

US EPA ARCHIVE DOCUMENT

The HAPEM5 User's Guide

Hazardous Air Pollutant Exposure Model,
Version 5

March 2005

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 Consulting

60 Broadway

San Francisco, CA 94111

(415) 677-7100

This page intentionally left blank.

Contents

New Features in HAPEM5	iv
1. Introduction	1-1
1.1. Organization of the User's Guide	1-1
1.2. Background	1-2
1.2.1. <i>Population Data</i>	1-2
1.2.2. <i>Activity Data</i>	1-2
1.2.3. <i>Air Quality Data</i>	1-4
1.2.4. <i>Microenvironmental Data</i>	1-5
1.2.5. <i>Stochastic Elements</i>	1-5
1.3. Strengths and Limitations of HAPEM5	1-7
1.3.1. <i>Strengths</i>	1-7
1.3.2. <i>Limitations</i>	1-8
1.4. Applicability	1-8
1.5. Brief History of the Hazardous Air Pollutant Exposure Model	1-9
2. Getting Started—An Overview of HAPEM5	2-1
2.1. Model Structure	2-1
2.1.1. <i>Parameter Files</i>	2-2
2.1.2. <i>DURAV and the Activity and Cluster Files</i>	2-10
2.1.3. <i>INDEXPOP and the Population File</i>	2-11
2.1.4. <i>COMMUTE and the Commuting File</i>	2-11
2.1.5. <i>AIRQUAL and the Air Quality File</i>	2-12
2.1.6. <i>HAPEM, the Microenvironmental Factors File, and the Activity ClusTrans File</i>	2-12
2.1.7. <i>Statefip File</i>	2-14
2.1.8. <i>Background Concentration</i>	2-15
2.1.9. <i>Exposure Output Files</i>	2-15
2.2. Changing the Parameter Settings	2-16
2.2.1. <i>Changing the Number of Microenvironments</i>	2-16
2.2.2. <i>Changing the Number And/or Definitions of the Demographic Groups</i>	2-16
2.2.3. <i>Changing the Number And/or Definitions of Day Types</i>	2-17
2.2.4. <i>Changing the Number And/or Definitions of Time Blocks</i>	2-17
2.3. Setting Up a HAPEM5 Run	2-18
2.3.1. <i>Running HAPEM5 as a "Batch" Job</i>	2-20
2.3.2. <i>Running HAPEM5 Programs Individually</i>	2-20
3. HAPEM5 Input Files	3-1
3.1. Parameter Files	3-1
3.1.1. <i>Specifying the Location and Names of Input and Output Files</i>	3-2
3.1.2. <i>Identifying the Uniform Component of the Background Concentration</i>	3-3
3.1.3. <i>Selecting the Commuting Option</i>	3-4
3.1.4. <i>Setting the Internal Parameters</i>	3-4
3.2. Activity File	3-4
3.2.1. <i>Variables and Format of the Default File</i>	3-5
3.2.2. <i>Replacing or Modifying the Default File</i>	3-8
3.3. Cluster File	3-10

3.3.1.	<i>Variables and Format of the Default File</i>	3-10
3.3.2.	<i>Replacing or Modifying the Default File</i>	3-10
3.4.	Population File	3-12
3.4.1.	<i>Variables and Format of the Default File</i>	3-12
3.4.2.	<i>Replacing or Modifying the Default File</i>	3-12
3.5.	Commuting Flow File	3-15
3.6.	Air Quality File	3-16
3.7.	Microenvironmental Factors File	3-18
3.8.	ClusTrans File	3-24
3.9.	Statefip File	3-25
4.	HAPEM5 Output Files	4-1
4.1.	Log File	4-1
4.1.1.	<i>DURAV Output to the Log File</i>	4-2
4.1.2.	<i>INDEXPOP Output to the Log File</i>	4-3
4.1.3.	<i>COMMUTE Output to the Log File</i>	4-3
4.1.4.	<i>AIRQUAL Output to the Log File</i>	4-3
4.1.5.	<i>HAPEM Output to the Log File</i>	4-4
4.2.	Counter File	4-4
4.3.	Mistract File	4-5
4.4.	Final Exposure File	4-6
4.4.1.	<i>File Format</i>	4-6
5.	HAPEM5 Programs	5-1
5.1.	Programming Guidelines Used to Develop HAPEM5	5-1
5.1.1.	<i>Common Structural Elements</i>	5-2
5.2.	Program Descriptions	5-2
5.2.1.	<i>DURAV</i>	5-3
5.2.2.	<i>INDEXPOP</i>	5-5
5.2.3.	<i>COMMUTE</i>	5-6
5.2.4.	<i>AIRQUAL</i>	5-7
5.2.5.	<i>HAPEM</i>	5-9
6.	References	6-1
Appendix: Development of Algorithm for Creating Longitudinal Activity Patterns with Cluster Analysis		A-1

