the *commuting* file that are not matched by tracts in the *air quality* file. Only tracts included in both the *population* and *air quality* files are processed by HAPEM6 since both these pieces of information about a tract (population and air quality) are needed to make an exposure estimate. If commuting is included in the simulation and the tract is missing from the *commuting* file, it is assumed that all workers residing in that tract stay in the home tract for work.

Figure 2-2a.
Example *parameter* file for running HAPEM6 programs 1-3

```
INPUT FILES:
 activity file            =      input\activity pattern\durhw.txt
 cluster file             =      input\activity pattern\durhw_cluster.txt
 population file          =      input\population\census2000.txt
 commuting file           =      input\commute\comm2000.txt
 CommutTime file          =      input\others\commtime_new.txt
 CommutFrac file          =      input\others\commfrac.txt
 DistToRoad file          =      input\others\distprob.txt
 statefip file            =      input\statefip.dat
OUTPUT FILES:
 log file                 =      output\log_file.txt.
 counter file             =      output\counter.dat
 mistract file            =      output\mistract.dat
PARAMETER SETTINGS:
 region1        =      1
 region2        =      53
 keep           =      YES
 nmicro         =      14            ! Number of microenvironments
 nblock         =      24            ! Number of time blocks/day in activity file
 hblock         =       8            ! Number of time blocks/day for analysis
 ntype          =       3            ! Number of day types
 ngroup         =       6            ! Number of demographic groups
```

Figure 2-2b.
Example *parameter* file for running HAPEM6 programs 4-5

```
INPUT FILES:
 activity file     =      input\activity pattern\durhw.txt
 ClusTrans file    =      input\activity pattern\clustertransa.txt
 population file   =      input\population\census2000.txt
 commuting file    =      input\commute\comm2000.txt
 air quality file  =      input\airqual\benzene.txt
 factors file      =      input\factor\gas_factors.txt
 mobiles file      =      input\factor\mbfactors_ALL.txt
 DistToRoad file  =       input\others\distprob.txt
 statefip file    =     input\statefip.dat
 product file     =     input\
 AutoPduct file   =     input\empty.txt

OUTPUT FILES:
 log file         =     output\log_file.txt
 counter file     =     output\counter.dat
 mistract file    =     output\mistract.dat
 afile            =     \output\
PARAMETER SETTINGS:
 pollutant        =     benzene
 CAS              =     99999
 units            =     ug/m3
 year             =     1999
 region1          =     1
 region2          =     53
 EPA region       =     All
 sarod            =     45201
 backg            =     0.00
 nmicro           =     14          ! Number of microenvironments
 hblock           =      8          ! Number of time blocks/day for analysis
 ntype            =      3          ! Number of day types
 ngroup           =      6          ! Number of demographic groups
 nsource          =      4          ! Number of source categories
 nmobiles         =      3          ! Sequence # of up to 10 air quality source categories which are on-road
mobile sources
 nbmicro          =      1          ! Sequence # of first indoor microenvironment
 nemicro          =     10          ! Sequence # of last indoor microenvironment
 nvehicles        =       8         ! Sequence #s of up to 10 microenviroments used for private commuting
(e.g., cars and trucks)
 npublict         =      9          ! Sequence #s of up to 10 microenviroments used for public
transportation commuting (e.g., busses and trains)
 nreplic          =     30          ! Number of replicates for each demographic group/tract
 Rseed1           =     -10         !Random seed (negative for selecting activity pattern data
 Rseed2           =      -1         !Random seed (negative) for selecting microenvironment factors
 Rseed3           =      -1         !Random seen (negative) for selecting air quality data
```

DEMOGRAPHIC GROUP DEFINITIONS:

B_00	=	Ages 0-1
B_02	=	Ages 2-4
B_05	=	Ages 5-15
B_16	=	Ages 16-17
B_18	=	Ages 18-64
B_65	=	Ages > 65

2.1.2. DURAV and the Activity and Cluster Files

The DURAV program performs three main functions.

- It categorizes and groups population activity data extracted from CHAD into demographic groups, day types (season, day-of-week), commuting status groups, and cluster categories.
- If a different number of daily time blocks is specified for the analysis than in the activity data file, it processes the activity records so that the number of time blocks matches the number specified for the analysis.
- It creates a sequential ASCII file of the activity pattern records for use by the HAPEM program.

The *activity* file is the primary input file for the DURAV program. The default file, *durhw.txt*, contains data extracted from CHAD, describing the amount of time spent in various microenvironments by individuals. Each record in the *activity* file consists of one person-day (*i.e.*, 1,440 minutes of data for an individual) of activity data. This information is not an activity sequence, rather it is the total number of minutes spent in each microenvironment during each block of time throughout the day (*i.e.*, the time increments used per twenty-four hour period).

For example, in the default *activity* file, *durhw.fix_2005.txt*, there are 14 microenvironments, (24) one- hour time blocks, and 2 exposure districts (home and work), resulting in a total of 1,776 duration values. The duration in each of the 15 microenvironments for the first hour comes first in the *activity* file, followed by the 14 durations for the second hour, etc. This pattern is repeated for all twenty-four hours for the home exposure district, and then for the 24 hours and 14 microenvironments of the work district.

The number of time blocks in the *activity* file is specified by the user in the *parameter* file of DURAV as ***nblock***. The number of microenvironments in both the *activity* file and the *factors* and *mobiles* files (discussed below) must be the same and is specified in the *parameter* files of DURAV and HAPEM as ***nmicro***.[12] The number of duration values in the *activity* file must equal twice the product of the values of the ***nmicro*** and ***nblock*** settings in the *parameter* file. The

[12] As explained in section 2.1.6, there must be ***nmicro*** records for each onroad mobile source category in the *mobiles* file.

sum of the duration values for each individual profile should always equal 1,440 minutes (*i.e.*, there should be no unaccounted time); otherwise, the program will stop. Each duration must be specified as an integral (*i.e.*, no decimals) number (this number can be zero) of minutes in each microenvironment.

The number of time blocks for the analysis is specified in the *parameter* files of DURAV, AIRQUAL, and HAPEM as ***hblock***. The number may be less than or equal to ***nblock***; however, it must be an integral factor of ***nblock***, so that the activity time blocks can be combined if necessary to match to match ***hblock*.** For example, if ***nblock*** is 24 and ***hblock*** is set to 8, DURAV will combine the (24) one-hour activity time blocks into (8) three-hour activity time blocks.

Each record in the *activity* file also contains information about the individual from whose activities the data were derived, so that the records can be classified into demographic groups. The definitions of these groups are part of the DURAV source code, so that in order to change the demographic group definitions the source code must be modified and recompiled. Similarly, the definitions of day types, pertaining to season and day-of-week for categorizing activity patterns, are part of the DURAV source code. The number of demographic groups, ***ngroup***, is specified in the *parameter* files of DURAV, INDEXPOP, AIRQUAL, and HAPEM. The number of day types, ***ntype***, is specified on the *parameter* files of DURAV and HAPEM.

The cluster category for each CHAD record, identified by CHAD identification code, is specified in the *cluster* file. The current version of DURAV divides the activity data into 12 demographic groups, based on age (6 categories) and commuting status (yes or no). Activity pattern data are also separated into 3 day types: summer weekdays, other weekdays, and weekends. The number of clusters, derived from a statistical cluster analysis procedure, ranges from 1 to 3, depending on the demographic group and day type.

2.1.3. INDEXPOP and the Population File, DistToRoad File, CommutTime File, and CommutFrac File

The INDEXPOP program performs three main functions:

- It creates a direct access file of population data to be used in AIRQUAL.
- It creates sequential ASCII index files for the population data census tracts, to facilitate file searching in COMMUTE and AIRQUAL.
- It creates direct access files and associated index files of the data in the *DistToRoad*, *CommutTime*, and *CommutFrac* files, to be used in COMMUTE and AIRQUAL

The main input file to INDEXPOP is the *population* file, which provides the number of people in each demographic group (defined in the DURAV source code) for each census tract in the study area under investigation. The data must be sorted according to state FIPS, county FIPS, and tract code. These data are typically obtained from the U.S. Census Bureau's census surveys. For example, the default *population* file contains 2000 US Census population counts for each of the demographic groups defined in the current version of the DURAV source code for each Census tract in the US.

Other files with tract-specific information about the population, such as *DistToRoad*, *CommutTime*, and *CommutFrac* are also first processed in this program. *DistToRoad* provides information on the fraction of each age-group in each tract that resides within 3 different distance categories of major roadways, as well as the fraction of the tract area that is within each distance category. *CommutTime* provides information on the average commuting time for commuters in each tract. *CommutFrac* provides information on the fraction of each age-group that commutes to work in each tract.

2.1.4. COMMUTE and the Commuting File, DistToRoad File, CommutTime File, and CommutFrac File

The COMMUTE program performs three main functions:

- It creates a file identifying for each Census tract (i.e., home tract) the associated set of work tracts (i.e., tracts in which the residents of the home tract work), the fraction of home tract workers in each work tract, and the normalized centroid-to-centroid distance between home tract and each work tract. The normalized distance is the distance/(average distance). The normalized distance is combined with the average commuting time for the tract to estimate the commuting time for the home-tract/work-tract pair in HAPEM.
- It creates a sequential index file to facilitate file searching in HAPEM.
- It adds the tract–specific information from the *DistToRoad*, *CommutTime*, and *CommutFrac* direct access files (created in INDEXPOP) to the commuting index file

The *commuting* file is the main input file to the COMMUTE program. The default *commuting* file was derived by US EPA Office of Research and Development from the 2000 Census data collected as part of the Census Transportation Planning Package (CTPP). The data files specify the number of residents of each tract that work in that tract and every other US Census tract (i.e., the population associated with each home tract/work tract pair) and the distance between the centroids of the two tracts. While there are approximately 500 million pairs of tracts nationwide within a reasonable commuting distance of each other, only about 5 million of these pairs have a non-zero flow of commuters. Only those pairs with non-zero flows are included in the *commuting* file.

An important issue pertaining to this commuting data is that workers do not always travel between their home and work locations on a daily basis. The larger the distance between home and work, the greater the likelihood that daily commuting does not occur. For example, places of residence in the lower 48 states appear with Alaskan places of work. These workers are almost surely not commuting on a daily basis between the continental U.S. and Alaska. To address this issue the commuting flows were examined as a function of distance. To examine how the decline in commuting flow is affected by distance, researchers plotted the natural log of the natural log of the total flow versus distance. This plot revealed that the ln(ln(total flow)) is nearly linear for distances ranging from 0 to about 100 km. For distances greater than 100 km, the graph exhibits a decreasingly negative slope with distance (*i.e.*, the curve “flattens out”). These findings suggest that people’s “commuting behavior” is fairly consistent, on an aggregate basis, to a distance of approximately 100 km. Then, at greater distances, factors other than daily commuting may become increasingly important. Therefore, within the COMMUTE program a limit for the distance between home and work distance is specified, such that commuting flows for greater distances are not processed. The distance limit is currently set at 120 km.

2.1.5. AIRQUAL and the Air Quality File, and DistToRoad File

The AIRQUAL program performs three main functions:

- It creates a sequential file of air quality data to be used in HAPEM.
- It determines the number of data records for each census tract in the *air quality* file
- It creates index files to facilitate file searching in HAPEM.
- It adds the tract–specific information from the DistToRoad direct access file (created in INDEXPOP) to the air quality index files

The *air quality* file contains the ambient air concentrations that are used by the AIRQUAL program. The file records have concentration contributions from multiple emission source categories for multiple time blocks for a census tract, as well as a time-invariant location-specific background concentration. There may be multiple such records for each tract, representing spatial variability throughout the tract. AIRQUAL requires a separate *air quality* file for each pollutant being evaluated. Details about the format of the *air quality* file can be found in Chapter 3.

The number of outdoor emission source categories is specified in the *parameter* files of AIRQUAL and HAPEM as ***nsource***, and must match the number in the *factors* file, discussed below. The user specifies the number of time blocks for the analysis in the *parameter* files of DURAV, AIRQUAL, and HAPEM as ***hblock***. As discussed above, this value must be an integral factor of ***nblock***, the number of time blocks in the *activity* file, so that the activity time blocks can be combined if necessary to match to match ***hblock*****.** Similarly, ***hblock*** may also be greater than or equal to the number of time blocks in the *air quality* file. But it must be an integral multiple of the number of air quality time blocks, so that the air quality values can be replicated if necessary to create ***hblock*** air quality values. For example, suppose the *air quality* input file has (8) three-hour time blocks per day. If ***hblock*** is set to 24 AIRQUAL will create 24 air quality time blocks with three replicates of each of the 8 air quality values.

2.1.6. HAPEM, the Microenvironmental Factors and Mobiles Files, and the Activity ClusTrans File

The HAPEM program performs six main functions:

1. For each demographic group in each census tract, it randomly selects ***nreplic*** sets of microenvironment (ME) factors based on the distribution data provided in the *factors* and *mobiles* file. Each set contains a subset of ME factors randomly selected for each of the time blocks (for the *PEN* and *ADD* factors) or each of the sources (for the *PROX* and *LAG* factors). Each subset contains randomly selected ME factors for each of ***nmicro*** microenvironments.
2. For each demographic group in each census tract, it randomly selects ***nreplic*** sets of air quality data from the data sets available for a census tract.
3. For each demographic group in each census tract, it creates ***nreplic*** sets of average activity patterns, where a set contains one average pattern for each day type. An average activity pattern for each day type is calculated as a weighted average of activity patterns randomly

selected from each cluster in a demographic-group/day-type/commuting-status combination. The weights are determined by the relative frequencies of cluster types randomly selected in a one-stage Markov process[13], based on the cluster transition probabilities provided in the C*lusTrans* file.

4. For each activity pattern for a commuting demographic group, it randomly selects a work census tract with probability weighting based on the fraction of residents that work in that tract.

5. For each census tract it estimates the concentration in each microenvironment based on microenvironment factors and outdoor concentrations.

6. It combines activity patterns, commuting, and microenvironment concentration estimates to calculate ***nreplic*** annual average exposure concentrations for each demographic group in each census tract

The microenvironment (ME) *factors* and *mobiles* files provide the factors used to calculate an estimated microenvironmental concentration from an outdoor concentration. This methodology allows the user to specify values (distributions or point estimates) for three types of ME factors: penetration factors, proximity factors, and additive factors. These factors are combined with the outdoor concentration estimates according to the following algorithm.

$$\text{ME concentration} = PROX \times PEN \times \text{outdoor concentration} + ADD$$

The outdoor concentration is the sum of the concentration contributions from each outdoor emission source category and background.

The penetration factor, *PEN,* is an estimate of the ratio of the ME concentration contribution (from a given emission source category) to the concurrent outdoor concentration contribution in the immediate vicinity of the ME.

The proximity factor, *PROX,* is an estimate of the ratio of the outdoor concentration in the immediate vicinity of the ME to the outdoor concentration represented by the air quality data. The air quality data represent an average over some geographic area (i.e., some subset of a census tract). For most situations the default *factors* file specifies a *PROX* value of 1.0, i.e., an outdoor concentration contribution in the immediate vicinity of the Census tract equal to the Census tract average concentration contribution. However, when assessing exposure to motor vehicle emissions, for MEs near roadways (e.g., in-vehicle, residences near major roadways) the pollutant concentration contribution in the immediate vicinity of the ME is expected to be higher than the average pollutant concentration contribution over the Census tract, i.e., *PROX* is expected to be greater than 1.0, and this is reflected in the default *factors* and *mobiles* files.

ADD is an additive factor that accounts for emission sources within or near to a microenvironment, i.e., indoor emission sources. Unlike the other two factors, the *ADD* factor is itself a concentration and therefore has units of mass/volume. The actual units used must be the same as those in the *air quality* file.[14]

[13] A one-stage Markov process is a sequence of events, such that at every step in the Markov chain the probability distribution for the next event depends on what the current event is.

[14] A data base of distributions of indoor source concentration contributions for several indoor source categories and subcategories is currently under development. The current version of the HAPEM program contains new, but untested algorithms to utilize the developing data base. Therefore, it is currently recommended that indoor sources

A fourth factor, *LAG*, is used to account for the possibility of very slow pollutant diffusion and penetration, so that the relevant air quality concentration value may be from the previous time block. A value of zero for *LAG* indicates no time lag, i.e., use the concurrent air quality value; otherwise, the previous time block value is used.

The *factors* file includes distributions for each of these factors for each ME/emission source category combination, with the exception of *PROX* and *LAG* factors for onroad mobile source emissions. These are contained in the *mobiles* file, with separate distributions specified for 3 distance-from-roadway categories. As noted above, the number of MEs in the *factors* and *mobiles* file must match the number in the *activity* file (i.e., ***nmicro***). Similarly, the number of outdoor emission source categories (i.e., ***nsource***) must match the number in the *air quality* file. And the *mobiles* file must contain ***nmicro*** records for each onroad mobile source category specified with ***nmobiles***.

There are three default *factors* files: one each for gaseous, particulate, and semi-volatile HAPs. And there are four default *mobiles* files: one each for benzene, 1-3-butadiene, diesel particle, as well as one for non-specific HAPs. Each of these default files is formatted for a single onroad mobile source category.

The default *factors* and *mobiles* files contains ME factors applicable to all the MEs included in the default *activity* file, for ***nsource*** emission source categories (e.g., point, area, onroad mobile, and nonroad mobile). These category-specific estimates were derived from reported measurement and modeling studies. Because, as noted above, a new approach to evaluating indoor sources is in development, the *ADD* factors are uniformly set to zero. And due to lack of data, *LAG* is uniformly set to zero.

The C*lusTrans* file specifies for each demographic group/day type combination the number of activity patterns in each of 2 to 3 clusters (derived from cluster analysis on the activity pattern data from CHAD), and the cluster-to-cluster transition probabilities (derived from the transition frequencies for multiple-day activity pattern records from CHAD). These values are used to create weights for averaging selected activity patterns, one from each cluster, to represent an individual within the demographic group for that day type.

2.1.7. Statefip File

The *statefip* file cross-references the 2-digit state FIPS codes for each US state to its numerical ranking on the list. The default *statefip* file contains 53 codes: one for each US state, the District of Columbia, Puerto Rico, and the US Virgin Islands. Therefore, the numerical rankings range from 1 to 53, although the FIPS codes in the file range from 01 to 78, since several possible codes in the sequence are skipped (i.e., not assigned to a state, district, or territory).

The *statefip* file is used in conjunction with the parameters ***region1*** and ***region2*** specified in the *parameter* files of INDEXPOP, COMMUTE, AIRQUAL, and HAPEM to specify the group of states to be included in the analysis, according to numerical ranking. For example, setting ***region1*** to 1 and ***region2*** to 53 results in assessment of all the states, districts, and territories in the default *statefip* file (assuming the input files contain all the necessary data). Alternatively,

be omitted from HAPEM6 applications until the database and algorithms have been tested and reviewed. To disable the indoor source algorithms, set keyword ***CAS*** to 99999.

setting both ***region1*** and ***region2*** to 5 results in assessment of the fifth state only: California with FIPS code 06.

The region range need not be the same for each of the five HAPEM programs; the range for each program may be the same as or smaller than the range for the preceding program, where the order of the programs is as specified above. For example, INDEXPOP and COMMUTE could be run for region range 1 to 53, while AIRQUAL and HAPEM are run for a single state.

Note that the ***region1*** and ***region2*** parameters specify the states for which the program will look for data in the input files. The input files need not contain data for every tract within the specified states, however. For example, if the *air quality* file contains data for only a subset of tracts within a state, AIRQUAL and HAPEM will simply make estimates for those tracts, as long as the state or states are specified within the ***region1*** and ***region2*** range.

2.1.8. Background Concentration

In addition to estimating exposure concentration contributions for each emission source category for which data are provided in the *air quality* file, HAPEM also estimates the exposure concentration contribution from the background outdoor concentration. The background concentration is an estimate of the outdoor concentration that would occur in the absence of any anthropogenic emissions within the modeling domain. It includes concentration contributions from natural sources, re-entrainment, global transport, and other anthropogenic sources outside the modeling domain. This background exposure contribution is added together with the emission source category contributions; the total exposure concentration is reported in the exposure output files.

The background concentration is comprised of two parts, either or both of which may be used. The first is a uniform background concentration throughout the study area, with the single value is specified as ***backg*** in the *parameter* file of HAPEM. The units of measurement must be the same as those used in the *air quality* file.

The second background concentration specification is a single value for each location specified in the *air quality* file, representing a spatially variable component of the background concentration.

2.1.9. Exposure Output Files

As currently configured HAPEM6 creates an exposure output file for each state/pollutant combination. The names of these files are constructed by the program based on the pollutant SAROAD code and the state FIPS code as follows:

XXXXX.YY.dat

where	XXXXX	=	the 5-digit SAROAD pollutant code specified by the ***sarod*** parameter
and	YY	=	the 2-digit state FIPS code

These output files contain ***nreplic*** records for each Census tract/demographic group combination. Each record identifies the Census tract, the demographic group, the number of people to which the exposure estimates apply (i.e., 1/***nreplic*** of the population of the demographic group in the Census tract), and exposure concentration contribution estimates: one each for the ***nsource*** outdoor emission source categories, one for background, one for each of four indoor source categories, and a total of the contributions from all outdoor emission source categories, background, and indoor sources.

2.2. Changing the Parameter Settings

The HAPEM6 was designed to be as easy to use as possible. With this in mind, the model's structure is such that, for routine applications, no changes need be made to the model's computer code. For most applications the user need only supply the model with the appropriately formatted input files and parameter specifications declared in the *parameter* files.

However, there are several changes that a user can make to HAPEM6 to "tailor" the model to his or her needs. Changes or modifications to the model are most easily accomplished by altering the parameter settings. The following discussion describes those parameters that can be altered.