Figures

Figure 2-1. Overview of HAPEM5	2-3
Figure 2-2a. Example <i>parameter</i> file for running HAPEM5 programs 1-3.....	2-8
Figure 2-2b. Example <i>parameter</i> file for running HAPEM5 programs 4-5.....	2-8
Figure 2-3. HAPEM5 programs and input and output files	2-19
Figure 2-4. Example “batch” file for running the five HAPEM5 programs	2-20

Tables

Table 1-1. HAPEM5 microenvironments	1-3
Table 2-1. Keywords for parameter files and example file names.....	2-4
Table 3-1. Variables in the default <i>activity</i> file.....	3-5
Table 3-2. Variables in the default <i>population</i> file	3-13
Table 3-3. Format for the <i>factors</i> file	3-22
Table 5-1. Keywords for filenames recognized by HAPEM5 <i>parameter</i> files	5-3

New Features in HAPEM5

The Hazardous Air Pollutant Exposure Model, version 5 (HAPEM5) includes a number of new features. Some of these new features are designed to provide exposure estimates that better characterize the variability across the population. Others provide more recent data for improved accuracy in exposure estimates. Several aspects of HAPEM5 have also been modified to make it easier to apply. These new features are summarized here, and detailed in other portions of this User's Guide.

- Microenvironment factors are now characterized with probability distributions instead of point estimates, to better reflect the variability found in measurement studies. Default files are provided for each of three categories of HAPs: gaseous, particulate, and semi-volatile.
- New algorithms have been added for creating annual average activity patterns from daily activity pattern data, using cluster analysis and a Markov chain selection process. These new algorithms require two new input files. Default files are provided.
- The default population file has been updated with 2000 US Census data.
- The final output file now lists many of the parameter settings used for the simulation to help with file tracking.
- The number of replicates to be simulated for each demographic group in each census tract is no longer hard-coded into the HAPEM module, but is specified by the user in the *parameter* file. This means the user no longer has to modify the HAPEM source code and re-compile it in order to change the number of replicates.
- The name of the *parameter* file is no longer hard-coded, but is specified by the user on the command line when a HAPEM5 module is run. This means the user no longer has to modify the source code and re-compile it in order to change the name of the *parameter* file.

1. Introduction

The Hazardous Air Pollutant Exposure Model, version 5 (HAPEM5) User's Guide is designed to assist exposure analysts with running and interpreting results from HAPEM5. 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 HAPEM5, 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 HAPEM5. 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 HAPEM5.

<i>Chapter 1</i>	Provides a brief introduction to HAPEM5 modeling fundamentals including a brief history of the development of HAPEM5.
<i>Chapter 2</i>	Provides an overview of the various components of HAPEM5 and basic information needed to run the model.
<i>Chapter 3</i>	Provides a description of the format, data, and options for each of HAPEM5 input files.
<i>Chapter 4</i>	Provides a description of the format and data associated with each of HAPEM5 output files.
<i>Chapter 5</i>	Provides a description of the purpose, operations, inputs, and outputs, including a brief description of the computer code, for each of HAPEM5 computer programs.
<i>Chapter 6</i>	References.

1.2. Background

The Hazardous Air Pollutant Exposure Model, version 4 (HAPEM5) 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 urban¹ to national. HAPEM5 provides a relatively transparent set of exposure assumptions and approximations, as is appropriate for a screening level model.

HAPEM5 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.

HAPEM5 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.

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.

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 HAPEM5 uses 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.

1.2.2. Activity Data

HAPEM5 uses two types of population activity data: activity pattern data and commuting pattern 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' 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.

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

The default population activity file for HAPeM5 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 HAPeM5 are presented in Table 1-1.

**Table 1-1.
HAPeM5 microenvironments**

MICRO – ENVIRONMENT No.	MICROENVIRONMENT		MICRO – ENVIRONMENT No.	MICROENVIRONMENT	
	SPECIFIC	GENERAL		SPECIFIC	GENERAL
1	Car	In vehicle	20	Manufacturing facility	Indoors
2	Bus	In vehicle	21	School	Indoors
3	Truck	In vehicle	22	Church	Indoors
4	Other	In vehicle	23	Shopping mall	Indoors
5	Public garage	Indoors	24	Auditorium	Indoors
6	Parking lot/garage	Outdoors	25	Health care facility	Indoors
7	Near road	Outdoors	26	Other public building	Indoors
8	Motorcycle	Outdoors	27	Other location	Indoors
9	Service station	Indoors	28	Not specified	Indoors
10	Service station	Outdoors	29	Construction site	Outdoors
11	Residential garage	Indoors	30	Residential grounds	Outdoors
12	Other repair shop	Indoors	31	School grounds	Outdoors
13	Residence - no gas stove	Indoors	32	Sports arena	Outdoors
14	Residence - gas stove	Indoors	33	Park/golf course	Outdoors
15	Residence - attached garage	Indoors	34	Other location	Outdoors
16	Residential - stove and garage	Indoors	35	Not specified	Outdoors
17	Office	Indoors	36	Train/subway	In vehicle
18	Store	Indoors	37	Airplane	In vehicle
19	Restaurant	Indoors			

Because available activity data are not adequate to estimate the exposure of each individual in a population, HAPEM5 groups activity patterns data together for people with similar demographic characteristics that are expected to influence exposure to air pollutants (e.g., age, gender, work 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 HAPEM5 provides a distribution of exposure concentrations for each demographic group in each census tract.

HAPEM5 divides the population into 10 demographic groups, based on combinations of age (5 categories) and gender. Activity pattern data are also separated into 3 day types: summer weekdays, other weekdays, and weekends.

The commuting data contained in the HAPEM5 default file was derived from a special 1990 US Census study that specifies 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. HAPEM5 uses this data in coordination with the activity pattern data to place an individual either in the home tract or the work tract at each time step. For this version of HAPEM we have adjusted these 1990 data to reflect the modified Census tract designations for 2000. However, we have not updated the fundamental patterns from what was reflected in the 1990 data, because the corresponding data from the 2000 Census was not yet available at the time of this document was written.

1.2.3. Air Quality Data

Some previous versions of HAPEM relied on measured outdoor air pollutant 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 pollutants².

HAPEM4, the immediate predecessor to HAPEM5, however, incorporated the capability of using model predictions of outdoor air quality data in addition to measured data. This added flexibility, retained in HAPEM5, gives the model the ability to treat more pollutants and geographical areas than the earlier versions.

HAPEM5 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, HAPEM5 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 is then combined in HAPEM5 with similarly stratified activity data to estimate exposure concentrations. The default activity data is stratified into (8) 3-hour time blocks.

A new feature in this version of HAPEM is the ability 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 stratified), but may be characterized by a set of up to 500 diurnally

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.

stratified values. How HAPEM5 handles this data set is explained below in the section on 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 HAPEM5 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, 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 an input file. The default *factors* files for HAPEM5 contain estimates of distributions for *PEN*, *PROX*, and *ADD* for three categories of pollutants: gases, particles, and semi-volatiles. These were obtained from an extensive review of literature and databases on indoor/outdoor ratios of hazardous air pollutants (HAPs). How they are utilized in HAPEM5 is discussed below in the section on stochastic elements. There are also point estimates of *PEN* and *PROX* factors available for 32 specific pollutants from HAPEM4, the predecessor to HAPEM5.

As is the case with all other HAPEM5 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. HAPEM5 incorporates four stochastic elements.