2.2.1. Changing the Number of Microenvironments

In principle, HAPEM6 will work with any number of microenvironments. The number, specified as ***nmicro*** in the *parameter* files of DURAV and HAPEM, must match the number actually used in both the *activity* file and the *factors* and *mobiles* files. Definitions of the microenvironments do not appear anywhere in HAPEM6 program code.

The HAPEM6 programs should be able to accommodate anywhere from one up to at least 100 microenvironments. However, large numbers of microenvironments could result in input-file line lengths beyond a system's limits (particularly in the case of the *activity* file) if other parameters (such as the number of time blocks) are also set to large values.

2.2.2. Changing the Number And/or Definitions of the Demographic Groups

The number of demographic groups, specified as ***ngroup*** in the *parameter* files of DURAV, INDEXPOP, AIRQUAL and HAPEM, must be consistent in two places:

- The group definitions in the source code for the DURAV program,
- The number of columns in the *population* file,
- The number of columns in the *CommutFrac* file,
- The number of columns in the *DistToRoad* file, and
- The number of demographic groups specified in the *cluster* and *ClusTrans* files.

The definitions of the groups appear only once explicitly (in the DURAV program). However, these definitions are paired with the columns in the *population* file by numerical order, so if the group definitions are changed then the columns in the *population* file must also be changed.

The definitions are listed in the *parameter* file for the HAPEM program so that they can be repeated at the start of the final output file for tracking. The *parameter* file listing has no impact on the exposure results.

The 6 current age groups are as follows.

- 0 - 1
- 2 - 4
- 5 – 15
- 16-17
- 18 - 64
- 65+

The number of demographic groups is unlimited. However, the user is cautioned that for narrowly defined groups, there might not be enough activity pattern data to calculate a reliable group average or create meaningful activity pattern clusters. An extreme example of this is where no activity patterns fit a demographic group's definition, resulting in incorrect exposure calculations (*i.e.*, exposure concentrations equal to zero) for that group.

2.2.3. Changing the Number And/or Definitions of Day Types

Day types are used to guide the selection of the activity patterns. Demographic studies indicate that typical weekday (Monday-Friday) and weekend (Saturday-Sunday) activities differ significantly for most working people and school children. Furthermore, in certain respects, activities in summer (or warm weather) may differ from those in winter (or cold weather), especially for children or other non-workers. Currently, two variables, season and day of week, are used to determine three day types:

- Weekdays in summer (June - August).
- Other weekdays.
- Weekends.

In principle, year, month, day, season, temperature, rainfall, other meteorological variables, or even geographical variables could be used to assign day type. However, if there are too many day types, or if they are too narrowly defined, then there may not be enough activity pattern data fitting the day type definition to allow the determination of a reliable average or to create meaningful activity pattern clusters. If additional variables are used to define day types, then the programmer is advised to check that there are an adequate number of activity pattern profiles for each new day type.

2.2.4. Changing the Number And/or Definitions of Time Blocks

The traditional method for running HAPEM has been to use one hour time increments (referred to as time blocks). The HAPEM6, however, was designed to allow more flexibility in the selection of time blocks. Time blocks can range between one minute (the finest resolution available for the activity data) and one day, so in principle, there can be any number from one to 1,440 time blocks. In most practical applications, the number of time blocks will be twenty-four or less. In order to accommodate the possible adjustment of time blocks from ***nblock*** to ***hblock*** as discussed above, the time blocks must each be of equal size.

2.3. Setting Up a HAPEM6 Run

This section shows how to set-up and make a simple HAPEM6 run. Subsequent sections and chapters provide more detailed explanation about HAPEM6's input and output files and the model's programs.

The example shown in this section is for a hypothetical HAPEM6 analysis of benzene.

The most important consideration for making a HAPEM6 run is ensuring that the input files are accurate and correctly formatted. This is the responsibility of the user. To run the model, the user must provide twelve data input files, the *parameter* files (these files are used by the model to identify the name and location of the input and output files) and the five executable files that comprise the model. The programs can either be run consecutively by using a "batch" file, or they can be run independently.

Parameter files

The *parameter* files for this example, presented above in Figures 2-2a and 2-2b, can be used for running the five executables. The name of the *parameter* file must be specified in the command line immediately after the executable name.

Input/output files

As seen in Figures 2-2a and 2-2b, the input files (including full path names) are identified in the *parameter* files. The input files reside in a subdirectory named "input\". The main exposure output files (*afile*) are sent to a subdirectory named "\output\", along with the diagnostic output files (the *log* file, the *mistract* file, and the *counter* file). When the full path name is identified for an input or output file, it is not required that it reside in the same subdirectory as the executables.

The names of the input and output files must be identified in the *parameter* files before the parameter settings.

As noted above, an existing path name should be specified for the *product* files, and the full path name of any existing file (except other HAPEM input or output files) must be specified as the *AutoPduct* file in the *parameter* file used with HAPEM. In the future these file will comprise part of the input for evaluating indoor sources, but for now the file will not actually be utilized by the HAPEM program. To disable the indoor source algorithms, set keyword ***CAS*** to 99999.

Parameter settings

The "PARAMETER SETTINGS" in the *parameter* files shows that the region to be modeled is 1 through 53 (all states, the District of Columbia, Puerto Rico, and the US Virgin Islands), and the pollutant SAROAD code is 45201 (benzene).

The last group of information in the *parameter* file shows that there are fourteen microenvironments to be modeled (***nmicro***). This number of microenvironments must be consistent with the number of microenvironmental (ME) factors specified in the *factors* and *mobiles* files (*i.e.*, *gas_factors.txt* and *mbfactors_ALL.txt*) and the number of duration values specified in the *activity* file (*i.e.*, *durhw.txt*). The number of time blocks per day in the *activity* file is 24 (***nblock***), but the number of time blocks per day for the analysis is 8 (***hblock***), which is an integral factor of the ***nblock*** value, as explained above. The number of outdoor emission source categories is four (***nsource***). The data in the *air quality* file (for this example the file is *benzene.txt*) must be consistent with ***nsource***, and the number of time blocks must be an integral factor of ***hblock***, as explained above. The number of demographic groups (***ngroup***) must be consistent with the demographic groups specified in the DURAV source code and in the *population, CommutFrac, DistToRoad, cluster, and ClusTrans* files (*i.e.*, *census2000.txt, commfrac.txt, distprob.txt, durhw_cluster.txt, clustertransa.txt*). The number of replicates to be simulated for each demographic group in each tract is 30 (***nreplic***).

In addition there are five parameter settings that specify the sequence numbers of particular emission source categories and microenvironment types that are subject to special treatment in HAPEM6. In the example, the sequence number for the single on-road mobile source category in the *air quality* file is 3 (***nmobiles***). The sequence numbers of the indoor microenvironments (including in-vehicle) in the *factors* and *mobiles* files are 1-10 (***nbmicro*** through ***nemicro***). There is a single microenvironment for private commuting with a sequence number of 8 (***nvehicles***), and a single microenvironment for public transit commuting with a sequence number of 9 (***npublict***). (There may be up to 10 values each for ***nmobiles***, ***nvehicles***, and ***npublict***.)

The pollutant name (***pollutant***), measurement units (***units***), target year for the analysis (***year***), and the demographic group definitions are listed here by the user, so that they can be repeated at the beginning of the final output file for tracking. They have no effect on the exposure results.

2.3.1. Running HAPEM6 as a "Batch" Job

When running HAPEM6 by submitting batch jobs, each job should be allowed to finish before submitting the next job.

For this example a simple batch file was written to run the five HAPEM6 programs sequentially, with all five programs residing in the same directory as the batch file. The batch file is shown in Figure 2-3.

Because the *parameter* files specify the names or paths of all the input and output files, as well as the parameter settings, the batch file simply specifies the order in which the HAPEM6 executable programs will be run.

Figure 2-3.
Example "batch" file for running the five HAPEM6 programs

```
durav.exe p1.txt
indexpop.exe p1.txt
commute.exe p1.txt
airqual.exe p2.txt
hapem.exe p2.txt
```

2.3.2. Running HAPEM6 Programs Individually

Any of HAPEM6 programs can be run individually. The user must ensure that the required input files exist and are in the same location specified in the *parameter* file.

If a user is interested in running the DURAV program (this is typically the first program that is run when doing an exposure analysis), he or she would go to the subdirectory containing the executable program and type the following command on the DOS command line:

```
durav.exe p1.txt
```

The other HAPEM6 programs are run similarly.

As indicated in Table 2-1, COMMUTE, AIRQUAL and HAPEM all require input files that are generated from running other HAPEM6 programs. Therefore, if any of these programs is run alone, the user must ensure that the required HAPEM6 generated input files exist and are in the same subdirectory as the original input file from which their filenames were derived (see Table 2-1). For example, running AIRQUAL requires two files with filenames derived from the *population* file and two files with names derived from the *DistToRoad* file. For this example these files are *census.da*, *census_direct.ind*, *distprob.STIDX*, and *distprob.dat*, with filenames derived from *census.txt* and *distprob.txt*. Therefore, the *parameter* file for running AIRQUAL must specify the full path name of the *population* and *DistToRoad* files, and the four intermediate files must exist and reside in the subdirectories specified for *population* and *DistToRoad* files, respectively.

3. HAPEM6 Input Files

The HAPEM6 programs use twelve user-supplied input data files, and two or more *parameter* files. All are in ASCII format. The function of each of the files and their relationship to the structure of HAPEM6 are discussed in Chapter 2. The reader is referred to that chapter for an overview of HAPEM6 input files. This chapter summarizes that information, and presents the format of each of the user-supplied input files.

The *parameter* files are the central input files for HAPEM6 simulations and customized *parameter* files should be prepared for every simulation or set of simulations. It is best to save the *parameter* file used for each simulation or set of simulations under a unique name, so that the files from earlier simulations are not overwritten. A consistent naming system should be developed to pair each *parameter* file with the output files generated by the simulation or set of simulations. This pairing serves as one form of documentation for the model simulations, so the user can later determine which settings produced which results. Another form of documentation is the repetition of the parameter settings at the start of the final output file.

The remaining file names used by the HAPEM6 programs are input from the *parameter* file. Thus, the user must check that the *parameter* file refers to the correct file names before submitting a job. Which of the user-supplied files and HAPEM6-generated files are required for each of the five programs that comprise HAPEM6 is discussed in Chapter 2 and presented in Table 2-1.

As explained in Chapter 2, there are default files available for eleven of the twelve user-supplied input files. They are:

- The *activity* file;
- The *cluster* file;
- The *population* file (national scope);
- The *CommutTime* file (national scope);
- The *CommutFrac* file (national scope);
- The *DistToRoad* file (national scope);
- The *commuting* file (national scope);
- The *factors* files (one each for gaseous, particulate, and semi-volatile HAPs);
- The *mobiles* files (one each for benzene, 1-3-butadiene, diesel particles, and non-specific HAPs);
- The *ClusTrans* file; and

- The *statefip* file (national scope).

The user may provide his or her own files as replacements for any or all of these files, using the file formats described in this chapter.

The twelfth user-supplied file, the *air quality* file, must be provided by the user with the format described in this chapter.

3.1. Parameter Files

The *parameter* files contain eight types of information for use in HAPEM6 runs:

1. Paths and file names for the input data files and all output data files except the final exposure output files and the indoor source files (not currently used);
2. Path names for the final exposure output files and the indoor source files (not currently used);
3. Identification of the set of states (optionally including the District of Columbia, Puerto Rico, and the US Virgin Islands) for the simulation;
4. Identification of the pollutant, the units of measurement, the target year of the analysis, and the demographic group definitions;
5. A spatially uniform background concentration;
6. Internal parameter settings; and
7. Seed values for three random number generators.

All of this information is identified using keywords. The required *parameter* file information for running each of the five HAPEM6 programs is presented in Table 2-1 of Chapter 2 as user-defined files and user-defined parameters. The contents and format of each of the user defined files is described below. As explained in Chapter 2, with one exception any information in the *parameter* file in addition to that required by a program will be ignored by the program. (The exception is that programs 1 –3 will stop if the keyword ***nreplic*** -- required by the AIRQUAL and HAPEM programs -- is encountered in the parameter file.) Therefore, although a separate *parameter* file may be used for each program in the HAPEM set, it is possible to use the same parameter file for running programs 1-3 and another for running programs 4-5 by aggregating all the information needed for each program in the file. The format (including keywords) of a *parameter* file for running the HAPEM6 programs is presented in Figures 2-2a and 2-2b in Chapter 2.

The HAPEM6 programs only scan lines containing an equals sign. The word or words to the left of the equals sign identify which variable is being set and thus should not be changed. The data to the right of the equals sign are the values or settings that the user selects for the run. The pathnames should precede the parameter settings in the file. The user can add additional lines (*e.g.,* comments) anywhere to the *parameter* file. It is safest if these lines do not contain an equals sign, which could cause them to be parsed accidentally by HAPEM6 programs. To ensure that all the necessary information is specified, it is safest to edit an existing *parameter* file, changing only the comments and the right hand sides of the equations.

3.1.1. Specifying the Location and Names of Input and Output Files

In editing the *parameter* file, the user should typically provide the full path names for input and output files (except the final exposure output files and the indoor source files [not currently used]). The names can be up to 100 characters in length and should **not** use quotation marks to enclose the file names. If the full path names exceed 100 characters, the user may use abbreviated paths (location of the files relative to the *parameter* file), but must always update these paths if the *parameter* file is moved.

Forward slashes (/) are used in path names on Unix systems; and backslashes (\) are used in path names on PC systems.

In addition to the input files discussed above, there are three diagnostic output files and a set of final output files (i.e., one file for each state included in the simulation). The diagnostic output files are:

- The *log* file;
- The *counter* file; and
- The *mistract* file.

Full path names must be specified for these files.

As explained in Chapter 2, HAPEM6 creates an exposure output file for each state/pollutant combination. The names of these files are constructed by the program based on the pollutant SAROAD code, specified as the value of the ***sarod*** parameter, and the state FIPS code. Thus, the pathname, but not the filenames, for these files must be specified in the *parameter* file.

3.1.2. Identifying the Uniform Component of the Background Concentration

In addition to estimating exposure concentration contributions for each emission source category for which data are provided in the *air quality* file, HAPEM also estimates the exposure concentration contribution from the background outdoor concentration. This background exposure contribution is also added together with the emission source category contributions to calculate the total exposure concentration. One component of the background concentration is assumed uniform throughout the study area, i.e., a single value is specified as the ***backg*** parameter. The units of measurement must be the same as those used in the *air quality* file. The uniform component of the background concentration is an estimate of the outdoor concentration that would occur in the absence of any local anthropogenic emissions. It includes concentration contributions from natural sources, re-entrainment, or global transport. A second component of the background concentration is provided in the *air quality* file, as a single time-invariant value for each location specified in the *air quality* file. This component typically represents either the impact of anthropogenic emissions outside of the modeling domain or a combination of those emissions and the outdoor concentration that would occur in the absence of all anthropogenic emissions. In the latter case, the value of ***backg*** would be set to 0.0, since its constituents would be included in the location-specific background value.

3.1.3. Setting the Internal Parameters

The twelve internal parameter settings (***nmicro***, ***nblock***, ***hblock, ntype***, ***ngroup***, ***nsource***, ***nreplic***, ***nmobiles, nbmicro***, ***nemicro***, ***nvehicles***, and ***npublict***) are specified by the user in one or more of the *parameter* files and must be consistent with the structure of the other input data files. Each of these parameters are defined in the adjacent text box. Thus, if the user wishes to change the number of microenvironments, for example, the input files that specify microenvironments must also be altered in a consistent manner.

Internal Parameters

nmicro	number of microenvironments in the *activity* and *factors* files
nblock	number of time blocks per day in the *activity* file
hblock	number of time blocks per day for the analysis
ntype	number of day types in the DURAV source code
ngroup	number of demographic groups in the DURAV source code and the *population* file
nsource	number of emission source categories in the *air quality* file
nreplic	number of replicates to be simulated for each demographic group in each census tract
nmobiles	sequence numbers of up to 10 onroad mobile emission source categories in the *air quality* file
nbmicro	sequence number of the first indoor microenvironment (including in-vehicle) in the*factors* and *mobiles* files
nemicro	sequence number of the last indoor microenvironment (including in-vehicle) in the*factors* and *mobiles* files
nvehicles	sequence numbers of up to 10 microenvironments for private commuting in the*factors* and *mobiles* files (e.g., cars, trucks)
npublict	sequence numbers of up to 10 microenvironments for pubic transit commuting in the*factors* and *mobiles* files (e.g., busses, trains)

As explained in Chapter 2, the value of the ***hblock*** parameter, the number of time blocks per day for the analysis, must be selected to meet the following criteria.

- The value of ***hblock*** must be an integral factor of ***nblock***, the number of time blocks per day in the *activity* file, so that the activity time blocks can be combined if necessary to match to match ***hblock.***
- The value of ***hblock*** must be an

integral multiple of the number of time blocks per day in the *air quality* file, so that the air quality values can be replicated if necessary to create ***hblock*** air quality values.

3.2. Activity File

The *activity* file, the primary input to the DURAV program, contains information on the time spent in various microenvironments by individuals. This information is not presented as an activity sequence; rather it is presented in an *activity* file as the total time spent in each microenvironment during each block of time and at each location throughout the day.

3.2.1. Variables and Format of the Default File

The first line of the *activity* file is a text header that indicates the order of the variables in each record. The header in the default *activity* file, *durhw.txt*, is as follows.

Header of default *activity* file (in “wrapped” view)

```
CHADID ZIP DAYTYPE STATE COUNTY GENDER RACE EMPLOYED YEAR MONTH DAY AGE COMMUTE
DURATION (MICRO,BLOCK,HW) (NMICRO=14 NBLOCK=24 HW=2 IN FORTRAN ORDER)
```

Although most of the header record of the *activity* file is not used by the HAPEM6 programs, it provides documentation to inform the user of the meaning of the data fields. The exception is the specification of the number of time blocks per day, *nblock*, which the DURAV program checks against the value of the ***nblock*** parameter specified in the *parameter* file for consistency. If inconsistent, an error message is sent to the *log* file and the program stops.

Each space-delimited record following the header record consists of one person-day (1,440 minutes) of activity data. The variables in the default *activity* file, extracted from CHAD, are defined in Table 3-1.

Following the commuting indicator is a series of duration values. The values specify the integral number of minutes (possibly zero) spent in each microenvironment/time block/location combination, where locations are the vicinity of either home and work. The default *activity* file, *durhw.txt*, with 14 microenvironments (listed in Table 1-1), 24 time blocks per day, and two locations has a total of 672 duration values. These values are sequenced so that the 14 microenvironment durations for the first time block in the home location come first, followed by the 14 microenvironment durations for the second time block in the home location, and so on, until all the 336 values for the home location are specified. These are followed by the 336 values for the work location. An example of a record from *durhw.txt*. is presented below.

Table 3-1.
Variables in the default *activity* file

CHADID	9-character string identifying the data record; e.g., the corresponding person-day in the CHAD activity database. This information is used by the DURAV program only to identify faulty records in the diagnostic output files.
ZIP	5-character string identifying the zip code of respondent's residence. If a ZIP code is missing, it is reported as "00000 ". This information is not used by the current version of the DURAV program.
DAYTYPE	integer indicator of day type for classification, with values as follows: 1 = summer weekday 2 = non-summer weekday 3 = weekend
STATE	2-character string identifying the FIPS code of the state where the activities took place. This information is not used by the current version of the DURAV program.
COUNTY	3-character string identifying the FIPS code of the county where the activities took place. This information is not used by the current version of the DURAV program.
GENDER	1-character string, indicating gender, used to assign the respondent to a demographic group, with values as follows: "1" = female "2" = male "9" = unknown
RACE	1-character string, indicating race/ethnic group, used to assign the respondent to a demographic group, with values as follows: "1" = White (non-Hispanic) "2" = Black (non-Hispanic) "3" = Hispanic (any race) "4" = Asian or Other (non-Hispanic) "9" = unknown This information is not used by the current version of the DURAV program.

(continued)

Table 3-1.
Variables in the default *activity* file

EMPLOYED	1-character string, indicating employment status of respondent, with values as follows "Y" = Yes "N" = No "X" = missing This information is not used by the current version of the DURAV program.
YEAR, MONTH, DAY	numeric variables (four-digit year) that identify the date when the activities actually took place. This information is not used by the current version of the DURAV program.
AGE	integer indicator of the age of the subject (missing = -999.00)
COMMUTE	integer indicator of whether the respondent is a commuter, with values as follows: 0 = no commuting 1 = commuting

(concluded)

Example data record from default *activity* file (in “wrapped” view)

```
BAL97001A 21204  2  24 000 1 1 N 1997   1 21   77.00 0  60   0    0    0    0    0    0    0
0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0   60    0
0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0
0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0
0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0
0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0
0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0
30   0    0    0    0   30    0    0    0    0    0    0    0    0   45    0    0    0    0   15    0    0
0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0   60    0
0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0
0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0
0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0
0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0
0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0
60   0    0    0    0    0    0    0    0    0    0    0    0    0   60    0    0    0    0    0    0    0
0    0    0    0    0    0   60    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
0    0    0    0
```

3.2.2. Replacing or Modifying the Default File

If the user wishes to replace or modify the default *activity* file, he or she must ensure that the following two conditions are met.

- The number of duration values in each record must equal twice the product of the values of ***nmicro*** and ***nblock*** as specified in the *parameter* file.
- The sum of the duration values in each record must total 1,440 minutes (*i.e.*, no time is unaccounted); otherwise the DURAV program will stop.

In addition, the user must ensure that the *activity* file is consistent with several features of the DURAV source code. First, the record length of the *activity* file (unit 11) and two files derived from it (units 20 and 21) are specified in the DURAV program. Unit 21 is also used as input to the HAPEM program, where it's record length is again specified. If the user constructs a replacement activity file with a record length different from that of the default *activity* file, corresponding changes need to be made in both DURAV and HAPEM.

The variables used by the DURAV program for classifying activity records (i.e., DAYTYPE, AGE, and COMMUTE), as well as the activity duration values, are identified by the program by their position in the data record. If the user constructs a replacement activity file with these variables positioned differently, corresponding changes need to be made in DURAV.

The definitions of the demographic groups (presented in Section 2.2.2) are part of the DURAV source code, so that in order to change the demographic group definitions the source code must be modified. Similarly, the definitions of day types for categorizing activity patterns, presented above, are part of the DURAV source code. The number of demographic groups and day types is unlimited. However, the user is cautioned that for narrowly defined groups and day types, there might not be enough activity pattern data to calculate a reliable group average or create meaningful activity pattern clusters. An extreme example of this is where no activity patterns fit a demographic group's definition, resulting in incorrect exposure calculations (*i.e.*, exposure concentrations equal to zero) for that group.

The number, definition, and order of microenvironments must be the same in both the *activity* file and the *factors* and *mobiles* files (see Section 3.10). The number is specified in the *parameter* files as ***nmicro***.

The *activity* file is read by the DURAV program, which creates several intermediate output files with the same path and root file name, but with different filename extensions. Thus, the user should NOT name an *activity* file with any of the following filename extensions: "*.da"*, *"wrong_chad"*, and *".nonzero"*. There is also a file created with the root name of the *activity* file and the extension "*.draft"* that is used internally by the DURAV program but deleted at the end of the DURAV run.

As with other HAPEM6 input files, the user can add comments or other information after the last data record in the file. To prevent the program reading these comments as data, a blank line must be inserted after the last data record and before any comments.

3.3. Cluster File

This file provides information on demographic type, day type, cluster type of each complete (i.e., with 1440 minutes a day) CHAD record in *activity* file. The file is used in DURAV to group CHAD records according to cluster.

3.3.1. Variables and Format of the Default File

The first line of the *cluster* file is a text header that indicates the order of the variables in each record. The header in the default *cluster* file, *durhw_cluster.txt*, is as follows.

Header of default *cluster* file

```
CHADID   Demographic    DayType    "Comtype(1=non-commute,"   CLUSTER    Ncluster
```

Although the header record of the *cluster* file is not used by the HAPEM6 programs, it provides documentation to inform the user of the meaning of the data fields. "CLUSTER" refers to the cluster category number for that record, and "Ncluster" refers to the total number of cluster categories for that demographic-group/day-type/commuting-status combination.

An extract from the default *cluster* file is shown below. These cluster categories were determined using cluster analysis, as explained in Appendix A.[15]

3.3.2. Replacing or Modifying the Default File

If the user wishes to replace or modify the default *cluster* file, he or she must ensure that the file is properly formatted and the following two conditions are met.

- There should be one record for every valid record in the corresponding *activity* file (i.e., one with 1440 minutes, an age specification, and a day type designation of 1-3, and a commuting status specification). Any record in the *activity* file without a corresponding record in the *cluster* file will not be used.
- The records should be sorted by demographic group, day type, commuting status, and cluster.

[15] See footnote 4 above.

Extract from Default *cluster* File

```
CHADID   Demographic   DayType   "Comtype(1=non-commute,"   CLUSTER    Ncluster
NHW18890A    5      1      1      2      2
NHW18892A    5      1      1      2      2
NHW18915A    5      1      1      2      2
NHW18927A    5      1      1      2      2
NHW19058A    5      1      1      2      2
NHW19079A    5      1      1      2      2
NHW19131A    5      1      1      2      2
NHW19139A    5      1      1      2      2
NHW19207A    5      1      1      2      2
NHW19218A    5      1      1      2      2
NHW19245A    5      1      1      2      2
NHW19259A    5      1      1      2      2
NHW19332A    5      1      1      2      2
NHW19341A    5      1      1      2      2
NHW19343A    5      1      1      2      2
NHW19383A    5      1      1      2      2
CAA08971A    5      1      2      1      3
CAA10771A    5      1      2      1      3
CAA15631A    5      1      2      1      3
CAA29341A    5      1      2      1      3
CAA40691A    5      1      2      1      3
CAA41021A    5      1      2      1      3
CAA41031A    5      1      2      1      3
CAA41231A    5      1      2      1      3
CAA41361A    5      1      2      1      3
CAA41651A    5      1      2      1      3
CAA43491A    5      1      2      1      3
CAA43651A    5      1      2      1      3
CAA45011A    5      1      2      1      3
CAA45251A    5      1      2      1      3
CAA45401A    5      1      2      1      3
CAA45451A    5      1      2      1      3
CAA46201A    5      1      2      1      3
CAA46701A    5      1      2      1      3
CAA47271A    5      1      2      1      3
CAA47341A    5      1      2      1      3
CAA47781A    5      1      2      1      3
CAA47921A    5      1      2      1      3
CAA48221A    5      1      2      1      3
CAA48271A    5      1      2      1      3
CAA48281A    5      1      2      1      3
CAA48321A    5      1      2      1      3
CAA48391A    5      1      2      1      3
CAA48431A    5      1      2      1      3
CAA48501A    5      1      2      1      3
CAA48811A    5      1      2      1      3
CAA48891A    5      1      2      1      3
CAA48901A    5      1      2      1      3
CAA48941A    5      1      2      1      3
```

3.4. Population File

The *population* file, the primary input to the INDEXPOP program, provides the number of people in each demographic group residing in each census tract of the study area. The data must be sorted according to state FIPS, county FIPS, and tract code. The data are typically derived from the US Census data. The demographic group definitions are defined in the DURAV source code, and presented in Section 2.2.2.

3.4.1. Variables and Format of the Default File

The *population* file begins with two text header records, followed by one data record for each census tract. The first header record indicates the order of the variables in each of the data records. The first header record of the default *population* file is as follows.

First header record from the default *population* file

```
   TRACT           B_00    B_02    B_05    B_16    B_18    B_65
```

Although the header records of the *population* file are not used by the HAPEM6 programs, the first one provides documentation to inform the user of the meaning of the data fields. Each space-delimited data record following the header records consists of a census tract identifier and a population value for each of the indicated demographic groups in that tract. The definitions of the data fields in the default *population* file is presented in Table 3-2.

An extract from the default *population* file is presented below.

3.4.2. Replacing or Modifying the Default File

If the user wishes to replace or modify the default *population* file, he or she must ensure that the definitions and ordering of the demographic groups in the *population* file corresponds to the ordering in the output file from DURAV that is subsequently used in the HAPEM program.

In addition, the user must ensure that the record length is consistent with its specification in the INDEXPOP program (unit 14).

As noted elsewhere, the definitions of the demographic groups (presented in Section 2.2.2) are part of the DURAV source code, so that in order to change the demographic group definitions the source code must be modified.

The *population* file is read by the INDEXPOP program, which creates several intermediate output files with the same path and root file name, but with different filename extensions. Thus, the user should NOT name a *population* file with any of the following filename extensions: "*.da*",

".county_tract_pop_range", and *".state_county_pop_range"*. There is also an intermediate file with the characters "*_direct.ind"* attached to the *population* file root name.

As with other HAPEM6 input files, the user can add comments or other information after the last data record in the file. To prevent the program reading these comments as data, a blank line must be inserted after the last data record and before any comments.

Table 3-2.
Variables in the default *population* file

TRACT	11-character string uniquely identifying a US Census tract. The first two characters identify the state FIPS code, the next three characters the county FIPS code. The remaining 6 characters consist of the 4-digit tract code followed by its 2-digit extension. If there is no extension for the tract, "00" is used.
B_YY	Integer specifying the number of 2000 tract residents with age in category YY. The age category definitions are: 00 = 0-1 years old 02 = 2-4 years old 05 = 5-15 years old 16 = 16-17 years old 18 = 18-64 years old 65 = 65 years or older

Extract from Default *population* File

```
TRACT           B_00    B_02    B_05    B_16    B_18    B_65
                COM     COM     COM     COM     COM     COM
01001020100     44      50      357     68      1225    176
01001020200     58      94      321     57      1148    214
01001020300     98      147     609     106     1924    453
01001020400     104     145     740     134     2730    702
01001020500     166     280     1237    184     3751    422
01001020600     79      146     691     109     2045    308
01001020700     90      124     462     96      1815    313
01001020800     276     381     1733    324     5939    704
01001020900     124     205     873     171     2787    468
01001021000     78      109     484     89      1584    331
01001021100     99      126     529     97      1778    352
01003010100     104     177     683     126     2502    567
01003010200     61      105     454     85      1637    292
01003010300     183     280     1071    182     4051    744
01003010400     121     188     818     138     2759    418
01003010500     106     127     586     106     2844    796
01003010600     163     246     754     134     2107    320
01003010701     140     249     1158    181     3748    1033
01003010703     141     252     1109    158     3245    399
01003010704     130     218     773     169     3090    452
01003010705     163     225     889     156     4012    557
01003010800     118     202     1059    220     3756    898
01003010901     267     363     1618    287     6172    938
01003010902     259     414     1659    301     5494    1081
01003011000     131     202     708     114     2465    475
01003011100     187     306     1231    231     4295    1273
01003011201     61      93      543     124     2276    1088
01003011202     102     172     703     158     3052    1214
01003011300     91      123     536     101     2278    536
01003011401     194     246     1128    171     4265    906
```

3.5. Commuting Time File

The *CommutTime* file provides data for each tract on the proportion of commuting workers who take public transit and private transit, and their respective round-trip average commuting times (minutes). This information is combined with data on the centroid-to-centroid commuting distances for workers in the tract, provided in the *commuting* file described below, to estimate a commuting time for each replicate that is probabilistically selected to commute to work, according to the data provided in the *CommutFrac* file described below. The HAPEM program then adjusts the selected activity patterns for that replicate to reflect the estimated commuting time. (See Section 5.2.5 for more details on the algorithm.)

The data in the default *CommutTime* file are derived from the 2000 Census (P32).

The *CommutTime* file has no header records, only data records. Each space-delimited data record contains five variables, as follows.

- Tract ID (11-character string: state FIPS, county FIPS, and tract code)
- Proportion of commuters who travel by public transit (decimal number)
- Proportion of commuters who travel by private car or truck (decimal number)
- Average round-trip commuting time for public transit commuters
- Average round-trip commuting time for private transit commuters

The default *CommutTime* file is sorted by tract ID, smallest to largest in numerical order. Several example data records from the default *CommutTime* file are presented below.

Extract from the default *CommutTIme* file

```
 01001020100   0.0000   1.0000   0.0000   55.3124
 01001020200   0.0000   1.0000   0.0000   52.9643
 01001020300   0.0000   1.0000   0.0000   51.3188
01001020400   0.0052   0.9948  180.0000   47.4427
 01001020500   0.0000   1.0000   0.0000   51.7882
01001020600   0.0156   0.9844   30.0000   45.1287
 01001020700   0.0000   1.0000   0.0000   48.5694
 01001020800   0.0000   1.0000   0.0000   61.0869
 01001020900   0.0000   1.0000   0.0000   75.0188
 01001021000   0.0000   1.0000   0.0000   87.1034
01001021100   0.0008   0.9992   30.0000   56.4098
01003010100   0.0090   0.9910   74.0000   82.5042
01003010200   0.0068   0.9932  180.0000   65.7579
01003010300   0.0045   0.9955  104.0000   60.7361
01003010400   0.0031   0.9969   74.0000   67.2765
01003010500   0.0114   0.9886   30.0000   58.6531
01003010600   0.0101   0.9899   52.0000   54.2015
 01003010701   0.0000   1.0000   0.0000   53.1494
01003010703   0.0055   0.9945   50.3077   56.8456
01003010704   0.0026   0.9974   74.0000   54.1121
 01003010705   0.0000   1.0000   0.0000   52.6994
01003010800   0.0023   0.9977  180.0000   52.7853
01003010901   0.0056   0.9944   50.4545   61.3040
01003010902   0.0002   0.9998   30.0000   55.3814
01003011000   0.0055   0.9945   30.0000   62.6955
 01003011100   0.0000   1.0000   0.0000   55.0802
 01003011201   0.0000   1.0000   0.0000   49.8336
```

3.6. Commuting Fraction File

The *CommutFrac* file provides data for each tract on the proportion of each demographic group that commutes to work. This information is used by the HAPEM program to determine for each replicate in each demographic group whether they commute to work, and therefore, which set of activity patterns should be sampled to represent that replicate. The data in the default *CommutFrac* file are derived from the 2000 Census (P31 and PCT35).

The *CommutFrac* file has no header records, only data records. Each space-delimited data record contains thirteen variables, as follows.

- Tract ID (11-character string: state FIPS, county FIPS, and tract code)
- Proportion of demographic group 1 that does not commute to work (decimal number)
- Proportion of demographic group 1 that commutes to work (decimal number)
- Proportion of demographic group 2 that does not commute to work (decimal number)
- Proportion of demographic group 2 that commutes to work (decimal number)
- Proportion of demographic group 3 that does not commute to work (decimal number)
- Proportion of demographic group 3 that commutes to work (decimal number)
- Proportion of demographic group 4 that does not commute to work (decimal number)
- Proportion of demographic group 4 that commutes to work (decimal number)
- Proportion of demographic group 5 that does not commute to work (decimal number)
- Proportion of demographic group 5 that commutes to work (decimal number)
- Proportion of demographic group 6 that does not commute to work (decimal number)
- Proportion of demographic group 6 that commutes to work (decimal number)

The default *CommutFrac* file is sorted by tract ID, smallest to largest in numerical order. Several example data records from the default *CommutFrac* file are presented below.

Extract from the default *CommutFrac* file

```
01001020100  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5269  0.4731  0.3939  0.6061  0.7357  0.2643
01001020200  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5066  0.4934  0.4604  0.5396  0.8221  0.1779
01001020300  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.3787  0.6213  0.2992  0.7008  0.9644  0.0356
01001020400  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5981  0.4019  0.2902  0.7098  0.8606  0.1394
01001020500  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5785  0.4215  0.2906  0.7094  0.9011  0.0989
01001020600  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.4548  0.5452  0.2764  0.7236  0.9001  0.0999
01001020700  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7413  0.2587  0.2811  0.7189  0.6625  0.3375
01001020800  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6627  0.3373  0.2903  0.7097  0.8778  0.1222
01001020900  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7037  0.2963  0.3835  0.6165  0.9243  0.0757
01001021000  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7633  0.2367  0.3976  0.6024  0.9028  0.0972
01001021100  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5131  0.4869  0.3597  0.6403  0.9124  0.0876
01003010100  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7768  0.2232  0.4041  0.5959  0.9184  0.0816
01003010200  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6186  0.3814  0.3114  0.6886  0.9444  0.0556
01003010300  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6186  0.3814  0.3080  0.6920  0.8636  0.1364
01003010400  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5518  0.4482  0.3279  0.6721  0.8498  0.1502
01003010500  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7235  0.2765  0.3925  0.6075  0.8931  0.1069
01003010600  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6021  0.3979  0.3740  0.6260  0.8730  0.1270
01003010701  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5755  0.4245  0.2642  0.7358  0.9148  0.0852
01003010703  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5857  0.4143  0.3144  0.6856  0.8175  0.1825
01003010704  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5475  0.4525  0.2769  0.7231  0.8883  0.1117
01003010705  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6255  0.3745  0.2506  0.7494  0.8894  0.1106
01003010800  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7060  0.2940  0.3475  0.6525  0.9263  0.0737
01003010901  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6148  0.3852  0.3967  0.6033  0.8757  0.1243
01003010902  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6163  0.3837  0.2879  0.7121  0.8610  0.1390
01003011000  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5833  0.4167  0.3126  0.6874  0.8188  0.1812
01003011100  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6201  0.3799  0.3658  0.6342  0.9156  0.0844
01003011201  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5989  0.4011  0.2654  0.7346  0.9206  0.0794
01003011202  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.5966  0.4034  0.3243  0.6757  0.9367  0.0633
01003011300  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7399  0.2601  0.3186  0.6814  0.8405  0.1595
01003011401  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7440  0.2560  0.3542  0.6458  0.9067  0.0933
01003011403  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.4799  0.5201  0.3332  0.6668  0.8501  0.1499
01003011404  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.4438  0.5562  0.3155  0.6845  0.8943  0.1057
01003011500  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.4195  0.5805  0.2871  0.7129  0.9045  0.0955
01003011600  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.4604  0.5396  0.4045  0.5955  0.9149  0.0851
01005950100  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6552  0.3448  0.3031  0.6969  0.9165  0.0835
01005950200  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.8727  0.1273  0.6816  0.3184  0.8778  0.1222
01005950300  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.8710  0.1290  0.5227  0.4773  0.9066  0.0934
01005950400  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.8195  0.1805  0.6508  0.3492  0.9081  0.0919
01005950500  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6960  0.3040  0.3870  0.6130  0.9093  0.0907
01005950600  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.6863  0.3137  0.3207  0.6793  0.9555  0.0445
01005950700  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7462  0.2538  0.4811  0.5189  0.9433  0.0567
01005950800  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.7865  0.2135  0.3322  0.6678  0.8159  0.1841
01005950900  1.0000  0.0000  1.0000  0.0000  1.0000  0.0000  0.8597  0.1403  0.3986  0.6014  0.7908  0.2092
```

3.7. Distance to Road File

The *DistToRoad* file provides data for each tract on the tract area and the proportion of each demographic group that resides within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, and greater than 200 meters. This information is used by the HAPEM program to determine for each replicate for each ME the distance from a major roadway, and therefore, which PROX factor distributions in the *mobiles* file, described below, to sample from for the onroad mobile source categories.

The data in the default *DistToRoad* file are derived from the Environmental Sciences Research Center (ESRI) StreetMap US roadway geographic database and a geographic database of US Census block boundaries. In developing this data base a “major roadway” was defined as a “Limited Access Highway”, “Highway”, “Major Road” or “Ramp”, according to the Census Feature Class Codes (CFCC). Because Census blocks in some locations can be large and have non-uniform population density, in order to estimate population locations more precisely, it was assumed that all residences are located within 150 meters of some roadway, whether a major roadway or a local roadway. (See Appendix B for more details.)

The *DistToRoad* file has no header records, only data records. Each space-delimited data record contains twenty-two variables, as follows.

- Tract ID (11-character string: state FIPS, county FIPS, and tract code)
- Fractions of tract area within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)
- Fractions of demographic group 1 that reside within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)
- Fractions of demographic group 2 that reside within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)
- Fractions of demographic group 3 that reside within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)
- Fractions of demographic group 4 that reside within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)
- Fractions of demographic group 5 that reside within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)
- Fractions of demographic group 6 that reside within 3 distance categories from a major roadway: 0-75 meters, 75-200 meters, greater than 200 meters (3 decimal numbers)

The default *DistToRoad* file is sorted by tract ID, smallest to largest in numerical order. Several example data records from the default *DistToRoad* file are presented below.