Activity Patterns

The first stochastic element is the selection of activity patterns to represent each demographic group. HAPEM5 estimates long-term average concentrations, but the available population activity data sequences are specified for 24-hour periods only. HAPEM5 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.)

The general approach used by HAPEM5 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 30 times for each census tract/demographic group combination, to create a set of 30 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 and day type. Then for each demographic group/day type combination the activity patterns are stratified into two or 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 HAPEM5 algorithms construct an average activity pattern for each tract/demographic-group/day-type combination, by randomly selecting one activity pattern from each category and combining them with weighted averaging. 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 HAPEM5 for demographic groups whose activity patterns indicate time spent at 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.

Air Quality

A new feature of HAPEM5 is the ability to characterize air quality concentrations as spatially variable within a census tract. The data for each tract is entered as a data set with up to 500 sets of value (i.e., sets of eight annual average 3-hour time blocks). For each demographic group/replicate, a different set of ambient air quality concentrations is selected for the tract to reflect the variability in air quality among residential locations within the tract.

Microenvironment Factors

Another new feature of HAPEM5 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. Several distribution types may be used, as discussed in Section 3.7. For each tract/demographic group/replicate, and in some cases source type, a different set of microenvironment factors is randomly selected.

1.3. Strengths and Limitations of HAPEM5

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 HAPEM5. 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. The most recent of these 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.

An important feature of HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 benzene. (Uniform ME factors are also available for more than 30 HAPs.)

1.3.2. Limitations

HAPEM5 calculates long-term average exposure concentrations in order to address exposures to pollutants with carcinogenic and other long-term effects. Thus, HAPEM5 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.

HAPEM5 only estimates exposures experienced through inhalation. For certain HAPs, inhalation might not be the major route of exposure, and therefore, HAPEM5 may underestimate exposures in these instances. Also, although HAPEM5 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 is 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 was derived from the 1990 US Census, and therefore may not reflect current commuting patterns. Moreover, the commuting 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 HAPEM5 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 factors incorporated into HAPEM5 were derived from reported measurement studies. The data available was 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.
- 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

HAPEM5 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, HAPEM5 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, HAPEM5 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)³ 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

³ The EPA changed this name to the Office of Transportation and Air Quality in 1999.

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 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.

This previous version of the model, called 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 is 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 current version of the model, HAPEM5, incorporates additional enhancements. These include the use of variable ME factors and air quality data that is spatially variable within Census tracts. It also contains 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 will be applied as part of the NATA national-scale assessment for 1999.

NOTE: The current version of HAPEM5 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.

2. Getting Started—An Overview of HAPEM5

This chapter provides the user the basic information needed to run HAPEM5. The topics addressed in this chapter include the functions of the programs that comprise the HAPEM5, the contents of the various input and output files, and the meanings of parameter values. The chapter has been separated into the following sections.

- | | |
|--------------------|--|
| <i>Section 2.1</i> | Describes the general structure of HAPEM5, the input and output files, and the parameter settings. |
| <i>Section 2.2</i> | Discusses considerations for changing parameter settings. |
| <i>Section 2.3</i> | Provides instructions for setting up and running HAPEM5. |

Figure 2-1 presents a graphical overview of the HAPEM5 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 HAPEM5. 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 HAPEM5 programs use eight 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 **nreplc** (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 files⁴), or a parameter value is specified. As currently configured HAPEM5 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 HAPEM 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 2-2 presents an example *parameter*

⁴ Indoor source algorithms are included in the HAPEM5 module, 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.

file that can be used to run all 5 HAPEM5 programs. Note that the input and output file names must be listed before the parameter settings.