Extract from the default *DistToRoad* file (in “wrapped” view)

```
01001020100   0.3268  0.2578  0.4154  0.6570  0.3318  0.0112  0.7354  0.2496  0.0150  0.6920  0.2961
       0.0119  0.7101  0.2791  0.0108  0.6966  0.2893  0.0141  0.6536  0.3305  0.0159
01001020200   0.6165  0.2142  0.1693  0.7997  0.2003  0.0000  0.7731  0.2269  0.0000  0.8016  0.1984
       0.0000  0.8392  0.1608  0.0000  0.8114  0.1886  0.0000  0.8255  0.1745  0.0000
01001020300   0.5894  0.2725  0.1381  0.7724  0.2226  0.0050  0.7500  0.2437  0.0063  0.7327  0.2612
       0.0061  0.7387  0.2547  0.0066  0.7540  0.2389  0.0071  0.7725  0.2225  0.0050
01001020400   0.5822  0.2046  0.2132  0.8608  0.1334  0.0058  0.8584  0.1360  0.0056  0.8529  0.1412
       0.0059  0.8116  0.1824  0.0060  0.8472  0.1479  0.0049  0.8716  0.1242  0.0042
01001020500   0.4175  0.2381  0.3444  0.6969  0.2515  0.0516  0.7122  0.2375  0.0503  0.7311  0.2251
       0.0438  0.7395  0.2188  0.0417  0.7227  0.2326  0.0447  0.7036  0.2415  0.0549
01001020600   0.5169  0.2952  0.1879  0.7351  0.2496  0.0153  0.7545  0.2289  0.0166  0.7337  0.2486
       0.0177  0.7091  0.2736  0.0173  0.7239  0.2551  0.0210  0.7283  0.2518  0.0199
01001020700   0.2132  0.1680  0.6188  0.8004  0.1851  0.0145  0.7538  0.2241  0.0221  0.7653  0.2290
       0.0057  0.7510  0.2340  0.0150  0.7703  0.2201  0.0096  0.7529  0.2401  0.0070
01001020800   0.1574  0.1451  0.6975  0.5187  0.3968  0.0845  0.5197  0.3904  0.0899  0.5385  0.3860
       0.0755  0.5344  0.3919  0.0737  0.5279  0.3934  0.0787  0.5011  0.4131  0.0858
01001020900   0.1195  0.1210  0.7595  0.4542  0.4101  0.1357  0.4430  0.4336  0.1234  0.4169  0.4123
       0.1708  0.4469  0.4053  0.1478  0.4218  0.4060  0.1722  0.4330  0.4073  0.1597
01001021000   0.1354  0.1252  0.7394  0.4855  0.4396  0.0749  0.4864  0.4420  0.0716  0.4833  0.4390
       0.0777  0.4907  0.4493  0.0600  0.4894  0.4367  0.0739  0.5102  0.4086  0.0812
01001021100   0.1329  0.1262  0.7409  0.5202  0.4138  0.0660  0.5107  0.4208  0.0685  0.5284  0.3901
       0.0815  0.5084  0.3931  0.0985  0.5183  0.4089  0.0728  0.5235  0.3782  0.0983
01003010100   0.0344  0.0378  0.9278  0.1082  0.1297  0.7621  0.1148  0.1341  0.7511  0.1121  0.1377
       0.7502  0.1210  0.1472  0.7318  0.1224  0.1503  0.7273  0.1234  0.1649  0.7117
01003010200   0.0414  0.0487  0.9099  0.1819  0.2169  0.6012  0.1848  0.2070  0.6082  0.1744  0.2163
       0.6093  0.1642  0.2105  0.6253  0.1752  0.2098  0.6150  0.1926  0.2207  0.5867
01003010300   0.0311  0.0396  0.9293  0.1112  0.1606  0.7282  0.0976  0.1549  0.7475  0.1079  0.1558
       0.7363  0.1074  0.1457  0.7469  0.1058  0.1615  0.7327  0.1281  0.1707  0.7012
01003010400   0.0384  0.0416  0.9200  0.1238  0.1631  0.7131  0.1405  0.1845  0.6750  0.1314  0.1785
       0.6901  0.1048  0.1389  0.7563  0.1396  0.1806  0.6798  0.1350  0.1779  0.6871
01003010500   0.1396  0.2045  0.6559  0.2395  0.3845  0.3760  0.2227  0.3501  0.4272  0.2453  0.3598
       0.3949  0.2231  0.3688  0.4081  0.3157  0.3799  0.3044  0.2291  0.4040  0.3669
01003010600   0.0938  0.1396  0.7666  0.2459  0.2884  0.4657  0.2849  0.2869  0.4282  0.2505  0.2947
       0.4548  0.2202  0.2608  0.5190  0.2336  0.2794  0.4870  0.1948  0.3221  0.4831
01003010701   0.1497  0.1694  0.6809  0.1370  0.2566  0.6064  0.1365  0.2176  0.6459  0.1388  0.2278
       0.6334  0.1515  0.2127  0.6358  0.1535  0.2310  0.6155  0.1727  0.2555  0.5718
01003010703   0.0625  0.0689  0.8686  0.1321  0.2078  0.6601  0.1268  0.1939  0.6793  0.1202  0.1874
       0.6924  0.1172  0.1883  0.6945  0.1243  0.1868  0.6889  0.1308  0.1740  0.6952
01003010704   0.2935  0.3197  0.3868  0.1445  0.5054  0.3501  0.1990  0.4766  0.3244  0.2165  0.4518
       0.3317  0.2338  0.4464  0.3198  0.1912  0.4686  0.3402  0.2152  0.4605  0.3243
01003010705   0.0833  0.1021  0.8146  0.0125  0.0901  0.8974  0.0326  0.1024  0.8650  0.0205  0.0784
       0.9011  0.0187  0.0901  0.8912  0.0362  0.1153  0.8485  0.0442  0.1107  0.8451
01003010800   0.1302  0.1924  0.6774  0.1112  0.1878  0.7010  0.1220  0.1879  0.6901  0.1039  0.2026
       0.6935  0.1235  0.2171  0.6594  0.1277  0.2147  0.6576  0.1184  0.2142  0.6674
01003010901   0.0488  0.0582  0.8930  0.1736  0.2180  0.6084  0.1776  0.2087  0.6137  0.1642  0.2092
       0.6266  0.1499  0.1933  0.6568  0.1574  0.2010  0.6416  0.1633  0.2201  0.6166
```

3.8. Commuting Flow File

The *commuting* file, the main input file to the COMMUTE program, provides data on the commuting flows (*i.e.*, the number of commuters) between pairs of census tracts. The default *commuting* file was derived from 2000 U.S. Census Bureau data identifying the tract of work and tract of residence for individuals in all fifty states and the District of Columbia. Although there are approximately 500 million pairs of tracts nationwide within a reasonable commuting distance of each other, only about 5 million of these pairs have a non-zero flow of commuters. Only those pairs with non-zero flows are included in the default *commuting* file.

The *commuting* file has no header records, only data records. Each space-delimited data record contains four variables, as follows.

- Home tract ID (11-character string: state FIPS, county FIPS, and tract code)
- Work tract ID (11-character string: state FIPS, county FIPS, and tract code)
- Distance apart in kilometers (decimal number)
- Fraction of workers in the commuting flow (decimal)

The default *commuting* file is sorted by home tract ID, smallest to largest in numerical order. Several example data records from the default *commuting* file are presented below.

The *commuting* file is read by the COMMUTE program, which creates several intermediate output files with the same path and root file name, but with different filename extensions. Thus, the user should NOT name a *commuting* file with any of the following filename extensions: "*.da*", "*.ind*", and "*.st_comm1_fip_range*".

As with other HAPEM6 input files, the user can add comments or other information after the last data record in the file. To prevent the program reading these comments as data, a blank line must be inserted after the last data record and before any comments.

Extract from the default *commuting* file

```
01001020100 01001020100    0.00 0.02896871
01001020100 01001020200    1.40 0.06952491
01001020100 01001020300    2.80 0.04866744
01001020100 01001020400    3.90 0.04287370
01001020100 01001020500    6.10 0.07647740
01001020100 01001020600    3.50 0.08342990
01001020100 01001020700    5.50 0.09965237
01001020100 01001020800    3.60 0.01853998
01001020100 01001020900   19.30 0.00695249
01001020100 01021060101   41.30 0.00347625
01001020100 01051031300   10.60 0.00463499
01001020100 01081041300  104.90 0.00579374
01001020100 01085999999 9999.00 0.00811124
01001020100 01101000100   19.60 0.09269989
01001020100 01101000200   20.10 0.01506373
01001020100 01101000600   21.50 0.00579374
01001020100 01101000700   20.70 0.01969873
01001020100 01101000900   15.40 0.03823870
01001020100 01101001100   19.20 0.00926999
01001020100 01101001500   22.60 0.00695249
01001020100 01101001600   22.60 0.00347625
01001020100 01101001800   22.70 0.01274623
01001020100 01101001900   23.60 0.00579374
01001020100 01101002000   25.00 0.00811124
01001020100 01101002500   20.90 0.00926999
01001020100 01101002700   26.40 0.01853998
01001020100 01101002800   26.40 0.02201622
01001020100 01101003000   17.70 0.00926999
01001020100 01101005101   24.00 0.01390498
01001020100 01101005301   23.60 0.01853998
01001020100 01101005302   26.60 0.00695249
01001020100 01101005401   32.20 0.00811124
01001020100 01101005402   28.30 0.03707995
01001020100 01101005406   32.90 0.00926999
01001020100 01101005901   29.50 0.02317497
01001020100 01101006000   13.50 0.03012746
```

3.9. Air Quality File

The *air quality* file contains the ambient air concentrations that are used by the AIRQUAL program. AIRQUAL requires a separate *air quality* file for each pollutant being evaluated.

The *air quality* file must begin with at least one text header record, followed by one or more data records for each census tract to be evaluated. The required text header is used by the

AIRQUAL program to determine the number of time blocks per day (of equal size) in the air quality data. This value should be indicated immediately following the last instance of the character string “block”. For example, the sixth header record of the ASPEN-derived *air quality* files used for the recent NATA analysis, which indicates the order of the variables in each of the data records, is as follows.

Example header record from an *air quality* File (in “wrapped” view)

```
FIPS   Tract  Backgrd_Conc   Conc_block1   Conc_block2   Conc_block3
Conc_block4   Conc_block5   Conc_block6   Conc_block7   Conc_block8
Conc_block1   Conc_block2   Conc_block3   Conc_block4   Conc_block5
Conc_block6   Conc_block7   Conc_block8   Conc_block1   Conc_block2
Conc_block3   Conc_block4   Conc_block5   Conc_block6   Conc_block7
Conc_block8   Conc_block1   Conc_block2   Conc_block3   Conc_block4
Conc_block5   Conc_block6   Conc_block7   Conc_block8   Conc_block1
Conc_block2   Conc_block3   Conc_block4   Conc_block5   Conc_block6
Conc_block7   Conc_block8
```

For this example, AIRQUAL will interpret the number of time blocks per day as 8. As noted elsewhere, the number of time blocks per day in the *air quality* file must be an integral factor of ***hblock***, the number of time blocks per day for the analysis as specified in the *parameter* file; otherwise the program will stop. If the number of time blocks per day in the air quality file is less than ***hblock***, AIRQUAL will replicate the values to create ***hblock*** concentration values.

The other information in this header record and all other header records is ignored by AIRQUAL.

After the required header information is found, AIRQUAL identifies data records by finding a numerical digit in the fourth data field. To avoid a mistaken identification, the user should insure that header records do NOT contain a numerical digit in the fourth data field.

The fields in the data records are defined as follows.

Columns 4-5	state FIPS code (2-character string)
Columns 6-8	county FIPS code (3-character string)
Columns 11-16	census tract ID (6-character string)
Columns 17- 30	space-delimited concentration contribution value for spatially-variable (but temporally constant) background concentration (decimal number, optionally in exponential format)

Columns 31-Last	space-delimited concentration contribution values for each emission source category/time block combination (decimal numbers, optionally in exponential format)

The number of non-background concentration values in each data record must equal the product of the number of outdoor emission source categories (i.e., the value of ***nsource*** in the *parameter* file), and the number of time blocks per day, as indicated in the text header record discussed above. The values are ordered beginning with the first time block of the first emission source, followed by the second time block of the first emission source, and so on. An example data record is presented below for ***nsource*** = 4.

Example data record from an *air quality* file (in "wrapped" view)

```
   48201  212100  0.442917E+00  0.423370E+00  0.184805E+01  0.779232E+00  0.670257E+00
0.961739E+00  0.139123E+01  0.118992E+01  0.102914E-05  0.154827E-05  0.640535E+00  0.461527E+00
0.347169E+00  0.405696E+00  0.189035E+00  0.201626E-01  0.448010E+00  0.509662E+00  0.491137E+00
0.212838E+00  0.163410E+00  0.221771E+00  0.625868E+00  0.698943E+00  0.131965E+00  0.156899E+00
0.202974E+00  0.884305E-01  0.679746E-01  0.827090E-01  0.139021E+00  0.160499E+00  0.249700E+00
```

The *air quality* file is read by the AIRQUAL program, which creates several intermediate output files with the same path and root file name, but with different filename extensions. Thus, the user should NOT name an *air quality* file with any of the following filename extensions: "*.da"*, *".air_da"*, *".pop_air_da"*, *".state_air_fip_range", ".state_air1_fip_range",* and *".state_air2_fip_range"*.

As with other HAPEM6 input files, the user can add comments or other information after the last data record in the file. To prevent the program reading these comments as data, a blank line must be inserted after the last data record and before any comments.

3.10. Microenvironmental Factors Files

The microenvironmental (ME) *factors* and *mobiles* files provide probability distributions for the factors used to calculate an estimated microenvironmental concentration from an outdoor concentration. The files contain probability distributions for three of the four factors for each microenvironment, and a single value for the fourth factor. These factors are used in the HAPEM algorithm, as follows.

$$ME(m,c,t,s,d) = PROX(m,s,d) \times PEN(m,t) \times AMB(c,t_{LAG(m)},s) \qquad (3\text{-}1a)$$

$$ME(m,t,i) = ADD(m,t) \qquad (3\text{-}1b)$$

$$ME(m,c,t,b) = PROX(m.,s) \times PEN(m,t) \times [bckgd\_u + bckgd\_v(c)] \qquad (3\text{-}1c)$$

$$ME(m,c,t.d) = \sum_{s} ME(m,c,t,s,d) + ME(m,t,i) + ME(m,c,t,b) \qquad (3\text{-}1d)$$

where:

ME(m,c,t,s,d): concentration in microenvironment *m* located in census tract *c* at time *t* due to source category *s* and at distance-from-source category *d*,

PROX(m,s,d): proximity factor for microenvironment *m*, source category *s,* and distance-from-source category *d* (defined below),

PEN(m,t): penetration factor for microenvironment *m* at time *t* (defined below),

AMB(c,t,s): ambient concentration for census tract *c* at time *t* for source category *s* from the *air quality* file,

$t_{LAG(m)}$: time *t* if *LAG(m)* = 0; time *t-1*, otherwise,

ME(m,t,i): concentration in microenvironment *m* at time *t* due to indoor sources,

ADD(m,t): additive factor for microenvironment *m* at time *t* (defined below),

ME(m,c,t,b): concentration in microenvironment *m* located in census tract *c* at time *t*, due to the background concentration,

Bckgd_u: uniform component of ambient background concentration,

Backgd_v(c): spatially variable component of background concentration, and

ME(m,c,t,d): total concentration in microenvironment *m* located in census tract *c* at time *t*, and distance-from-source category *d*.

The penetration factor, *PEN,* is an estimate of the ratio of the ME concentration contribution (from a given emission source category) to the concurrent outdoor concentration contribution in the immediate vicinity of the ME. That is,

$$PEN = \frac{\text{indoor or in - vehicle ME concentration}}{\text{outdoor concentration in immediate vicinity of indoor or in - vehicle ME}} \qquad (3\text{-}2)$$

The proximity factor, *PROX,* is an estimate of the ratio of the outdoor concentration in the immediate vicinity of the ME (or in the ME for outdoor MEs) to the outdoor concentration represented by the air quality data. That is,

$$PROX_I = \frac{\text{outdoor concentration in immediate vicinity of indoor or in - vehicle ME}}{\text{air quality file concentration}} \tag{3-3a}$$

$$PROX_O = \frac{\text{outdoor ME concentration}}{\text{air quality file concentration}} \tag{3-3b}$$

As explained elsewhere, the air quality data from ASPEN represent a spatial average over some portion of a census tract. For most MEs the default *factors* file specifies a *PROX* value of 1.0, i.e., an outdoor concentration contribution in the immediate vicinity of the ME equal to the spatial average contribution over that part of the census tract. However, when assessing exposure to motor vehicle emissions, for MEs near roadways (e.g., in-vehicle, indoor MEs situated near roadways) the pollutant concentration contribution in the immediate vicinity of the ME is expected to be higher than the spatial average pollutant concentration contribution over surrounding portion of the census tract, i.e., *PROX* is expected to be greater than 1.0. This is because the concentration gradient near roadways tends to be relatively steep. This condition for onroad mobile emissions is reflected in the default *mobiles* file, which contains *PROX* factor distributions and *LAG* factors for onroad mobile emissions.

ADD is an additive factor that accounts for emission sources within or near to a microenvironment, i.e., indoor emission sources. Unlike the other two factors, the *ADD* factor is itself a concentration and therefore has units of mass/volume. The actual units used must be the same as those in the *air quality* file.[16]

LAG, is used to account for the possibility of very slow pollutant diffusion and penetration, so that the relevant air quality concentration value may be from the previous time block. A value of zero for *LAG* indicates no time lag, i.e., use the concurrent air quality value; otherwise, the previous time block value is used. Due to lack of sufficient data to make estimates for *LAG*, the default file contains a uniform value of zero for all MEs.

The *factors* and *mobiles* files have no header records. The *factors* file contains one data record for each ME, with the same number, definition and order as the MEs in the *activity* file. The *mobiles* file contains a set of records (one for each ME) for each onroad mobile source category

[16] A data base of distributions of indoor source concentration contributions for several indoor source categories and subcategories is currently under development. The current version of the HAPEM program contains new, but untested algorithms to utilize the developing data base. Therefore, it is currently recommended that indoor sources be omitted from HAPEM6 applications until the database and algorithms have been tested and reviewed. To disable the indoor source algorithms, set keyword ***CAS*** to 99999.

identified with ***nmobiles***. The files are read in free format, once for each ME, with fields as specified in Tables 3-3a and 3-3b. All values are decimal numbers.

Distributions for *PROX* factors in the *factors* file are specified for each of the source categories separately. The number, definitions, and order of the emission source categories in the *factors* file must match those in the air quality file. The number of outdoor emission source categories is specified in the *parameter* file as ***nsource***.

A single set of *PEN* distributions, one for each ME, is specified in the *factors* file to be applied for all source categories. A single set of *ADD* factor distributions, one for each ME, is specified. And a single *LAG* factor, either 0 or 1, is specified for each ME.

The *mobiles* file contains PROX factor distributions and LAG factors for onroad mobile emissions. Each ME-specific record contains distributions for each of 3 distance-from-source categories: 0-75 meters, 75-200 meters, and greater than 200 meters. As noted above, the sequence numbers of the air quality concentrations associated with onroad mobile sources in the air quality file is specified in the parameters file as ***nmobiles***, and there must be a set of records (one for each ME) for each specified onroad mobile source category.[17] This information is combined with the data in the *DistToRoad* file (described below) in the HAPEM program to determine from which probability distribution the *PROX* factor should be selected for a given tract/ME combination. (See section 5.2.5 for more details.)

Distributions can take any of 5 different forms: normal, lognormal, uniform, triangular, or data set. The data set is comprised of up to 10 values, each of which is selected with equal probability.

The parameters that need to be specified for each type of distribution are as follows.

- Normal: arithmetic mean, arithmetic standard deviation, lower bound (optional), upper bound (optional) [Note: If both the lower and upper bounds are set to 0.0, then the distribution is sampled as if unbounded.]
- Lognormal: geometric mean, geometric standard deviation, lower bound (optional), upper bound (optional) [Note: If both the lower and upper bounds are set to 0.0, then the distribution is sampled as if unbounded.]
- Uniform: minimum, maximum
- Triangular: minimum, maximum, mode
- Data set: number of data values in the set (1-10), each value

[17] Note that a *PROX* factor distribution is specified in the *factors* file for the onroad mobile source category, but is not used by the HAPEM program. The *PROX* factor distributions in the *mobiles* file are used instead. Therefore, the *PROX* factor distributions for the onroad mobile source category may be specified arbitrarily in the *factors* file, as long as they meet the format requirements described below.

For HAPEM6 default *factors* files are provided for each of three categories of HAPs: gaseous, particulate, and semi-volatile. And default *mobiles* files are provided for benzene 1,3-butadiene, diesel particles, and a non-specific HAP (formatted for a single onroad mobile source category). Because, as noted above, a new approach to evaluating indoor sources is in development, the *ADD* factors are uniformly set to zero. And due to lack of data, *LAG* is uniformly set to zero. Excerpts from the default *factors* and *mobiles* files for gaseous HAPs and non-specific HAPs, respectively, are presented below.

As with other HAPEM6 input files, the user can add comments or other information after the last data record in the file. In this case a blank line need NOT be inserted after the last data record before the comments.

Extract from default *gas_factors* file (in wrapped view)

1	5	3	0.8	0.8	1	0	0	0	0	0	0	0	
	5	1	0	0	0	0	0	0	0	0	0	0	
	5	1	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
	4	1	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
2	5	3	0.8	0.8	1	0	0	0	0	0	0	0	
	5	1	0	0	0	0	0	0	0	0	0	0	
	5	1	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
	4	1	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
3	5	3	0.33	0.67	0.71	1	1	0	0	0	0	0	
	5	1	0	0	0	0	0	0	0	0	0	0	
	5	1	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
	4	1	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0

Extract from default *gas_mobiles* file (in wrapped view)

6	2	2.5263	2.0783	0	1	8.4161	0	0	0	0	0	0	0
	2	1.6404	1.9802	0	1	5.0469	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
7	2	2.5263	2.0783	0	1	8.4161	0	0	0	0	0	0	0
	2	1.6404	1.9802	0	1	5.0469	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0	0
8	2	2.5263	2.0783	0	1	8.4161	0	0	0	0	0	0	0
	2	2.5263	2.0783	0	1	8.4161	0	0	0	0	0	0	0
	2	2.5263	2.0783	0	1	8.4161	0	0	0	0	0	0	0

Table 3-3a.
Format for the *factors* file

ME Factor	Data Fields	Parameter	
	Field 1	Number of Microenvironment (1-14)	
PEN	Field 2	Distribution Type 1 - Normal 2 - Log-normal 3 - Uniform 4 - Triangular 5 - Data Set	
	Field 3	Distribution Type Normal Lognormal Uniform Triangular Dataset	Parameter Mean Mean Minimum Minimum Number of data points
	Field 4	Distribution Type Normal Lognormal Uniform Triangular Dataset	Parameter Standard deviation Standard deviation Maximum Maximum First data point in the set
	Field 5	Distribution Type Normal Lognormal Triangular Dataset	Parameter 0 (always) 0 (always) Mode Second data point in the set
PEN **(continued)**	Field 6	Distribution Type Normal Lognormal Dataset	Parameter Lower bound (optional) Lower bound (optional) Third data point in the set
	Field 7	Distribution Type Normal Lognormal Dataset	Parameter Upper bound (optional) Upper bound (optional) Fourth data point in the set

(continued)

Table 3-3a.
Format for the *factors* file

PEN **(concluded)**	Field 8	Distribution Type Parameter Dataset Fifth data point in the set
	Field 9	Distribution Type Parameter Dataset Sixth data point in the set
	Field 10	Distribution Type Parameter Dataset Seventh data point in the set
	Field 11	Distribution Type Parameter Dataset Eighth data point in the set
	Field 12	Distribution Type Parameter Dataset Ninth data point in the set
	Field 13	Distribution Type Parameter Dataset Tenth data point in the set
ADD	Fields 14-25	Repeat fields 2-13 for additive factor
PROX		
Source 1	Fields 26-37	Repeat fields 2-13 for proximity factor
Source 2	Fields 39-50	Repeat fields 2-13 for proximity factor
Source 3	Fields 52-63	Repeat fields 2-13 for proximity factor
Source 4	Fields 65-76	Repeat fields 2-13 for proximity factor
LAG		
Source 1	Field 38	hours
Source 2	Field 51	hours
Source 3	Field 64	hours
Source 4	Field 77	hours

(concluded)

Table 3-3b.
Format for the *mobiles* file (one onroad mobile source category)

ME Factor	Data Fields	Parameter	
	Field 1	Number of Microenvironment (1-14)	
PROX for Onroad Mobile Source Category: **Distance-from-Source Category 1**	Field 2	Distribution Type	
		1 - Normal	
		2 - Log-normal	
		3 - Uniform	
		4 - Triangular	
		5 - Data Set	
	Field 3	Distribution Type	Parameter
		Normal	Mean
		Lognormal	Mean
		Uniform	Minimum
		Triangular	Minimum
		Dataset	Number of data points
	Field 4	Distribution Type	Parameter
		Normal	Standard deviation
		Lognormal	Standard deviation
		Uniform	Maximum
		Triangular	Maximum
		Dataset	First data point in the set
	Field 5	Distribution Type	Parameter
		Normal	0 (always)
		Lognormal	0 (always)
		Triangular	Mode
		Dataset	Second data point in the set
	Field 6	Distribution Type	Parameter
		Normal	Lower bound (optional)
		Lognormal	Lower bound (optional)
		Dataset	Third data point in the set
	Field 7	Distribution Type	Parameter
		Normal	Upper bound (optional)
		Lognormal	Upper bound (optional)
		Dataset	Fourth data point in the set

(continued)

Table 3-3b.
Format for the *mobiles* file (one onroad mobile source category)

PROX for Onroad Mobile Source Category: **Distance-from-Source Category 1** **(concluded)**	Field 8	Distribution Type Dataset	Parameter Fifth data point in the set
	Field 9	Distribution Type Dataset	Parameter Sixth data point in the set
	Field 10	Distribution Type Dataset	Parameter Seventh data point in the set
	Field 11	Distribution Type Dataset	Parameter Eighth data point in the set
	Field 12	Distribution Type Dataset	Parameter Ninth data point in the set
	Field 13	Distribution Type Dataset	Parameter Tenth data point in the set
LAG for Onroad Mobile Source Category: **Distance-from-Source Category 1**	Field 14	Hours	
Distance-from-Source Category 2	Fields 15-27	Repeat fields 2-14	
Distance-from-Source Category 3	Fields 28-40	Repeat fields 2-14	

(concluded)

3.11. ClusTrans File

The C*lusTrans* file specifies for each demographic-group/day-type/commuting-status combination the number of activity patterns in each of 1 to 3 clusters (derived from cluster analysis on the activity pattern data from CHAD), and the cluster-to-cluster transition probabilities (derived from the transition frequencies for multiple-day activity pattern records from CHAD). These values are used to create weights for averaging selected activity patterns, one from each cluster, to represent an individual within the demographic group for that day type.