**Figure 2-1.
Overview of HAPEM5**

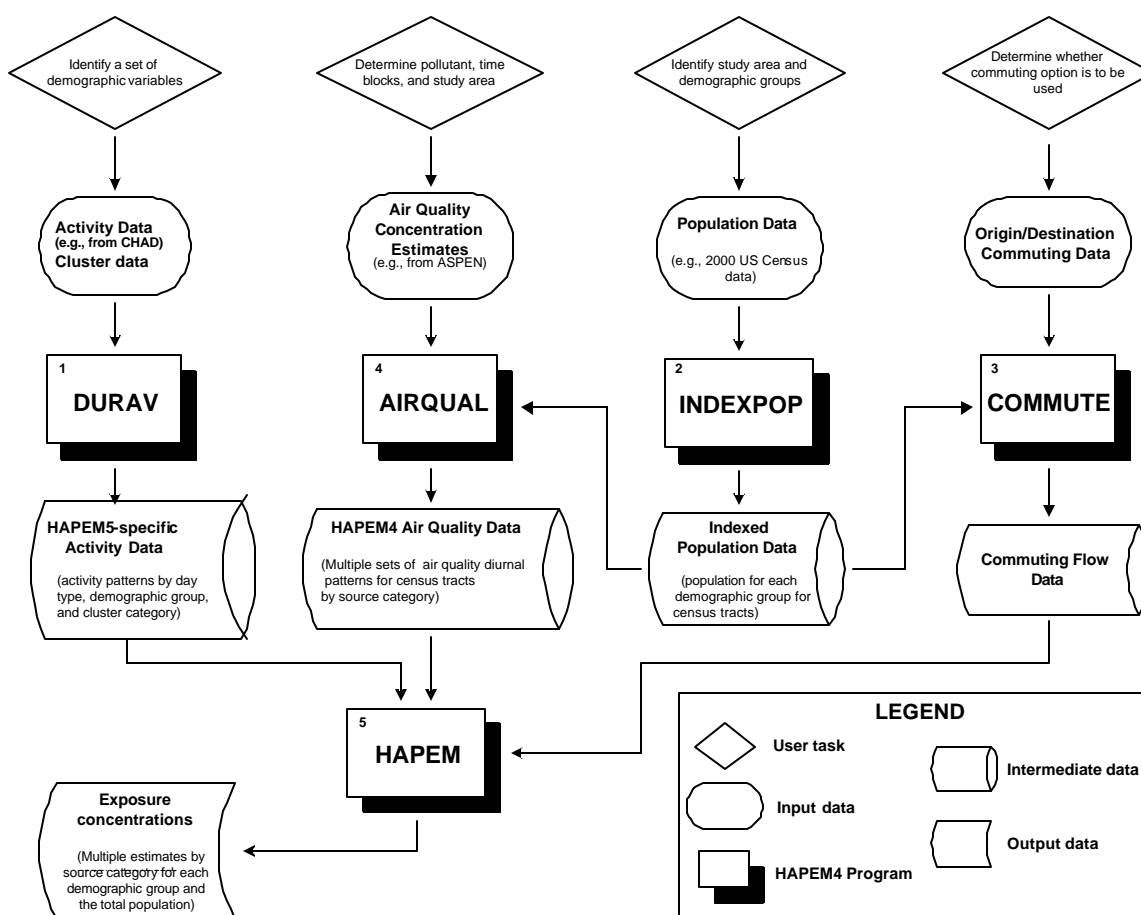


Table 2-1.
Keywords for parameter files and example file names

User/Model Defined	Inputs	Outputs
DURAV.f90		
User defined files ?	<i>activity</i> file (e.g., <i>durhw.txt</i>) <i>cluster</i> file (e.g., <i>durhw_cluster.txt</i>)	<i>log</i> file <i>counter</i> file
User defined parameters ?	comm nmicro nblock hblock ntype ngroup	
HAPEM5 defined files ?		<i>durhw.wrong_chad</i> <i>durhw.da</i> <i>durhw.nonzero</i>
INDEXPOP.f90		
User defined files ?	<i>population</i> file (e.g., <i>census2000.txt</i>) <i>statefip</i> file	<i>log</i> file <i>counter</i> file
User defined parameters ?	region1 region2 ngroup	
HAPEM5 defined files ?		<i>census.da</i> <i>census_direct.ind</i> <i>census.county_tract_pop_range</i> <i>census.state_county_pop_range</i>
COMMUTE.f90		
User defined files ?	<i>commuting</i> file (e.g., <i>comm2000.txt</i>) <i>population</i> file (e.g., <i>census2000.txt</i>) <i>statefip</i> file	<i>log</i> file <i>counter</i> file <i>mistract</i> file
User defined parameters ?	region1 region2 keep	
HAPEM5 defined files ?	<i>census_direct.ind</i> <i>census.county_tract_pop_range</i> <i>census.state_county_pop_range</i>	<i>comm.da</i> ; <i>comm.ind</i> <i>comm.st_comm1_fip_range</i>