The *ClusTrans* file begins with a text header record, followed by one data record for each demographic group/day type combination. The header record indicates the order of the variables in each of the data records. The header record of the default *ClusTrans* file is as follows.

Header record from the default *ClusTrans* file

```
Demographic   DayType  "Comtype(1=non-commute,2=commuting)"    Ncluster  cluster1  cluster2
cluster3  prob11  prob12 prob13  prob21  prob22  prob23  prob31  prob32  prob33
```

Although the header record of the *ClusTrans* file is not used by the HAPEM6 programs, it provides documentation to inform the user of the meaning of the data fields.

The C*lusTrans* file is read in free format with the following for each demographic group/day type combination.

- Field 1 demographic group
- Field 2 day type
- Field 3 commuting status (1= non-commuting; 2= commuting)
- Field 4 number of clusters for the demographic group/day type (2 or 3)
- Field 5 cumulative fraction of demographic group/day type activity patterns in cluster 1
- Field 6 cumulative fraction of demographic group/day type activity patterns in cluster 2
- Field 7 cumulative fraction of demographic group/day type activity patterns in cluster 3
- Field 8 cumulative cluster 1 to cluster 1 transition probability
- Field 9 cumulative cluster 1 to cluster 2 transition probability
- Field 10 cumulative cluster 1 to cluster 3 transition probability
- Field 11 cumulative cluster 2 to cluster 1 transition probability
- Field 12 cumulative cluster 2 to cluster 2 transition probability

- Field 13 cumulative cluster 2 to cluster 3 transition probability
- Field 14 cumulative cluster 3 to cluster 1 transition probability
- Field 15 cumulative cluster 3 to cluster 2 transition probability
- Field 16 cumulative cluster 3 to cluster 3 transition probability

The default *ClusTrans* file is presented below.

Default *ClusTrans* file

```
1  1 1 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
1  1 2 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
1  2 1 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
1  2 2 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
1  3 1 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
1  3 2 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
2  1 1 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
2  1 2 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
2  2 1 3 0.50435  0.78261  1.00000 0.78947  0.84211 1.00000 0.06250 0.93750  1.00000 0.42857  0.57143  1.00000
2  2 2 1 1.00000  0.00000 0.00000 1.00000  0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
2  3 1 2  0.46025  1.00000 0.00000 0.25000 1.00000 0.00000  0.30000  1.00000 0.00000 0.00000 0.00000  0.00000
2  3 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000  0.00000
3  1 1 3  0.34314  0.73856  1.00000 0.60000 0.91111  1.00000  0.38298  0.91489  1.00000 0.04545 0.18182  1.00000
3  1 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000  0.00000
3  2 1 3  0.23145  0.76048  1.00000 0.77778 0.94444  1.00000  0.06897  0.82759  1.00000 0.14286 0.50000  1.00000
3  2 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000  0.00000
3  3 1 3  0.30561  0.88659  1.00000 0.39286 0.92857  1.00000  0.22500  0.95000  1.00000 0.26667 0.86667  1.00000
3  3 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000  0.00000
4  1 1 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
4  1 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
4  2 1 3  0.21656  0.76433  1.00000 0.68750 0.87500  1.00000  0.05556  0.77778  1.00000 0.18750 0.87500  1.00000
4  2 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
4  3 1 2  0.45536  1.00000  0.00000 0.75000 1.00000  0.00000  0.22222  1.00000 0.00000 0.00000 0.00000  0.00000
4  3 2 1  1.00000  0.00000  0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
5  1 1 2  0.60828  1.00000  0.00000 0.76923 1.00000  0.00000  0.56667  1.00000 0.00000 0.00000 0.00000  0.00000
5  1 2 3  0.79192  0.88822  1.00000 0.93846 0.98462  1.00000  0.70455  0.95455  1.00000 0.16667 0.33333  1.00000
5  2 1 3 0.27483  0.79628  1.00000 0.47727 0.90909  1.00000  0.23529  0.96471  1.00000 0.22222 0.33333  1.00000
5 2 2  3 0.63076  0.79685  1.00000 0.78218 0.85149  1.00000  0.42568  0.98649  1.00000 0.60606 0.69697  1.00000
5 3 1  3 0.66244  0.89170  1.00000 0.84615 0.91209  1.00000  0.33333  0.94444  1.00000 0.61538 0.84615  1.00000
5 3 2  3 0.41328  0.61378  1.00000 0.87500 0.92500  1.00000  0.08333  0.54167  1.00000 0.29412 0.52941  1.00000
6 1 1  2 0.60788  1.00000  0.00000 0.93868 1.00000  0.00000  0.21429  1.00000 0.00000 0.00000 0.00000  0.00000
6 1 2  1 1.00000  0.00000   0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
6 2 1  3 0.36886  0.61851  1.00000 0.27586 0.75862  1.00000  0.39130  0.91304  1.00000 0.44444 0.50000  1.00000
6 2 2  1 1.00000  0.00000   0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
6 3 1  2 0.67657  1.00000  0.00000 0.91525 1.00000  0.00000  0.33333  1.00000 0.00000 0.00000 0.00000  0.00000
6 3 2  1 1.00000  0.00000   0.00000 1.00000  0.00000  0.00000  0.00000  0.00000 0.00000 0.00000  0.00000 0.00000
```

3.12. Statefip File

The *statefip* file cross-references the 2-digit state FIPS codes for each US state to its numerical ranking on the list. The default *statefip* file contains 53 codes: one for each US state, the District of Columbia, Puerto Rico, and the US Virgin Islands. It is presented below. The format of each record is as follows.

Fields 1-2 numerical rank (integer)

Fields 9-10 state FIPS code (2-character string)

As discussed in section 2.1.7, the *statefip* file is used in conjunction with the parameters ***region1*** and ***region2*** specified in the *parameter* files of INDEXPOP, COMMUTE, AIRQUAL, and HAPEM to specify the group of states to be included in the analysis, according to numerical ranking.

Default *statefip* file and corresponding state names

1	01	Alabama
2	02	Alaska
3	04	Arizona
4	05	Arkansas
5	06	California
6	08	Colorado
7	09	Connecticut
8	10	Delaware
9	11	District Of Columbia
10	12	Florida
11	13	Georgia
12	15	Hawaii
13	16	Idaho
14	17	Illinois
15	18	Indiana
16	19	Iowa
17	20	Kansas
18	21	Kentucky
19	22	Louisiana
20	23	Maine
21	24	Maryland
22	25	Massachusetts
23	26	Michigan
24	27	Minnesota
25	28	Mississippi
26	29	Missouri
27	30	Montana
28	31	Nebraska
29	32	Nevada
30	33	New Hampshire
31	34	New Jersey
32	35	New Mexico
33	36	New York
34	37	North Carolina
35	38	North Dakota
36	39	Ohio
37	40	Oklahoma
38	41	Oregon
39	42	Pennsylvania
40	44	Rhode Island
41	45	South Carolina
42	46	South Dakota
43	47	Tennessee
44	48	Texas
45	49	Utah
46	50	Vermont
47	51	Virginia
48	53	Washington
49	54	West Virginia
50	55	Wisconsin
51	56	Wyoming
52	72	Puerto Rico
53	78	US Virgin Islands

This page intentionally left blank.

4. HAPEM6 Output Files

The HAPEM6 creates three diagnostic output files and a set of final exposure output files. The diagnostic files record information about the parameters of the simulations, as well as error messages. The names for these three files are specified by the user in the *parameter* files of each HAPEM6 program. The final exposure output files contain all of the exposure estimates from a HAPEM6 run. The path name for these files are specified by the user in the *parameter* file for the HAPEM program.

4.1. Log File

The *log* file contains a record of a HAPEM6 analysis. Three of the HAPEM6 programs (INDEXPOP, COMMUTE, and HAPEM) will append records onto an existing *log* file, as specified its *parameter* file, without overwriting previous records. DURAV and AIRQUAL will overwrite any records on an existing *log* file. Therefore, if a single *log* file name is used to run all the HAPEM6 programs, a running record will be kept for DURAV, INDEXPOP, and COMMUTE, but a new record will be started by AIRQUAL and then added to by HAPEM. To maintain a complete *log* file record of a HAPEM6 simulation, two alternatives are:

- If a single *parameter* file is for a complete simulation, so that the *log* file name is the same for all 5 programs, rename the *log* file created by the first 3 programs before AIRQUAL is run.
- Use a different *parameter* file for running AIRQUAL and HAPEM than for the other programs, with a different name specified for the *log* file.

If the HAPEM6 programs experience no fatal errors during a simulation, there are several items written to the *log* file by each of the programs. The first record written to the file by each program identifies the program and its start time. The time consists of three numbers; the first is the current time, the second is the size of the time increment equivalent to one second, and the third is the maximum value allowed for the current time before it is reset to zero. All three of these quantities are system dependent. An example record of this type is presented below.

Example *log* file program and start time record

```
Durav Start time= 34862630 1000 86399999
```

The last two records written to the *log* by each HAPEM6 program report the ending time and the total job time for the particular program. For the total job time record, the job time is converted into seconds. Note that the total job time will not be correct if the clock maximum is exceeded during the job. An example of these type of records is presented below.

Example *log* file stop time and run time record

```
Durav End time =  34880980
Durav Job time =  18.3500004
```

If a error occurs that the HAPEM6 considers to be fatal, a diagnostic message will be written to the *log* file and the program stopped. For example, if DURAV finds that the number of time blocks per day specified in the *activity* file header does not match the value of ***nblock*** specified in the *parameter* file, it will write a message to the *log* file and stop. An example of this type of record is presented below.

Example *log* file error message record

```
number of time blocks in activity file does not equal nblock 999
```

4.1.1. DURAV Output to the Log File

Apart from the text produced by all HAPEM6 programs, each program writes some specialized information to the *log* file. DURAV writes the name of the input *activity* file, and intermediate file used to sort activity patterns for each group. An example of these types of records is presented below.

Example *log* file input and intermediate file records

```
Data read from file=c:\HAPEM6\durhw.txt
direct access in file=c:\hapem\durhw.draft
```

DURAV also records the number of records (person-days) extracted from the *activity* file, and the table of frequency counts for each combination of demographic group and day type (a matrix whose elements should sum to the total number of records extracted). If any elements of this matrix are zero then there are groups that have no activity patterns and thus are undefined. If the numbers are positive but small (*e.g.*, less than ten), then there is a chance that the exposure results might not be representative for the group. An example of a part of this type of matrix is presented below.

Example *log* file matrix records

```
Total number of person-days processed= 16613
Frequency table for person-days:
By demographic group (rows) & day type (cols)
   1837 2853   4266
   1276 1735   1972

           .
           .
           .
```

Before completing execution, DURAV writes the name of the output file (the averaged activity database) in the *log* file following the table of frequency counts. An example of this type of record is presented below.

Example *log* file output file record

```
Data written to file=c:\HAPEM6\durhw.da
```

4.1.2. INDEXPOP Output to the Log File

In addition to the program name and the start, stop, and job time information provided to the *log* file by all the HAPEM6 programs, INDEXPOP writes 2 other records to the *log* file. The first confirms that all the input files were successfully opened, and the second records the total number of tract records in the *population* file. An example of these two records is presented below.

Example *log* file records from INDEXPOP

```
Finished opening files
total number of tracts is 60803
```

4.1.3. COMMUTE Output to the Log File

The COMMUTE program write no information to the *log* file other than the program name and the start, stop, and job time.

4.1.4. AIRQUAL Output to the Log File

In addition to the program name and the start, stop, and job time information provided to the *log* file by all the HAPEM6 programs, AIRQUAL writes several other records to the *log* file. First, a summary of the *air quality* file, which is produced by counting the number of census tracts and distinct counties found in the file, is written. These tracts are then paired with the tracts found in the *population* file. The number of tracts found in the *air quality* file but not in the *census* file are recorded in the line containing the phrase “unpaired air tracts”. This is followed by the list (if any) of unpaired tracts. Then, the number of tracts in the *population* file but not in the *air quality* file is reported, along with the number of matches and the number of repeated matches (two or more air tracts for one census tract). Next, similar statistics are given, for counties that have air quality data. Ideally, every tract in the *air quality* file should be paired with exactly one tract in the *population* file. An example of the log output produced by AIRQUAL is presented below.

Example *log* file records from AIRQUAL

```
# air tracts = 60803
 # counties on air file = 3111 # of air records =
60803
 There were  0  unpaired air tracts.
 Overall, there were:
 455  unpaired census tracts.
 60803  census tracts with a matching air tract.
 0  census tracts with 2 or more air tracts.
 Within the counties on the air file, there were:
 0  unpaired census tracts.
 60803  census tracts with a matching air tract.
 0  census tracts with 2 or more air tracts.
```

4.1.5. HAPEM Output to the Log File

In addition to the program name and the start, stop, and job time information provided to the *log* file by all the HAPEM6 programs, the HAPEM program writes two other records to the *log* file.

HAPEM reports the time when dynamic array allocation is complete (i.e., `HAPEM Allocation =`) and the number of tracts used in the analysis (i.e., that had data in the *air quality*, *population*, and *commuting* files). An example of the log output produced by HAPEM is presented below.

Example *log* file records from HAPEM

```
HAPEM Allocation = 35921930
There were  4046  tracts in the study area.
```

4.2. Counter File

A second diagnostic file created by HAPEM6 is the *counter* file. The *counter* file records the number of records in various data input and output files, which can also be a useful tool for keeping track of which files were used in the simulation, and for troubleshooting.

It is important to use same *counter* file for all the HAPEM6 programs in a simulation, because the programs use some of the information recorded by previous programs for dynamic memory allocation of arrays. If the expected records from previous programs are not in the *counter* file, an error will occur.

The HAPEM6 programs add records to the *counter* file by appending to the end of the records generated by the previous programs, where programs are run in the expected order, as described in section 2.1. (Running COMMUTE is optional). For example, INDEXPOP reads the DURAV-generated records, and then begins its own recording. If INDEXPOP is run a second time, using the same counter file, the second run will overwrite the previous INDEXPOP-generated records.

The specific information recorded in the *counter* file is as follows.

DURAV:	record 1	-	number of data records in the *activity* file (*durhw.txt*)
		-	number of *activity* file data records with 1440 minute total
INDEXPOP	record 2	-	number of data records (tracts) in the *population* file (*census.txt*)
		-	number of counties in the *population* file (*census.txt*)
COMMUTE	record 3	-	number of records in the population index file (*census_direct.ind*)
		-	number of records in the *commuting* file (*comm.txt*)
	record 4	-	number of records in the work tract file (*comm.da*)
		-	number of records in the commuting index file *(comm.ind*)
AIRQUAL	record 5	-	number of matching tracts in the *air quality* (e.g., *benzene.txt*) and population index (*census_direct.ind*) files

		- number of counties with matching tracts in the *air quality* (e.g., *benzene.txt*) and population index (*census_direct.ind*) files
AIRQUAL	record 6	- number of tracts in the *air quality* (e.g., *benzene.txt*)
		- number of data records in the air quality file (e.g., *benzene*.txt)

The following relationships are expected among the numbers in the *counter* file.

- The number of records in the *population* file (*census.txt*), the population index file (*census_direct.ind*), and the commuting index file *(comm.ind*) should all be the same.
- The number of records in the work tract file (*comm.da*) may be larger or smaller than the number of records in the commuting file (*comm.txt*). It may be larger, because, if a tract in the *population* file has no matching home tract in the *commuting* file, COMMUTE creates a “commuting” flow, using the population tract as both the home and work tract. It may also be smaller if the *population* file study area is smaller than the *commuting* file study area (all US states and the District of Columbia).

4.3. Mistract File

A third diagnostic file created by the COMMUTE, AIRQUAL, and HAPEM programs is the *mistract* file. If the same *mistract* file name is used for COMMUTE and AIRQUAL, The COMMUTE information will be overwritten by AIRQUAL. HAPEM will append records onto an existing *mistract* file. To maintain a complete record of this information for a HAPEM6 simulation, either different *mistract* file names should be used for COMMUTE and AIRQUAL (requiring different *parameter* files), or the *mistract* file should be re-named after the COMMUTE program is run.

Each of the three programs records a different set of information about the consistency of census tracts included in the input files.

- The COMMUTE *mistract* file records the state, county, and tract FIPS codes of each tract in the *population* file that is not matched by a home tract in the *commuting* file. These unmatched tracts are still processed by COMMUTE, as explained in the previous section, by creating a “commuting” flow, using the population tract as both the home and work tract.
- The AIRQUAL *mistract* file records the record number, state and county FIPS, and tract code of each tract in the *population* file that is not matched by a tract in the *air quality* file. Only tracts that are included in both the files are processed by HAPEM6, since both these pieces of information about a tract (population and air quality) are needed to make an exposure estimate.
- The HAPEM *mistract* file records the state, county, and tract FIPS codes of each home tract in the *commuting* file that is not matched by a tract in the air quality index files. These index files contain information on tracts that were included in both the *population* and *air quality* files. The unmatched home tracts are not processed further. The HAPEM *mistract* file also records each instance of a work tract that is not matched by a tract in the *air quality* file. For these cases, the work tract is assigned the air quality of the home tract.

4.4. Final Exposure File

As explained in section 2.1.9, HAPEM6 creates an exposure output file for each state/pollutant combination. The names of these files are constructed by the program based on the pollutant SAROAD code and the state FIPS code as follows:

XXXXX.YY.dat

where XXXXX = the 5-digit SAROAD pollutant code specified by the ***sarod*** parameter

and YY = the 2-digit state FIPS code

4.4.1. File Format

The final exposure output files each begin with a repetition of some of the information specified in the parameter file for the HAPEM program, as follows.

- State fips
- Pollutant SAROD code
- Pollutant name
- Pollutant CAS #
- Air quality data units
- Year of air quality data
- Number of outdoor air emission source categories (i.e., the value of ***nsource***)
- Random number seed for activity pattern selection
- Random number seed for microenvironment factors selection
- Random number seed for air quality data selection
- Number of indoor product emission sources types
- Number of indoor material emission source types
- Number of indoor combustion emission source types
- Number of vehicle in residential garage emission source types
- EPA Region of indoor emission source data
- Number of demographic groups (i.e., the value of ***ngroup***)
- Number of replicates for each demographic group (i.e., the value of ***nreplic***)
- Definition of each demographic group, ordered as in the *population* file

This information is followed by a header record defining the fields in the data records. An example header record is presented below.

Example header record for final exposure output file

```
ST CTY CENSUS GRUP POPUL    SOURCE1    SOURCE2    SOURCE3    SOURCE4
BackgConc IndCon_Pro IndCon_Mat IndCon_Com IndCon_Veh Total Conc
```

The header record is then followed by ***nreplic*** data records for each group/tract combination. The format of each data record, assuming ***nsource*** = 4, is as follows.