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

User/Model Defined	Inputs	Outputs
(continued)		
AIRQUAL.f90		
User defined files ?	<i>air quality</i> file (e.g., <i>benzene.txt</i>) <i>population</i> file (e.g., <i>census.txt</i>) <i>statefip</i> file	<i>log</i> file <i>counter</i> file <i>mistract</i> file
User defined parameters ?	Comm hblock nsource ngroup region1 region2	
HAPEM5 generated files ?	<i>census.da</i> <i>census_direct.ind</i>	<i>benzene.da</i> <i>benzene.air_da</i> <i>benzene.pop_air_da</i> <i>benzene.state_air_fip_range</i> <i>benzene.state_air1_fip_range</i> <i>benzene.state_air2_fip_range</i>
HAPEM.f90		
User defined files ?	<i>factors</i> file (e.g., <i>factors.txt</i>) <i>air quality</i> file (e.g., <i>benzene.txt</i>) <i>commuting</i> file (e.g., <i>comm2000.txt</i>) <i>activity</i> file (e.g., <i>durhw.txt</i>) <i>ClusTrans</i> file (e.g., <i>clustertransa.txt</i>) <i>Product</i> files ⁵ (specify path only) <i>AutoPduct</i> file ⁶	<i>log</i> file <i>counter</i> file <i>mistract</i> file <i>afile</i> file (specify path only)

⁵ A path to one or more indoor emission source inputs for the HAPEM5 indoor source algorithms is specified in this statement. These algorithms are included in the HAPEM5 module, 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.

⁶ The full path name of an existing file must be specified as the *AutoPduct* file in HAPEM5, although its only function is as input to the indoor source algorithms. These algorithms are included in the HAPEM5 module, 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 module.

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

User/Model Defined	Inputs	Outputs
User defined parameters ?	Pollutant CAS ⁷ units year comm backg sarod nmicro hblock ntype ngroup nsource nreplic region1 region2 Rseed1 Rseed2 Rseed3 M_00 M_05 M_12 M_18 M_65 F_00 F_05 F_12 F_18 F_65	

⁷ 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 HAPEM5 module, 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
HAPEM5 generated files ?	<i>durhw.da</i> <i>durhw.nonzero</i> <i>benzene.da</i> <i>benzene.air_da</i> <i>benzene.pop_air_da</i> <i>benzene.state_air_fip_range</i> <i>benzene.state_air1_fip_range</i> if commuting is included: <i>comm.da;</i> <i>comm.ind</i> <i>comm.st_comm1_fip_range</i> otherwise: <i>benzene.state_air2_fip_range</i>	