Fields 2-3 state FIPS code (2-character string)

Fields 5-7 county FIPS code (3-character string)

Fields 9-14 tract ID (6-character string)

Fields 16-17 demographic group indicator (integer from 1-10, ordered as in the *population* input file)

Fields 19-25 number of people to which the exposure estimates in the data record apply, equal to the population of the group/tract combination divided by ***nreplic*** (decimal number)

Fields 27-36 estimated exposure concentration contribution from emission source category 1 (decimal number in scientific notation, units of measurement as in the *air quality* file)

Fields 38-47 estimated exposure concentration contribution from emission source category 2

Fields 49-58 estimated exposure concentration contribution from emission source category 3

Fields 60-69 estimated exposure concentration contribution from emission source category 4

Fields 71-80 estimated exposure concentration contribution from background (derived from the sum of the uniform background – ***backg*** – and the variable background)

Fields 82-91 estimated exposure concentration contribution from indoor product emission sources

Fields 82-91 estimated exposure concentration contribution from building materials indoor emissions

Fields 93-102 estimated exposure concentration contribution from indoor combustion emission sources

Fields 104-113 estimated exposure concentration contribution from vehicles in attached garages

Fields 115-124 estimated total exposure concentration, the sum of the preceding 9 values

Examples of one full and one partial set of replicate data records are presented below. The total population for demographic group 1 in this tract is 113 and ***nreplic*** = 30, so that the number of people to which the exposure estimates in each record apply is 3.767 = 113/30.

Example sets of replicate data records from a final exposure output file

```
48 201 212100  1   3.767 0.9604E+00 0.2540E+00 0.3881E+00 0.1140E+00 0.2102E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1927E+01
48 201 212100  1   3.767 0.9867E+00 0.2622E+00 0.7042E+00 0.1157E+00 0.2135E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2282E+01
48 201 212100  1   3.767 0.1129E+01 0.3303E+00 0.4181E+00 0.9131E-01 0.2093E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2178E+01
48 201 212100  1   3.767 0.1024E+01 0.2163E+00 0.4663E+00 0.9332E-01 0.2085E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2008E+01
48 201 212100  1   3.767 0.8908E+00 0.2251E+00 0.4172E+00 0.1186E+00 0.2198E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1872E+01
48 201 212100  1   3.767 0.9728E+00 0.2667E+00 0.4287E+00 0.1139E+00 0.2090E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1991E+01
48 201 212100  1   3.767 0.9284E+00 0.2298E+00 0.5169E+00 0.1091E+00 0.2231E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2007E+01
48 201 212100  1   3.767 0.1194E+01 0.3273E+00 0.5084E+00 0.9688E-01 0.2138E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2341E+01
48 201 212100  1   3.767 0.1049E+01 0.2174E+00 0.5472E+00 0.9774E-01 0.2184E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2130E+01
48 201 212100  1   3.767 0.9183E+00 0.2561E+00 0.5116E+00 0.1043E+00 0.2081E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1998E+01
48 201 212100  1   3.767 0.8146E+00 0.2190E+00 0.3807E+00 0.1078E+00 0.2058E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1728E+01
48 201 212100  1   3.767 0.1007E+01 0.2039E+00 0.4805E+00 0.9197E-01 0.2051E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1989E+01
48 201 212100  1   3.767 0.8943E+00 0.2012E+00 0.3477E+00 0.1434E+00 0.2067E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1793E+01
48 201 212100  1   3.767 0.1141E+01 0.2049E+00 0.4629E+00 0.1350E+00 0.2130E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2157E+01
48 201 212100  1   3.767 0.8135E+00 0.2207E+00 0.4548E+00 0.1096E+00 0.2134E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1812E+01
48 201 212100  1   3.767 0.1122E+01 0.2852E+00 0.5223E+00 0.9469E-01 0.2154E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2239E+01
48 201 212100  1   3.767 0.1124E+01 0.1984E+00 0.4117E+00 0.1310E+00 0.2037E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2069E+01
48 201 212100  1   3.767 0.8670E+00 0.2233E+00 0.4679E+00 0.1163E+00 0.2154E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1890E+01
48 201 212100  1   3.767 0.1182E+01 0.2086E+00 0.4138E+00 0.1386E+00 0.2163E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2160E+01
48 201 212100  1   3.767 0.1185E+01 0.2732E+00 0.4205E+00 0.1134E+00 0.2191E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2211E+01
48 201 212100  1   3.767 0.2687E+01 0.2293E+00 0.4610E+00 0.1030E+00 0.2208E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3702E+01
48 201 212100  1   3.767 0.8996E+00 0.1995E+00 0.3616E+00 0.1463E+00 0.2107E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1818E+01
48 201 212100  1   3.767 0.8803E+00 0.2157E+00 0.3859E+00 0.1248E+00 0.2100E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1817E+01
48 201 212100  1   3.767 0.8458E+00 0.2069E+00 0.4599E+00 0.1158E+00 0.2114E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1840E+01
48 201 212100  1   3.767 0.1190E+01 0.2062E+00 0.4067E+00 0.1376E+00 0.2139E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2154E+01
48 201 212100  1   3.767 0.9004E+00 0.1963E+00 0.3892E+00 0.1459E+00 0.2084E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1840E+01
48 201 212100  1   3.767 0.1083E+01 0.2642E+00 0.3631E+00 0.1028E+00 0.2021E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2015E+01
48 201 212100  1   3.767 0.9782E+00 0.2152E+00 0.3841E+00 0.1588E+00 0.2256E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1962E+01
48 201 212100  1   3.767 0.1011E+01 0.3019E+00 0.4500E+00 0.1037E+00 0.2161E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2082E+01
48 201 212100  1   3.767 0.1052E+01 0.2248E+00 0.5081E+00 0.9568E-01 0.2156E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2096E+01
48 201 212100  2   5.733 0.9436E+00 0.2130E+00 0.5572E+00 0.1486E+00 0.2154E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2078E+01
48 201 212100  2   5.733 0.1703E+01 0.1965E+00 0.3623E+00 0.1374E+00 0.2129E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2612E+01
48 201 212100  2   5.733 0.1473E+01 0.2037E+00 0.3782E+00 0.1488E+00 0.2200E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2423E+01
48 201 212100  2   5.733 0.9426E+00 0.1964E+00 0.3624E+00 0.1340E+00 0.2034E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1839E+01
48 201 212100  2   5.733 0.1507E+01 0.2025E+00 0.3933E+00 0.1559E+00 0.2295E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2488E+01
48 201 212100  2   5.733 0.9131E+00 0.2487E+00 0.5014E+00 0.1051E+00 0.2099E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1978E+01
48 201 212100  2   5.733 0.1087E+01 0.2228E+00 0.5952E+00 0.9670E-01 0.2182E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2220E+01
48 201 212100  2   5.733 0.1073E+01 0.2352E+00 0.5623E+00 0.9799E-01 0.2236E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2192E+01
48 201 212100  2   5.733 0.8830E+00 0.2277E+00 0.3946E+00 0.1358E+00 0.2114E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1852E+01
48 201 212100  2   5.733 0.1152E+01 0.2820E+00 0.3900E+00 0.1085E+00 0.2131E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2146E+01
48 201 212100  2   5.733 0.3379E+01 0.1904E+00 0.3562E+00 0.1305E+00 0.2075E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4264E+01
48 201 212100  2   5.733 0.9684E+00 0.2834E+00 0.4670E+00 0.1039E+00 0.2166E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2039E+01
48 201 212100  2   5.733 0.9111E+00 0.2324E+00 0.4539E+00 0.1388E+00 0.2151E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1951E+01
48 201 212100  2   5.733 0.1739E+01 0.2001E+00 0.3652E+00 0.1395E+00 0.2168E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2660E+01
48 201 212100  2   5.733 0.1079E+01 0.2309E+00 0.5047E+00 0.9671E-01 0.2199E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2131E+01
48 201 212100  2   5.733 0.9045E+00 0.2019E+00 0.3773E+00 0.1460E+00 0.2111E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1841E+01
```

5. HAPEM6 Programs

This section presents detailed descriptions of the five programs that comprise the model: DURAV, INDEXPOP, COMMUTE, AIRQUAL, and HAPEM. The first four programs (DURAV, INDEXPOP, COMMUTE, and AIRQUAL) are essentially preprocessers that convert input data files into the form required for efficient exposure calculations. The final program (HAPEM) actually performs the exposure calculations and summarizes the results.

NOTE TO USERS

It is important to note that some knowledge of Fortran programming is necessary to understand all the programming details discussed in this chapter. However, all of the general concepts related to the programs should be clear to all users.

5.1. Programming Guidelines Used to Develop HAPEM6

The source code for each of the five HAPEM6 programs is written in Fortran 90 and designed so that it can be compiled and executed with little or no programming changes on various platforms (*e.g.,* UNIX, DOS, Windows).

The HAPEM6 programs incorporate a structured programming style as summarized by the following attributes:

- No “GO TO” statements or line numbers in any of the programs. Program flow is direct from the beginning to the end within each program, thus making the code easy to follow. The only looping is within “DO” blocks.
- No filenames appear in source code of any of the HAPEM6 programs. Instead, this information is specified in the *parameter* file.
- Most parameter values are input from the *parameter* file so that the programs themselves only allocate space for and carry out as many calculations as are necessary.
- All arrays depending on variable parameters are dynamically allocated.
- To avoid limitations such as hard-coded numbers of variables on a “FORMAT” statement, many of the “FORMATS” are dynamically written during the run.
- All variables are declared (no implicit typing), with comments at the end of most declarations to assist in interpretation. Comment lines are inserted between the logical blocks of code for clarity.

- Generally, the “READ” statements use unformatted (list) input so that the data do not have to be in predetermined columns, but the “WRITE” statements are usually formatted for clarity.

5.1.1. Common Structural Elements

All of the HAPEM6 programs consist of a declarations section, a parameters section, a setup section, a primary section that processes the data, and a wrap-up section. In the declarations section, all variables are explicitly typed. Most lines include a trailing comment to indicate the general purpose of the variable(s). Arrays that are to be dynamically allocated are fixed in rank (number of dimensions), with a colon used to defer the size specification.

The second program section, referred to as the params section, reads the *parameter* file to determine the specific input file names and the parameter settings. This section is similar in all the HAPEM6 programs, except that only the names of files needed by each job are retained as variables. Each line of the *parameter* file is read in as a character string (maximum length of 120 characters) and inspected for an equals sign “=”. If there is no equals sign then the line is ignored. This allows the programmer to add comments and other lines directly to the *parameter* file without altering its performance. Lines containing an equals sign are divided into two parts at the equals sign. The part to the left of the sign is scanned for keywords. All keywords are in lower case. If the string ‘file’ is found, then the line is assumed to specify one of the input or output files. For these lines, a second keyword is searched for. Possible keywords are presented in Table 5-1. Which filenames and paths are required by each HAPEM6 program are presented in Table 2.1 as user-defined files.

The HAPEM6 user can use the above keywords in lines that do not contain an equals sign, or in comments containing an equals sign as long as the word “file” does not also appear left of the equals sign. The strings containing the directory and file names should not exceed 100 characters. If they do, then use an alias or a logical drive specification to identify most of the path, and thereby reduce the length to less than 100 characters. As described earlier in this guide, each of the input files requires a certain format for the data. It is the responsibility of the user to ensure that this format specification is met.

The setup section allocates and initializes the dynamic arrays that can be sized from the parameter settings specified in the *parameter* file. Other arrays that are dependent on the number of records in an input file are allocated elsewhere. The dynamic allocation serves three purposes, it saves space and time by only using as much space as is necessary, it allows for the parameters to be increased or decreased without recompiling the program, and it allows vector and array operations to be programmed more simply since they can be applied to the entire array rather than only to certain elements.

5.2. Program Descriptions

As mentioned previously, HAPEM6 is composed of five programs: DURAV, INDEXPOP, COMMUTE, AIRQUAL ,and HAPEM. This section describes the purpose and structure of the processing section of each of these programs.

Table 5-1.
Keywords for filenames recognized by HAPEM6 *parameter* files

Keyword	Definition
activity	name of the *activity* file (input)
cluster	name of the *cluster* file (input)
ClusTrans	name of the *cluster transition probability* file (input)
populat	name of the *population* file (input)
CommutTime	name of the *commuting time* file (input)
CommutFrac	name of the *commuting fraction* file (input)
DistToRoad	name of the *distance to major roadway* file (input)
commut	name of the *commuting flow* file (input)
quality	name of the *air quality* file (input)
factors	name of the *factors* file (input)
mobiles	name of the *mobiles* file (input)
statefip	name of the *statefip* file (input)
Log	name of the *log* file (output)
counter	name of the *counter* file (output)
mistract	name of the *mistract* file(output)
afile	path of final exposure file (output)
product[18]	path of indoor source files (input)
AutoPduct[19]	Name of file for automobile-related consumer products (input)

5.2.1. DURAV

As explained in section 2.1.2, the DURAV program performs three main functions.

[18] A path to one or more indoor emission source inputs for the HAPEM6 indoor source algorithms is specified in this statement. These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword ***CAS*** to 99999. Since no indoor source files will then actually be utilized by the HAPEM program, any existing path may be specified.

[19] The full path name of an existing file must be specified as the *AutoPduct* file in HAPEM6, although its only function is as input to the indoor source algorithms. These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword ***CAS*** to 99999. Since the *AutoPduct* file will then not actually be utilized by the HAPEM program, any existing file name may be specified, other than those otherwise specified for input or output for the HAPEM program.

- It categorizes and groups population activity data extracted from CHAD into demographic groups, day types (season, day-of-week), commuting status groups, and cluster categories.
- If a different number of daily time blocks is specified for the analysis than in the activity data file, it processes the activity records so that the number of time blocks matches the number specified for the analysis.
- It creates a sequential file of the activity pattern records for use by the HAPEM program.

The 6 demographic groups are defined by the following age categories.

- 0 – 1
- 2 - 4
- 5 - 15
- 16 - 17
- 18 - 64
- 65+

Two variables, season and day of week, are used to determine three day types:

- Weekdays in summer (June - August)
- Other weekdays
- Weekends.

Cluster types are used to represent variation in activity pattern within each combination of demographic group, day type, and commuting status. There are 1 or 3 cluster types for each demographic-group/day-type/commuting-status combination. Each CHAD record in the *activity* file has been assigned a cluster type based on the cluster analyses.

DURAV Processing Operations

In addition to the operations discussed above, the params section of DURAV conducts the following operation.

- The values of ***nblock***, the number of time blocks per day in the *activity* file, and ***hblock***, the number of time blocks per day for the analysis, are checked for compatibility. As explained above, ***hblock*** must be an integral factor of ***nblock***, so that the activity time blocks can be combined if necessary to match to match ***hblock.*** If not, an error message is written to the *log* file and the program is stopped.

In addition to the operations discussed above, the setup section of DURAV conducts the following operation.

- The number of time blocks per day in the *activity* file is determined from the header record, as explained in section above. This number is checked against the value of ***nblock*** specified in the *parameter* file. If the values are different, an error message is written to the *log* file and the program is stopped.

The main processing section of DURAV conducts several operations, as follows.

- The number of data records in the *activity* file is determined, so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- Each activity record is checked to ensure that the total activity time is 1440 minutes. If not the record is recorded in an intermediate output file with filename extension "*.wrong_chad*" and dropped from further processing.
- The ***nblock*** time blocks in each activity record are aggregated, if necessary, to create ***hblock*** time blocks.
- The compressed activity records are checked to ensure the total activity time is still 1440 minutes. If not, an error message is written to the *log* file and the program is stopped.
- The compressed activity records are written into a direct access file with a file name extension of "*.draft".*
- Each compressed record is classified by demographic group, day type, and commuting status, as defined in the DURAV source code. If any records cannot be classified, and error message is written to the *log* file and the program is stopped. The sequence numbers of the compressed records for each category are recorded in the array "ntest".
- The total number of data records in the *activity* file, and the total number with activity durations of 1440 minutes are recorded in the *counter* file.
- The number of compressed records in each demographic-group/day-type/commuting-status category is determined.
- The compressed activity records are read from the direct access *.*draft* file and each is additionally classified by cluster within its demographic group/day type category, according to the specifications in the *cluster* file.
- The number of compressed activity records in each demographic-group/day-type/commuting-status/cluster combination, and the number of clusters in each demographic-group/day-type/commuting-status combination are recorded in an intermediate file with filename extension "*.nonzero*".[20] This information is used in the HAPEM program, as described below.
- The total number of compressed activity records processed and their allocation among demographic groups, day types, and commuting status is written into the *log* file.
- The activity patterns are written into a sequential file with filename extension "*.da*" sorted by demographic group, day type, commuting status, and cluster type, and the filename is recorded in the *log* file.

5.2.2. INDEXPOP

As explained above, the INDEXPOP program performs two main functions:

[20] The "*.nonzero*" file also records a flag for each demographic group/day type combination indicating whether 10% of the activity patterns include commuting. This flag was used by an earlier version of the HAPEM program, but is not used in this version.

- It creates a direct access file of population data to be used in AIRQUAL.
- It creates sequential index files for the population data Census tracts, to facilitate file searching in COMMUTE and AIRQUAL.
- It creates direct access files and associated index files of the data in the *DistToRoad*, *CommutTime*, and *CommutFrac* files, to be used in COMMUTE and AIRQUAL

INDEXPOP Processing Operations

The specific operations performed in the main processing section of INDEXPOP are as follows.

- Each record in the *DistToRoad* file is read and written into a direct access file with filename extension "*.dat*", and an associated index file is created with filename extension "*.STIDX*".
- Each record in the *CommutTime* file is read and written into a direct access file with filename extension "*.dat*", and an associated index file is created with filename extension "*.STIDX*".
- Each record in the *CommutFrac* file is read and written into a direct access file with filename extension "*.dat*", and an associated index file is created with filename extension "*.STIDX*".
- The number of data records in the *population* file is determined so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- Each data record in the *population* file is read. The population array is recorded in a direct access file with the filename extension "*.da*". The state FIPS, county FIPS, tract code, and serial record number are recorded in a direct access file with the filename extension "*_direct.ind*".
- The total number of tract records in each county is determined.
- The total number of counties included in the *population* file that fall into each state is determined.
- A sequential index file is created with filename extension "*.county_tract_pop_range*". For each county in the *population* file, there is a record in this file indicating the serial record numbers of the first and last data record for tracts in that county in the "*.da*" and "*_direct.ind*" files.
- A sequential index file is created with filename extension "*.state_county_pop_range*". For each county there is a record in this file indicating the serial record numbers of the first and last data record for counties in that state in the "*.county_tract_pop_range*".
- The total number of records (tracts) and counties in the *population* file is added to the *counter* file.

5.2.3. COMMUTE

As explained in section 2.1.4, the COMMUTE program performs two main functions.

- It creates a file identifying for each Census tract (i.e., home tract) the associated set of work tracts (i.e., tracts in which the residents of the home tract work), the fraction of home tract workers in each work tract, and the normalized centroid-to-centroid distance between home tract and each work tract. The normalized distance is the distance/(average distance). The normalized distance is combined with the average commuting time for the tract to estimate the commuting time for the home-tract/work-tract pair in HAPEM.

- It creates sequential index files to facilitate file searching in HAPEM.
- It adds the tract–specific information from the *DistToRoad*, *CommutTime*, and *CommutFrac* direct access files (created in INDEXPOP) to the commuting index file.

COMMUTE Processing Operations

The specific operations performed in the main processing section of COMMUTE are as follows.

- The *DistToRoad* index file (filename extension "*.STIDX*", created in INDEXPOP) is read twice: first to determine the number of records for array allocation and then to populate the arrays with the data in the file
- The *CommutTime* index file (filename extension "*.STIDX*", created in INDEXPOP) is read twice: first to determine the number of records for array allocation and then to populate the arrays with the data in the file
- The *CommutFrac* index file (filename extension "*.STIDX*", created in INDEXPOP) is read twice: first to determine the number of records for array allocation and then to populate the arrays with the data in the file
- The number of data records in the *commuting* file is determined so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- The number of *commuting* file records with home tracts in each state is determined.
- For each state the sequence numbers of the first and last data record indicating a home tract in that state are determined.
- The number of records in the *population* file is read from the *counter* file, so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- All the tract IDs are read from the "*_direct.ind*" file created by INDEXPOP, using the indices from the "*.state_county_pop_range*" and "*.county_tract_pop_range*" files created by INDEXPOP.
- For each tract in the "*_direct.ind*" file created by INDEXPOP, all matching home tracts in the *commuting* file are found. (There is one home tract record for every commuting flow originating in that tract). For each matched home tract, the ID and number of work tracts within 120 km are determined. For each home tract the fractions of total commuting flow to work tracts, which are specified in the *commuting* file, are adjusted to the fractions of the total commuting flow within 120-km.
- For each home-tract/work-tract pair the centroid-to-centroid distance from the *commuting* file is determined and a normalized distance is calculated as distance/(average distance).

- Each work tract ID, its adjusted flow fraction, and its normalized distance is recorded in a sequential file with filename extension "*.da*" (one record for each work tract).
- If no matching home tracts are found in the *commuting* file for a population tract, an entry is recorded in the *mistract* file, indicating the indices of the tract in the "*.state_county_pop_range*", "*.county_tract_pop_range*", and the "*_direct.ind*" files, as well as the tract ID.
- For population tracts with no matching commuting home tracts, a record is recorded in the "*.da*" file indicating the population tract as the work tract, with fractional commuting flow of 1.0, i.e., all work takes place in the home tract.
- For each population tract, a record is written into a temporary index file. The fields in the record are the population tract ID, the sequence numbers of the first and last work tract record in the "*.da*" file, and a flag indicating whether the population tract was matched by a home tract in the *commuting* file (0=no; 1=yes).
- Two records are added to the *counter* file. The first indicates the number of records found in the "*_direct.ind*" file created by INDEXPOP, and the number of data records found in the *commuting* file. The second records the number of records in the "*.da*" file and the number in the "*.ind*" file.
- A sequential index file is created with filename extension "*.st_comm1_fip_range*". For each state there is a record in this file indicating the sequence numbers of the first and last data record for tracts for that state in the temporary index file.
- The temporary index file is rewound and read from the beginning. Each record is matched by tract with a record in the *DistToRoad*, *CommutTime*, and *CommutFra*c direct access files (filename extensions of "*.dat*", created in INDEXPOP). The combined data for each tract are written into a direct access file with the root file name of the *commuting* file and the filename extension "*.ind*".

5.2.4. AIRQUAL

As explained in section 2.1.5, the AIRQUAL program performs three main functions.

- It creates a sequential file of air quality data to be used in HAPEM.
- It determines the number of data records for each census tract in the *air quality* file.
- It creates index files to facilitate file searching in HAPEM.

AIRQUAL Processing Operations

The specific operations performed in the main processing section of AIRQUAL are as follows.

- The number of data records in the *air quality* file is determined so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- The number of time blocks in the *air quality* file is determined from the header record. It is checked for compatibility with the value of ***hblock***, the number of time blocks for the

analysis, as specified in the *parameter* file. As explained in section 2.1.5, ***hblock*** must be an integral multiple of the number of air quality time blocks, so that the air quality values can be replicated if necessary to create ***hblock*** air quality values. If this is not the case, an error message is written to the *log* file and the program is stopped.

- Each data record in the *air quality* file is read and if necessary, the concentration values for each time block are replicated to create ***hblock*** values.
- The concentrations in each record are recorded in a sequential file with the root name of the *air quality* file and the filename extension "*.da*", (e.g., *benzene.da*) to be used in HAPEM.
- The index ranges for the multiple data records in each tract are determined and stored in an index array.
- All the unique county FIPS in the *air quality* file are counted and the values saved into an array.
- The number of records in the *population* file is read from the *counter* file
- An attempt is made to match each population tract specified in the "*_direct.ind*" file created by INDEXPOP with a tract in the *air quality* file. If a match is found, the population array from the "*.da*" file created by INDEXPOP is recorded in a sequential file with the root name of the *population* file and the filename extension "*.pop_air_da*" (e.g., *census.pop_air_da)*. The tract ID (state FIPS, county FIPS, and tract code) and the indices range for data records in a tract from the index array are recorded in a sequential file with the root name of the *air quality* file and the filename extension "*.air_da*", (e.g., *benzene.air_da*). If no match is found, the serial record number of the tract in the "*_direct.ind*" file created by INDEXPOP and the tract ID are recorded in the *mistract* file.
- For each state the number of tracts in the "*.air_da*" file is determined.
- For each county in the "*.air_da*" file the number of tracts is determined
- A sequential index file is created with filename extension "*.state_air_fip_range*". For each county there is a record in this file indicating the serial record numbers of the first and last data records in the "*.pop_air_da*" and "*.air_da*" files.
- A sequential index file is created with filename extension "*.state_air1_fip_range*". For each state there is a record in this file indicating the serial record numbers of the first and last data records in the "*.state_air_fip_range*" file.
- A sequential index file is created with filename extension "*.state_air2_fip_range*". For each state there is a record in this file indicating the serial record numbers of the first and last data records in the "*.pop_air_da*" and "*.air_da*" files.
- Two records are added to the *counter* file. The first record indicates the number of tracts in the "*.pop_air_da*" and "*.air_da*" files, and the number of counties in the "*.state_air_fip_range*" file. The second record indicates the number of census tracts in the *air quality* file and the number of data records in the *air quality* file.

5.2.5. HAPEM

As explained in section 2.1.6, the HAPEM program performs six main functions.

- For each demographic group in each census tract, it randomly selects ***nreplic*** sets of microenvironment (ME) factors based on the distribution data provided in the *factors* file. Each set contains a subset of ME factors randomly selected for each of time blocks (for the *PEN* and *ADD* factors) or each of sources (for the *PROX* factor). Each subset contains randomly selected ME factors for each of ***nmicro*** microenvironments.
- For each demographic group in each census tract, it randomly selects ***nreplic*** sets of air quality data from the data sets available for a census tract.
- For each demographic group in each census tract, it creates ***nreplic*** sets of average activity patterns, where a set contains one average pattern for each day type. An average activity pattern for each day type is calculated as a weighted average of activity patterns randomly selected from each cluster in a demographic group/day type combination. The weights are determined by the relative frequencies of cluster types randomly selected in a one-stage Markov process, based on the cluster transition probabilities provided in the C*lusTrans* file. (A one-stage Markov process is a sequence of events, such that at every step in the Markov chain the probability distribution for the next event depends on what the current event is.)
- For each activity pattern for a commuting demographic group, it randomly selects a work census tract with probability weighting based on the fraction of residents that work in that tract.
- For each census tract it estimates the concentration in each microenvironment based on microenvironment factors and outdoor concentrations.
- It combines activity patterns, commuting, and microenvironment concentration estimates to calculate ***nreplic*** annual average exposure concentrations for each demographic group in each census tract

HAPEM Processing Operations

The specific operations performed in the main processing section of HAPEM are as follows.

- The distribution data of microenvironmental (ME) factors for each of ***nmicro*** microenvironments is read from the *factors* and *mobiles* files and saved into arrays.
- For the onroad mobile source *PROX* distributions in the *mobiles* file, the average *PROX* factor for the second distance category over all the indoor microenvironments is calculated. (This value will be used later to calculate the concentration ambient concentration for the third distance category, as described below.)
- For each demographic-group/day-type/commuting-status combination the number of activity patterns for each cluster is read from the "*.nonzero*" file created in DURAV.
- For each demographic-group/day-type/commuting-status combination, the frequency of each cluster, and the cluster-to-cluster transition probabilities are read from the *ClusTrans* file.

- For each demographic-group/day-type/commuting-status/cluster combination with a positive number of activity records, the activity pattern records are read from the "*.da*" file created in DURAV and the values saved into an array.
- Each activity pattern is checked to ensure a total activity time of 1440 minutes. If not, an error message is written to the *log* file and the program is stopped.
- Several values are read from the *counter* file to allocate memory for various arrays.
- Indices are read from the "*.state_air_fip_range*" and "*.state_air1_fip_range*" files created by AIRQUAL.
- Data are read from the "*.pop_air_da*" and the indices ranges for air records from "*.air_da*" files created by AIRQUAL.
- Air data records are read from ".da " files created by AIRQUAL.
- Indices are read from the "*.st_comm1_fip_range*" and "*.ind*" files created by COMMUTE, and data from the "*.da*" file created by COMMUTE.
- For each tract in the "*.ind*" file created by COMMUTE an attempt is made to find a matching tract in the "*.state_air_fip_range*" file created by AIRQUAL. If a match is not found, the commuting tract is recorded in the *mistract* file.
- For each demographic group in each census tract, ***nreplic*** sets of microenvironmental (ME) factors are randomly selected based on the distribution data provided in the *factors* and *mobiles* files, using subroutines DISTRIBUTION and DATASET. Each set contains a subset of ME factors randomly selected for each of time blocks (for the *PEN* and *ADD* factors) or each of sources (for the *PROX* factor). For onroad mobile source categories, first a distance-from-source category is selected for each indoor ME based on the population fractions in each distance category that were taken from the *DistToRoad* file and added to the *commuting* index file in COMMUTE[21]. Then a *PROX* factor for each indoor ME is selected from the appropriate distribution. Each subset contains randomly selected ME factors for each of ***nmicro*** microenvironments.
- For each demographic group in each census tract, ***nreplic*** sets of air quality data are randomly selected from the data sets available for the census tract in the "*.da*" file created by AIRQUAL.
- When a single set of ambient concentrations are provided for each tract in the *air quality* file (as is typically the case) they represent spatial averages over the tract, excluding locations in the very near vicinity of an emission source. For onroad mobile source categories it is assumed that the ambient concentrations in the *air quality* file represent spatial averages over the second and third distance categories for the *DistToRoad* and *mobiles* files (75-200 meters and greater than 200 meters). Because HAPEM estimates the ambient concentration for the second distance category by applying a *PROX* factor to the "tract average" ambient concentration, the ambient concentration for the third distance category is

[21] It is assumed that the spatial distribution of all indoor MEs in a tract with respect to distance from major roadways is the same as for residences.

also adjusted to make the area-weighted average over the two distance categories equal to the “tract average”. This is done as follows.

$$AREA_{D3} \times CONC_{D3} + AREA_{D2} \times CONC_{D2} = CONC_{AQ} \qquad \text{or}$$

$$AREA_{D3} \times CONC_{D3} + AREA_{D2} \times PROX_{D2} \times CONC_{D3} = CONC_{AQ} \qquad \text{or}$$

$$CONC_{D3} = {CONC_{AQ}} \Big/ {(AREA_{D3} + AREA_{D2} \times PROX_{D2})}$$

where

CONC

$CONC_{AQ}$ = the “tract average” concentration from the *air quality* file

$CONC_{D2}$ = average ambient concentration in second distance category

$CONC_{D3}$ = average ambient concentration in third distance category

$AREA_{D2}$ = fraction of the tract area in the second distance category (from the *DistToRoad* file)

$AREA_{D3}$ = fraction of the tract area in the second distance category (from the *DistToRoad* file), and

$PROX_{D2}$ = average PROX factor for the second distance category over all the indoor source categories (calculated above)[22].

- The randomly selected air quality data from the “*.da*” file created by AIRQUAL for each matched tract is combined with the randomly selected ME factors to estimate the concentrations for each ME/time block combination for that tract.
- For each demographic group in each census tract, the background exposure concentration contributions are calculated for each ME/ time block combination based on the uniform value of the ***backg*** parameter specified in the *parameter* file, variable background concentration values for each data record in *“.da”* file created by AIRQUAL, and the randomly selected ME factors.
- For each census tract/demographic-group/day-type replicate a commuting status is selected based on the data from the *CommutFrac* file that were added to the *commuting* index file in COMMUTE. If the replicate is a commuter, a commuting mode (public or private transit) is randomly selected based on the data from the *CommutTime* file that were added to the

[22] As implied by the equations above, the onroad mobile source *PROX* factor distributions are estimated as the ratios between the near-roadway concentration and the concentration distant from the roadway, rather than the ratios between the near-roadway concentration and the “tract average” concentration.

commuting index file in COMMUTE. This selection also determines an associated average commuting time for the tract.

- For each replicate that commutes, a work tract is randomly selected for each selected activity pattern, using an attached subroutine, RANDOMR. The work tract is selected from the set of work tracts corresponding to that home tract, as specified in the "*.da*" file created by COMMUTE. The air quality data for that work tract are randomly selected from the data sets available for the work tract in the "*air_da*" file created by AIRQUAL. If the work tract cannot be found in the "*.air_da*" file, the air quality data for the home tract are used. The air quality data are adjusted and combined with the ME factors randomly selected in the same way as the home tract to estimate the concentrations for each ME/time block combination for that work tract.
- For each replicate/day-type combination an average activity pattern is calculated as the weighted average of activity patterns randomly selected from each cluster in a demographic-group/day-type/commuting-status combination in the ".da" file created in DURAV. The weights are determined by the relative frequencies of cluster types randomly selected in a Markov process, based on the cluster transition probabilities provided in the *ClusTrans* file.
- The average activity pattern for the day-type is adjusted so that the commuting time for the replicate is equal to the product of the tract average commuting time for the commuting mode selected above, and the normalized home-tract/work-tract distance calculated in COMMUTE and recorded in the *commuting* direct access file created in COMMUTE. The adjustments are made by uniform scaling of the time in each time block for commuting MEs so that the sum matches the total calculated commuting time, and corresponding uniform scaling of the time in each time block for non-commuting MEs.
- The ME/time block time durations of the weight-averaged activity patterns are combined with the estimated ME/time block concentrations for the home tract and the work tracts to estimate ***nreplic*** exposure concentrations for each demographic-group/day-type combination. A separate set of estimates is made for each emission source category. The algorithm for each demographic-group/day-type combination in the tract is as follows.

$$ExpConc = \frac{\sum_{TimeBlocks}\sum_{Microenviroments} Conc_{t,m} \times Duration_{t,m}}{\sum_{TimeBlocks}\sum_{Microenviroment} Duration_{t,m}}$$

where $Conc_{t,m}$ is the emission source category concentration during time block *t* in microenvironment *m*; and

$Duration_{t,m}$ is the duration of activity during time block *t* in microenvironment *m.*

- The exposure concentrations for each day type are combined with weighted averaging to create an annual average exposure concentration. The weights are the relative frequencies of the day types: 0.181 for summer weekday, 0.534 for non-summer weekdays, and 0.285 for weekends.

- A total annual average exposure concentration is calculated by adding the annual average values for each emission source category, from the background contribution, and from the indoor source *ADD* factor.
- The results are written into the final exposure output files, ***nreplic*** records for each demographic group in each tract. The format of the files is described in section 4.4.

6. References

Glen, G., Y. Lakkadi, J.A. Tippett, and M. del Valle-Torres (Prepared by ManTech Environmental Technology, Inc.), 1997: Development of NERL/CHAD: The National Exposure Research Laboratory Consolidated Human Activity Database. EPA Contract No. 68-D5-0049.

This page intentionally left blank.

Appendix A: Development of Algorithm for Creating Longitudinal Activity Patterns with Cluster Analysis

This page intentionally left blank.

ICF CONSULTING

MEMORANDUM

To: **Ted Palma, US EPA**

From: Jonathan Cohen and Arlene Rosenbaum

Through: Rebecca Battye, ECR, Inc.

Date: July 23, 2002

Re: Proposed modification of HAPEM algorithm for creating longitudinal activity patterns: Results of data analysis.

Summary

The purpose of this task was to review the current modeling approach for developing annual average activity patterns from the CHAD database and recommend ways to improve the model's pattern selection process.

The data analysis grouped the CHAD daily activity patterns into either two or three categories of similar patterns for each of the 30 combinations of day type (summer weekday, non-summer weekday, and weekend) and demographic group (males or females; age groups: 0-4, 5-11, 12-17, 18-64, 65+). Under the proposed modification of the HAPEM algorithm, for each day type and demographic group, one daily activity pattern per category is randomly selected from the corresponding CHAD data to represent that category. The starting category is selected according to the relative frequencies of each category. The category for the second day is selected according to the transition probabilities from the starting category, which are the relative frequencies of each category among those days where the same individual was observed on the previous day and the previous activity pattern was in the starting category. The category for the third day is selected according to the transition probabilities from the second day's category. This is repeated for all days in the day type, producing a sequence of daily categories. For each day, the activity pattern is then given by the chosen representative activity pattern for that day's category.

Background

The approach used for the preliminary NATA simulations selected with replacement (365) 24-hour activity patterns for each demographic group in each census tract, with the patterns stratified by day-of-week and season. These were averaged together to create three averaged activity patterns for each group/tract combination: 65 Summer weekdays, 195 non-Summer weekdays, and 104 weekend days. The variability resulting for this approach represented uncertainty for the average activity pattern for the group/tract combination, rather than the variability of activity patterns among group members.

In response to SAB comments this approach was modified to try to represent the variability among individuals within a group/tract combination. For each group/tract combination three group-specific activity patterns were selected, one for each day-of-week/season combination. This approach implied that for any individual, the activity pattern is identical for every day in a day-of-week/season category, i.e., probability of transition to a different pattern equal zero. This approach tends to maximize the differences between individuals, perhaps to an unrealistic extent.

Proposed New Algorithm

To improve this approach we propose to treat transition probabilities in more detail. Information on the probability of changes among daily activity patterns for a single individual, stratified by day type could be used. For example, for some demographic group/day type combination suppose activity patterns could be grouped into two categories, A and B, based on differences among the times spent in various microenvironments. Further suppose that we could estimate the probability of transition from a type A pattern to a type B pattern, and from a type B pattern to a type A pattern. That is, suppose we could quantify

P_{AA}: probability that a type A pattern is followed by a type A pattern

P_{AB}: probability that a type A pattern is followed by a type B pattern ($P_{AB} = 1 – P_{AA}$)

P_{BB}: probability that a type B pattern is followed by a type B pattern

P_{BA}: probability that a type B pattern is followed by a type A pattern ($P_{BA} = 1 – P_{BB}$)

Then the HAPEM algorithms could be modified as follows to create an activity pattern sequence for an individual for a given day type, e.g., non-Summer weekdays.

1. For day 1 randomly select an initial activity pattern for non-Summer weekdays, type X
2. For the following day retain the same activity pattern with probability P_{XX}, or randomly select a type Y pattern with probability P_{XY}
3. If the activity pattern selected in the previous step is type X, repeat step 2 to find the activity pattern for the following day. If the activity pattern selected in the previous step is type Y retain the same activity pattern for the following day with probability P_{YY} and switch back to the type X pattern from day 1 with probability P_{YX}. In all cases, once a

type A or type B activity pattern has been selected, use that same pattern for each subsequent day that requires that activity pattern type.

4. Repeat step (3) until the desired number of activity patterns are selected, e.g., 193 for non-summer weekdays.

The averages of the selected activity patterns would then be used to evaluate the individual's exposure for non-Summer weekdays. This algorithm could be generalized to any number of activity pattern categories, as long as the transition probabilities can be quantified.

Use of a single activity pattern to represent each of the day types in the sequence will tend to minimize the mixing of activity patterns from different individuals while still accounting for some of the typical day-to-day variability for an individual. In lieu of data on long sequences of activity patterns for single individuals, we believe that this approach represents a reasonable compromise between over- and under-estimation of inter-individual differences in activity patterns with respect to factors that are likely to have an important influence on long-term average exposure.

Grouping Days Into Categories

The first data analysis task was to use the CHAD data to group activity pattern days into categories for each combination of day type and demographic group. First, each daily activity pattern was summarized by the total minutes in each of five micro-environments: indoors – residence; indoors – other building; outdoors – near road; outdoors – away from road; in vehicle. These five numbers are assumed to represent the most important features of the activity pattern for their exposure impact. (The five numbers are not independent since they sum to 1440). Each day type and demographic group was analyzed separately. In statistical terminology, grouping cases into categories based only on similarities or differences between measurements on those cases is referred to as classification or cluster analysis and the categories are called “clusters.” A cluster analysis was used to group the activity pattern days in each day type and demographic group into clusters of days with similar values for the minutes in each of the five microenvironments, a five-dimensional “time-spent” vector. The chosen analysis treated the time-spent vectors for different days and/or individuals as statistically independent, although, in principle, a complex statistical approach might take into account dependencies between the time-spent vectors for the same individual on different days.

The CHAD database was originally grouped into three day types, each with ten demographic groups (based on age group and gender). We reviewed the available data for estimating transition probabilities (i.e. data with more than one day per individual) and found that for the lowest (0-4) and highest (65 +) age groups, there were very limited numbers of consecutive day pairs, which would have led to very imprecise estimates of the transition probabilities. To reduce this problem, we decided to regroup the demographic groups so that males and females 0-4 were put into one group and males and females 65 + were put into another group. This corresponds to the assumption that the male and female activity patterns are approximately similar for those age groups, for each day type. This approach applies only for the purpose of allocating activity patterns to the clusters and estimating the transition probabilities. For the rest of the HAPEM model we recommend retaining the uncombined demographic group groupings, so that in particular, transitions between male and female activity patterns within the 0-4 or 65 + age groups would not be permitted. The following analyses are based on the reduced set of 24 day type and demographic group combinations.

There are dozens of possible methods of cluster analysis in the literature. In principle, the "best" method for a given problem depends on assumptions about the joint statistical distribution of the vector of measurement variables, which in turn give the expected shape of the clusters (using one dimension for each measurement variable), e.g., whether they are symmetric or elongated in one or more directions. If the number of clusters is given in advance, then several methods can be used. For example, the *k-means* method is designed to choose k clusters to minimize the total of the squared Euclidean distances from each vector to its cluster centroid vector. Since the number of possible configurations (groupings of the vectors into k clusters) is huge, the k-means algorithm chooses an initial set of k cluster seeds (points in the n-dimensional space of measurement vectors), assigns each case vector to the nearest cluster seed, redefines the cluster seeds as the new cluster centroids, and then repeats the last two steps until convergence. If the initial seeds are well chosen, then the cluster seeds will converge to a global minimum solution for the total squared Euclidean distance (from each case to its assigned cluster centroid). Initially, the k-means method with various values of k was applied to the CHAD data, but the results were not very useful because the method frequently produced very unbalanced cluster sizes, with some clusters having many cases and some clusters having just 1 or 2 cases. Applying these small clusters to HAPEM would likely lead to unstable model predictions, even if there were enough CHAD data for consecutive days to estimate the transition probabilities (described in the following section).

If the number of clusters is not given in advance, hierarchical, agglomerative clustering algorithms can be used. Starting with n cases, each case vector is initially assigned to its own cluster, producing n clusters. At the next stage, two nearby clusters are joined, producing n-1 clusters. This is repeated until the n'th stage, which has 1 cluster consisting of all n cases. Using the hierarchical approaches, the result is a tree structure showing at each stage which smaller clusters were joined together. The hierarchical methods differ by their definition of a "nearby" cluster. For example, *single linkage* defines the distance between clusters as the minimum Euclidean distance between pairs of vectors (one from each cluster), and *average linkage* uses the average distance, or average squared distance, between pairs of vectors. Another commonly used method, *Ward's method*, defines the distance as the squared Euclidean distance between the cluster centroids divided by (1/m + 1/n), where m and n are the numbers of cases in each cluster. Using Ward's method, at each stage in the hierarchy, the clusters to be joined are chosen to minimize the sum of squared Euclidean distances between cases and their cluster centroids. One advantage of Ward's method for the CHAD data is its tendency to produce clusters with roughly the same numbers of cases. Ward's method was chosen for these analyses.

An important consideration is whether or not to rescale the measurement variables before applying the clustering algorithm. If the different measurements are in different units (e.g. inches and feet, or inches and seconds), then rescaling is usually recommended to make the different variables comparable. For example, without rescaling, a measurement recorded in inches will have a much bigger impact on the clustering than the same measurement recorded in feet, assuming distances are defined using (equally weighted) Euclidean distances. The typical rescaling of each measurement variable subtracts the overall mean and then divides by the overall standard deviation, producing a new variable with mean zero and standard deviation one. If all measurements are in the same units, as in the present case (minutes in a micro-environment), then the statistical literature is less definitive on the need for rescaling. A classic textbook, *Clustering Algorithms* (Hartigan, J. A., Wiley, 1975) points out that rescaling to a constant variance often tends to downweight variables that cluster well. For these analyses, the five measurement variables were not rescaled.