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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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
statefp file     =      input\statefp.dat
OUTPUT FILES:
log file          =      output\log_file.txt.
counter file      =      output\counter.dat
mistract file     =      output\mistract.dat
PARAMETER SETTINGS:
comm              =      YES
region1           =      1
region2           =      53
keep              =      YES
nmicro            =      37          ! 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            =      10          ! Number of demographic groups
  
```

Figure 2-2b.
Example *parameter* file for running HAPEM5 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
statefp file       =      input\statefp.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	
comm	=	YES	
region1	=	1	
region2	=	53	
sarod	=	45201	
backg	=	0.00	
nmicro	=	37	! Number of microenvironments
hblock	=	8	! Number of time blocks/day for analysis
ntype	=	3	! Number of day types
ngroup	=	10	! Number of demographic groups
nsource	=	4	! Number of source categories
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:

M_00	=	Male, Ages 0-4
M_05	=	Male, Ages 5-11
M_12	=	Male, Ages 12-17
M_18	=	Male, Ages 18-64
M_65	=	Male, Ages > 65
F_00	=	Female, Ages 0-4
F_05	=	Female, Ages 5-11
F_12	=	Female, Ages 12-17
F_18	=	Female, Ages 18-64
F_65	=	Female, 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), 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.txt*, there are 37 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 37 microenvironments for the first hour comes first in the *activity* file, followed by the 37 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 37 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* file (discussed below) must be the same and is specified in the *parameter* files of DURAV and HAPEM as **nmicro**. 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 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.

Whether commuting will be included in the analysis is indicated by the value of **comm**, specified in the *parameter* files of DURAV, AIRQUAL, and HAPEM as either “yes” or “no”. If no commuting is indicated, DURAV will aggregate the home and work durations for each time block.

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 demographic group and day type are also specified in the *cluster* file but HAPEM5 does not use that information.)

The current version of DURAV divides the activity data into 10 demographic groups, based on combinations of age (5 categories) and gender. 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, is either 2 or 3 depending on the demographic group and day type.

2.1.3. INDEXPOP and the Population File

The INDEXPOP program performs two 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.

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.

2.1.4. COMMUTE and the Commuting File

The COMMUTE program performs two main functions:

- It creates a sequential file identifying for each home census tract the associated set of work tracts (i.e., tracts in which the residents of the home tract work), and the fraction of home tract workers in each work tract.
- It creates sequential index files to facilitate file searching in HAPEM.

The *commuting* file is the main input file to the COMMUTE program. The default *commuting* file was derived from 1990 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. These data were adjusted to match census tracts for 2000. 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 $\ln(\ln(\text{total flow}))$ 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

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.

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 File, 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* 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* factor). Each subset contains randomly selected ME factors for each of 37 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 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 *ClusTrans* 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.)
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* file provides the factors used to calculate an estimated microenvironmental concentration from an outdoor concentration. This version of HAPEM uses an updated ME factor methodology. It includes a data base of distributions of ME factors for each of three types of HAPs: gaseous, particulate, and semi-volatile. This updated 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.

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