After applying Ward's method to the CHAD data, the number of clusters needed to be chosen for each day type and demographic group. If the measurement variables are uncorrelated, then various statistical measures (e.g., pseudo-F statistic, pseudo-t^2 statistic, cubic clustering criterion) have been developed for use in determining the optimum number of clusters. For these analyses the five measurement variables are correlated (since they sum to 1440) and so the various statistical stopping rules could not be applied.[23] An important consideration for these analyses was the need for sufficient data on consecutive days to develop the transition probabilities, since most of the CHAD data had just one activity pattern day per individual and day type. On this basis, three clusters were chosen for 24 of the 30 demographic group and day type combinations. For the other 6 combinations, 2 clusters were chosen because otherwise there would have been no pairs or only one pair of consecutive day activity patterns available to estimate a transition probability.

The result of the Ward method cluster analysis was an assignment of every CHAD activity pattern day to a cluster, where each day type and demographic group had either two clusters (6 combinations) or three clusters (24 combinations). The Excel spreadsheet finaltree.xls gives the assigned cluster number (1, 2, or 3) for each CHADID and also includes the demographic group (original ten and recombined set of eight), day type, and number of clusters (Ncluster) for that day type and demographic group.

Estimating Transition Probabilities

In this step the transition probabilities were estimated from the clustered CHAD data. First, we extracted from each demographic group and day type all cases where an individual had a recorded activity pattern for two or more consecutive days. Define the following variables for each day type and demographic group:

trans*ij* = number of pairs of consecutive days for the same individual where the first day is in cluster *i* and the next day is in cluster *j*.

trans*i*x = number of pairs of consecutive days for the same individual where the first day is in cluster *i*.

= trans*i*1 + trans*i*2 + trans*i*3

prob*ij* = trans*ij* / trans*i*x

There are trans*i*x days where an individual is in cluster *i* on one day and where the next day is in the database. Of those trans*i*x days, the next day is in cluster *j* trans*ij* times. Therefore, prob*ij* is an estimate of the transition probability from cluster *i* to cluster *j*.

[23] *An alternative approach would have been to just use four of the measurement variables. This would have reduced the correlation problem but not removed it since the sum of the four is bounded above and below. Further, the results would then have depended upon which variable was not used. As a sensitivity study, the cluster analysis was repeated using all but the time in residence, which is usually where the greatest exposure time occurs. It was found that in many cases the vectors were assigned to the same clusters – more precisely, a vector assigned to the most populous cluster using all five variables would frequently also be assigned to the most populous cluster using the four variables, and similarly for the second most populous and least populous clusters.*

In a few cases, the estimated transition probability was zero. Although it is possible that the associated transitions cannot occur, it is more plausible that the true transition probabilities are small but non-zero, and these zero estimated probabilities were obtained because of the limited number of transition pairs. We therefore decided to replace each estimated zero probability by 0.5 / trans*i*x, which is one half the minimum observable non-zero probability, and subtract that probability equally from the one or two remaining non-zero values for the same *i*.

The results of this analysis are given in the spreadsheet finaltrans.xls that includes the transition counts and estimated transition probabilities for each day type and uncombined demographic group. For each day type, these counts and probabilities are the same for the male and female 0-4 and the male and female 65 + demographic groups, which were combined for the cluster and transition probability analyses. For the five combinations with only 2 clusters, the values with *i* or *j* equal to 3 are missing or zero, since cluster 3 is undefined. Also included are the variables cluster1, cluster2, and cluster3, giving the total numbers of CHAD activity patterns in clusters 1, 2, and 3, respectively.

Possible Algorithm Simplification

The algorithm described above may be characterized as a Markov chain model. Because for this application we are only interested in the average of the selected activity patterns and not their sequence, we considered whether it may be possible to simplify the algorithm considerably by applying well-established concepts from Markov chain theory. That is, we can estimate the expected fractions of days in each cluster for a lengthy sequence of selections analytically, based only on the transition probabilities, provided that the Markov chain converges to a steady state. We can then calculate the activity pattern average by simply selecting one pattern for each cluster and averaging them together with the calculated fractions as weights. The lengths of sequences for this application are 65 summer weekday, 104 weekend days, and 195 non-Summer weekdays. Whether steady state ratios exist and whether the sequences are long enough to converge to the steady-state fractions depends on the transition probabilities. In particular, if all the transition probabilities are not equal to zero or one, then the chain is irreducible, aperiodic, and recurrent, so that steady state probabilities exist.

Since we chose to replace all zero estimated transition probabilities by a suitably small positive number, the chains for each day type and demographic group are irreducible, aperiodic, and recurrent, so that steady state, limiting probabilities exist. These steady state probabilities are found by solving the linear equations:

steady*j* = limiting probability for cluster *j* = Σ_i steady*i* × prob*ij*

In a very long sequence of days, the proportions of days in each cluster will tend to the steady state probabilities, assuming the day to day transitions occur with the assigned transition probabilities. However, an analysis of simulated seasons for each day type and demographic group shows that the numbers of days per cluster varies significantly around the limiting value. For each day type and demographic group, the linear equations were solved for the steady state probabilities and 1000 sequences of daily clusters were simulated for each of the possible starting day clusters (either two or three). The resulting distribution of the numbers of days per cluster are shown in the spreadsheet sims.xls:

expdays*i* = steady state expected number of days for cluster *i*

low*i* = 5th percentile of the simulated number of days for cluster *i* days*i* = mean number of simulated days for cluster *i*

high*i* = 95th percentile of the number of days for cluster *i*

Although the steady state expected number of days in each cluster is very close to the mean number of days across all simulations, the variation around the mean value is quite large. (The cluster for a given day would be statistically independent of the cluster for the previous day if, for each *j*, prob*ij* is the same for all *i*. In this special case the number of days in each cluster would have a binomial distribution. For the independent case, the number of days in each cluster will also vary among simulated seasons, but the variation is generally greater for the non-independent cases simulated here.) This analysis leads to our recommendation that the proposed model revision uses the transition probabilities to directly simulate the cluster transitions, instead of using the steady state estimates of the number of days per cluster.

FINALTRANS – PART 1

DayType	DemographicType	trans11	trans12	trans13	trans21	trans22	trans23	trans31	trans32	trans33
1	1	4	0	0	0	2	0	0	0	2
1	2	5	5	1	2	29	4	0	2	4
1	3	12	1	1	2	1	0	1	0	13
1	4	23	9	2	10	64	2	2	0	1
1	5	11	1	0	2	1	0	0	0	0
1	6	5	2	0	1	8	0	0	0	2
1	7	18	2	0	1	6	0	0	0	0
1	8	32	3	0	2	9	1	2	0	1
1	9	90	7	10	8	38	19	8	16	75
1	10	21	0	2	0	2	1	3	1	1
2	1	15	1	1	1	3	0	0	1	3
2	2	5	1	0	2	13	3	0	2	1
2	3	3	2	1	1	5	6	0	9	11
2	4	35	18	2	9	102	5	2	5	14
2	5	10	4	0	1	3	0	0	0	0
2	6	7	5	0	0	11	1	0	0	6
2	7	4	1	1	0	8	2	2	2	1
2	8	40	1	7	0	9	1	6	0	9
2	9	209	12	13	16	212	15	12	16	77
2	10	58	2	0	5	18	0	0	1	3
3	1	1	1	0	3	2	0	0	0	0
3	2	10	1	3	1	0	1	2	0	1
3	3	12	1	3	2	1	1	4	0	2
3	4	36	2	5	5	3	1	9	2	8
3	5	5	1	0	0	1	0	0	0	0
3	6	0	1	0	6	9	0	0	0	0
3	7	4	3	1	5	2	1	2	1	0
3	8	12	0	5	5	5	2	1	0	3
3	9	29	3	17	1	31	5	12	6	127
3	10	3	1	1	2	4	0	0	1	6

FINALTRANS – PART 2

DayType	DemographicType	prob11	prob12	prob13	prob21	prob22	prob23	prob31	prob32	prob33
1	1	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
1	2	0.45	0.45	0.09	0.06	0.83	0.11	0.00	0.33	0.67
1	3	0.86	0.07	0.07	0.67	0.33	0.00	0.07	0.00	0.93
1	4	0.68	0.26	0.06	0.13	0.84	0.03	0.67	0.00	0.33
1	5	0.92	0.08	0.00	0.67	0.33	0.00			
1	6	0.71	0.29	0.00	0.11	0.89	0.00	0.00	0.00	1.00
1	7	0.90	0.10	0.00	0.14	0.86	0.00			
1	8	0.91	0.09	0.00	0.17	0.75	0.08	0.67	0.00	0.33
1	9	0.84	0.07	0.09	0.12	0.58	0.29	0.08	0.16	0.76
1	10	0.91	0.00	0.09	0.00	0.67	0.33	0.60	0.20	0.20
2	1	0.88	0.06	0.06	0.25	0.75	0.00	0.00	0.25	0.75
2	2	0.83	0.17	0.00	0.11	0.72	0.17	0.00	0.67	0.33
2	3	0.50	0.33	0.17	0.08	0.42	0.50	0.00	0.45	0.55
2	4	0.64	0.33	0.04	0.08	0.88	0.04	0.10	0.24	0.67
2	5	0.71	0.29	0.00	0.25	0.75	0.00			
2	6	0.58	0.42	0.00	0.00	0.92	0.08	0.00	0.00	1.00
2	7	0.67	0.17	0.17	0.00	0.80	0.20	0.40	0.40	0.20
2	8	0.83	0.02	0.15	0.00	0.90	0.10	0.40	0.00	0.60
2	9	0.89	0.05	0.06	0.07	0.87	0.06	0.11	0.15	0.73
2	10	0.97	0.03	0.00	0.22	0.78	0.00	0.00	0.25	0.75
3	1	0.50	0.50	0.00	0.60	0.40	0.00			
3	2	0.71	0.07	0.21	0.50	0.00	0.50	0.67	0.00	0.33
3	3	0.75	0.06	0.19	0.50	0.25	0.25	0.67	0.00	0.33
3	4	0.84	0.05	0.12	0.56	0.33	0.11	0.47	0.11	0.42
3	5	0.83	0.17	0.00	0.00	1.00	0.00			
3	6	0.00	1.00	0.00	0.40	0.60	0.00			
3	7	0.50	0.38	0.13	0.63	0.25	0.13	0.67	0.33	0.00
3	8	0.71	0.00	0.29	0.42	0.42	0.17	0.25	0.00	0.75
3	9	0.59	0.06	0.35	0.03	0.84	0.14	0.08	0.04	0.88
3	10	0.60	0.20	0.20	0.33	0.67	0.00	0.00	0.14	0.86

SIMS.XLS

DayType	Demographic group (uncombined)	expdays1	low1	days1	high1	expdays2	low2	days2	high2	expdays3	low3	days3	high3
1	1	39.00	17	37.58	57	15.29	2	15.89	32.5	10.71	1	11.53	25
1	2	6.81	2	7.33	14	43.33	32	42.66	53	14.86	6	15.01	25
1	3	29.29	10	29.60	49	4.29	1	4.61	9	31.43	10	30.79	52
1	4	21.89	11	22.10	34	39.99	26	39.42	52	3.12	0	3.47	7
1	5	28.77	12	28.41	46	28.97	14	28.79	44	7.27	1	7.80	17
1	6	39.00	17	37.77	57	15.29	3	15.91	32	10.71	1	11.32	25
1	7	38.24	20	37.68	54	26.76	11	27.32	45		0	0.00	0
1	8	46.05	30.5	45.31	58	16.26	5	16.60	31	2.68	0	3.09	7
1	9	24.41	9	23.99	41	14.14	5	14.52	24	26.46	13	26.50	41
1	10	28.77	13	28.42	45	28.97	14	28.72	44	7.27	1	7.86	17
2	1	29.03	17	29.53	43	100.90	73	100.19	125	65.07	38	65.28	96
2	2	72.84	49	72.97	98	93.33	71	92.90	115	28.83	20	29.13	39
2	3	18.36	8	18.73	31	81.70	71	81.76	93	94.94	81	94.50	108
2	4	35.16	20	35.28	53	138.14	117	137.58	158	21.70	8	22.14	38
2	5	148.10	126	146.78	166	40.59	24	41.29	60	6.31	1	6.94	15
2	6	29.03	16	29.37	44	100.90	75	100.50	125	65.07	37	65.13	95
2	7	56.95	37	56.88	78	103.54	81	103.47	126	34.51	25	34.65	44
2	8	116.61	87	115.56	141	31.85	7	32.80	65	46.55	30	46.64	64
2	9	84.19	50	84.04	121	75.74	42.5	75.60	109	35.07	17	35.36	57
2	10	148.10	125	146.42	166	40.59	25	41.50	60	6.31	1	7.08	16
3	1	60.97	54	60.93	68	43.03	36	43.07	50		0	0.00	0
3	2	66.94	57	66.40	76	11.95	6	12.26	19	25.10	18	25.34	33
3	3	73.06	63	72.50	81	8.57	4	8.91	15	22.37	16	22.58	30
3	4	78.46	67	77.84	88	8.21	3	8.57	15	17.33	9	17.59	27
3	5	49.92	35	50.02	65	54.08	39	53.98	69		0	0.00	0
3	6	60.97	54	60.99	68	43.03	36	43.02	50		0	0.00	0
3	7	57.28	50	57.02	65	33.16	26	33.10	40	13.57	8	13.88	20
3	8	44.96	32	45.26	58	12.29	6	12.45	20	46.75	34	46.29	59
3	9	14.97	6	15.15	27	22.60	5	23.49	46	66.43	45	65.36	84
3	10	49.92	35	49.90	65	54.08	39	54.10	69		0	0.00	0

Appendix B: Estimating near roadway populations and areas for HAPEM6

This page intentionally left blank.

ICF
CONSULTING

MEMORANDUM

To: Chad Bailey

From: Arlene Rosenbaum and Kevin Wright

Date: December 28, 2005

Re: Estimating near roadway populations and areas for HAPEM6

PURPOSE AND BACKGROUND

In its 2001 regulation of mobile source air toxics (the “MSAT Rule”) EPA’s Office of Transportation and Air Quality (OTAQ) committed to further study of the range of concentrations to which people are exposed for consideration in future rulemaking. As part of the Technical Analysis Plan outlined in that research, OTAQ undertook research activity looking at the air quality in immediate proximity of busy roadways and highways. Concentrations of pollutants directly emitted by motor vehicles show statistically significant elevation in concentrations with increased proximity to busy roadways.

The Hazardous Air Pollutant Exposure Model (HAPEM) is a screening-level exposure model appropriate for assessing average long-term inhalation exposures of the general population, or a specific sub-population, over spatial scales ranging from urban to national. HAPEM uses the general approach of tracking representatives of specified demographic groups as they move among indoor and outdoor microenvironments and among geographic locations. The estimated pollutant concentrations in each microenvironment visited are combined into a time-weighted average concentration, which is assigned to members of the demographic group.

Indoor microenvironment concentrations are estimated by applying scalar factors to outdoor tract concentrations, which are some of the required inputs. These scalar factors are derived from published studies of concurrent concentration measurements indoors and outdoors.

In the previous version, HAPEM5, if only a single outdoor concentration is provided for each Census tract, as is typical, this concentration is assumed to uniformly apply to the entire Census tract. For this version, HAPEM6, we refined the model to account for the spatial variability of outdoor concentrations within a tract due to enhanced outdoor concentrations of onroad mobile source pollutants at locations near major roadways. The term “major roadway” is used to describe a “Limited Access Highway”, “Highway”, “Major Road” or “Ramp”, as defined by the Census Feature Class Codes (CFCC). The new version of HAPEM more accurately reflects the average and variability of exposure concentrations within each Census tract by accounting for some of the spatial variability in the outdoor concentrations within the tract, and by extension some of the spatial variability in indoor concentrations within the tract.

Accomplishing this refinement to HAPEM required several activities, including the development and implementation of an approach for creating a database of the fraction of people within each US Census tract living near major roadways. This memorandum describes that activity.

OVERVIEW AND SPECIFICATIONS

The objective of this task was to estimate the fraction of people in each of 6 demographic groups in each US Census tract living near major roadways.

The basic analysis was conducted at the US Census block level for populations stratified by age, gender, and race/ethnicity. The block level data was then aggregated up to the tract level for populations stratified by age only for use in HAPEM6.

The data bases used for this task were:

- The Environmental Sciences Research Center (ESRI) StreetMap US roadway geographic database (which includes NavTech, GDT and TeleAtlas rectified street data)
- A geographic database of US Census block boundaries, extracted using the PCensus 2000 Census data extraction tool for Census file SF1
- A geographic data for US Census block boundaries in Puerto Rico and the US Virgin Islands obtained from Proximity

Although the block file is an intermediate product for this project, it will be retained to facilitate the re-specification of demographic groups for possible future analyses. Therefore, this file contains the most resolved age-gender groups available at the block level from the US Census STF1. The age groups for the block level data are as follows:

- 19 single-year age groups from 0-19 (P14)
- 2 single-year age groups 20-21 (P12)
- 16 age groups (P12)
    - 22 to 24 years
    - 25 to 29 years
    - 30 to 34 years
    - 35 to 39 years
    - 40 to 44 years
    - 45 to 49 years
    - 50 to 54 years
    - 55 to 59 years
    - 60 and 61 years
    - 62 to 64 years
    - 65 and 66 years

  - 67 to 69 years
  - 70 to 74 years
  - 75 to 79 years
  - 80 to 84 years
  - 85 years and over.

The aggregated age groups for the tract level data are:

- 0-1
- 2-4
- 5-15
- 16-17
- 18-64
- 65+

The race/ethnic groups (block level only) are:

- non-Hispanic White (alone or in combination - P010003)
- non-Hispanic Black (alone or in combination - P010004)
- non-Hispanic American Indian /Alaskan Native (alone or in combination - P010005)
- non-Hispanic Asian (alone or in combination - P010006)
- non-Hispanic Native Hawaiian/ Pacific Isalander (P010007)
- non-Hispanic other (alone or in combination - P010008)
- Hispanic (alone or in combination - P010009)

The spatial stratifications of the populations (block and tract level) are:

- Those residing within 75 meters of a major roadway
- Those residing from 75 to 200 meters from a major roadway
- Those residing at greater than 200 meters from a roadway.

In addition, the fraction of the area of each Census block and tract that is located within the same distance ranges from a major roadway was determined.

PROCEDURES

For all the spatial modeling and geoprocessing operations in this study ICF utilized ArcInfo software. ArcInfo is the most extensive version of ArcGIS 9.1, the industry's standard for Geographic Information Systems, produced by ESRI of Redlands, CA.

Due to the size of the roadway and block geography files, most of the processing was conducted on a county-by-county basis. The files for some counties, however, still exceeded ArcInfo's capacity and were processed tract-by-tract. A few counties in Arizona needed special handling because even at the tract level they exceeded ArcInfo's capacity and were disaggregated into smaller pieces for processing.

1. Because populations are not generally evenly distributed within blocks, it was assumed that the block populations all reside within 150 meters of *any* road within the block of designation "local" or greater as defined by the Census Feature Class Codes (CFCC). Thus, the first step was to create a 150-meter buffer around all roadways within the block. This buffer served as a "clipped" block boundary defining the portion of the block containing residential populations. The block population was assumed to be uniformly distributed within the "clipped" block boundary.

2. Next a 75-meter buffer and a 200-meter buffer were created around all major roadways within the block. These buffers were overlaid on the "clipped" block boundary, and the fraction of the "clipped" block area that that fell within each buffer was calculated. This area fraction was assumed to equal the population fraction that fell within each buffer, and the fractions were applied to each population stratification.

3. The 75-meter buffer and the 200-meter buffer were also overlaid on the unclipped block boundary to determine the fraction of the total block area that fell with each of the buffers.

4. The block level fractions for area and populations were then aggregated up to the tract level, and the population stratifications were aggregated up to the 6 tract age groups only.

RESULTS

The resulting database consists of 2 files types: (1) a block file for each state, and (2) a nation-wide tract file.

The block files contains the following 249 fields for each block:

- block FIPS code
- total population
- total area
- area within 75 meters of a major roadway
- area from 75 to 200 meters from a major roadway

- for each of 74 age-gender groups:
    - population residing within 75 meters of a major roadway
    - population residing between 75 and 200 meters from a major roadway
    - population residing more than 200 meters from a major roadway
- sum of race/ethnic populations (note; this may differ slightly from the total population due to some double-counting of persons with more than 1 race/ethnicity)
- for each of 7 race/ethnic groups:
    - population residing within 75 meters of a major roadway
    - population residing between 75 and 200 meters from a major roadway
    - population residing more than 200 meters from a major roadway

Note that because of the limitations of the US Census data the block level populations could not be stratified by age, gender, and race together,

The tract file contains the following 22 fields for each tract

- tract FIPS code
- fraction of area within 75 meters of a major roadway
- fraction of area between 75 and 200 meters from a major roadway
- fraction of area more than 200 meters from a major roadway
- for each of 6 age groups:
    - fraction of population residing within 75 meters of a major roadway
    - fraction of population residing between 75 and 200 meters from a major roadway
    - fraction of population residing more than 200 meters from a major roadway

To date only a subset of states have been completely processed. For this subset state summaries of the fraction of population living within various distances of major roadways are presented in Table 1.

Table 1. Fraction of population residing at various distances from major roadways for selected states.

STATE	**Distance from major roadways**		
	< 75 meters	**75 – 200 meters**	**> 200 meters**
Colorado	**0.22**	**0.33**	**0.45**
Georgia	**0.17**	**0.24**	**0.59**
New York	**0.31**	**0.36**	**0.33**