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) 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* file.

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.⁸

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 values for each of these factors for each ME/emission source category combination. As noted above, the number of MEs in the *factors* 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. There are three default factors files: one each for gaseous, particulate, and semi-volatile HAPs. The default *factors* files contains ME factors applicable to all the MEs included in the default *activity* file, for four emission source categories (e.g., point, area, onroad mobile, and nonroad mobile). These category-specific estimates were derived from reported measurement 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 *ClusTrans* 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

⁸ 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 HAPEM5 application until the database and algorithms have been tested and reviewed. To disable the indoor source algorithms, set keyword **CAS** to 99999.

the default *statefip* file (assuming the input files contain all the necessary data). Alternatively, 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 is 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 HAPEM5 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 30 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., one-thirtieth 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 HAPEM5 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 HAPEM5 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, HAPEM5 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* file. Definitions of the microenvironments do not appear anywhere in HAPEM5 program code.

The HAPEM5 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, and
- The number of columns in the *population* file.

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 module 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 10 current demographic groups are defined by combinations of the following age and gender categories.

Age:

- 0 - 4
- 5 - 11
- 12 - 17
- 18 - 64
- 65+

Gender:

- Male
- Female

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. 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. 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 HAPEM5, 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 HAPEM5 Run

This section shows how to set-up and make a simple HAPEM5 run. Subsequent sections and chapters provide more detailed explanation about HAPEM5's input and output files and the model's programs.

The example shown in this section is for a hypothetical HAPEM5 analysis of benzene. Figure 2-3 presents a detailed picture of all of the inputs, outputs, and programs associated with HAPEM5.

The most important consideration for making a HAPEM5 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 eight 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.

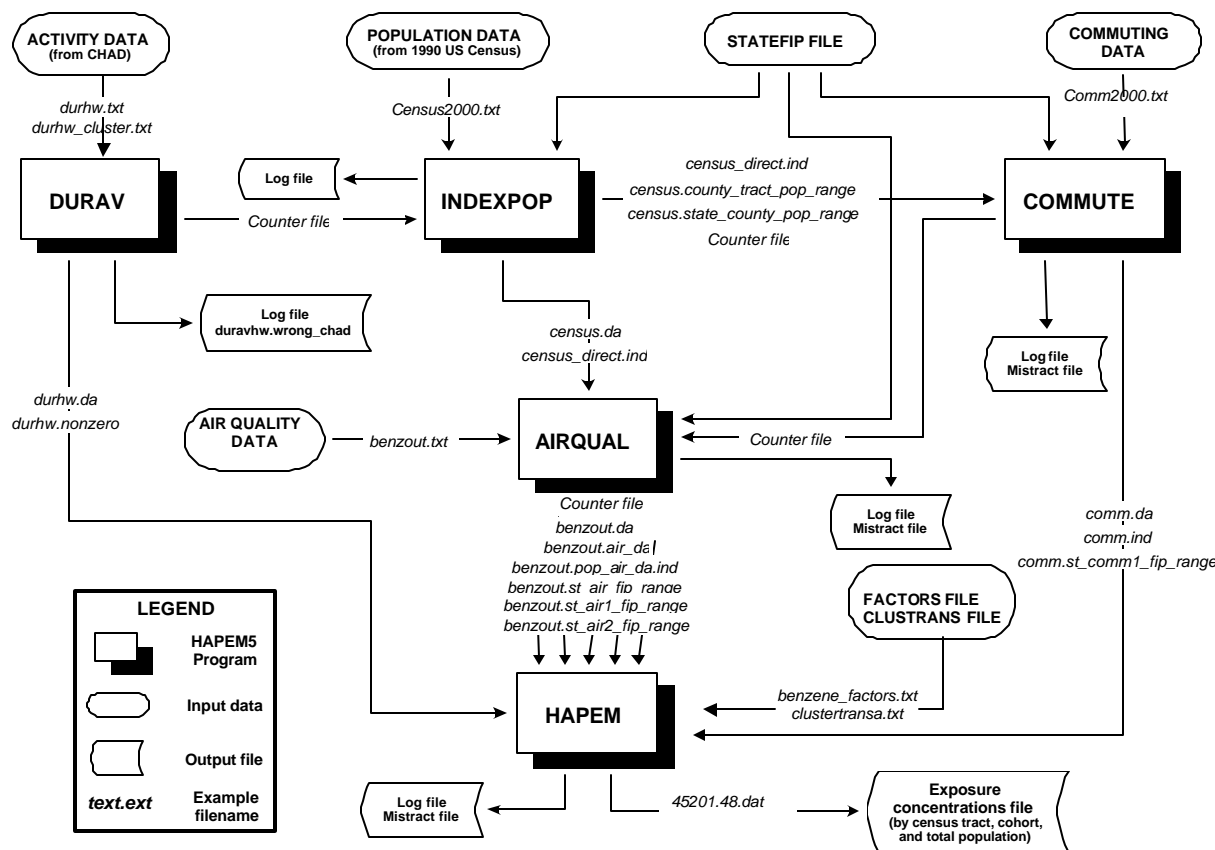
As noted above, an existing path name should be specified for the *product* files, and the full path name of any existing file (except other HAPEM input or output files) must be specified as the *AutoPduct* file in the *parameter* file used with HAPEM. In the future these file will comprise part of the 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 **comm** = YES (commuting is to be included in the analysis), 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 thirty-seven microenvironments to be modeled (***nmicro***). This number of microenvironments must be consistent with the number of microenvironmental (ME) factors specified in the *factors* file (i.e., *gas_factors.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* file (i.e., *census2000.txt*).

Figure 2-3.
HAPEM5 programs and input and output files



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 HAPEM5 as a “Batch” Job

When running HAPEM5 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 HAPEM5 programs sequentially, with all five programs residing in the same directory as the batch file. The batch file is shown in Figure 2-4.

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 HAPEM5 executable programs will be run.

Figure 2-4.
Example “batch” file for running the five HAPEM5 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 HAPEM5 Programs Individually

Any of HAPEM5 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 HAPEM5 programs are run similarly.

As indicated in Table 2-1 and Figure 2-3, INDEXPOP, COMMUTE, AIRQUAL and HAPEM all require input files that are generated from running other HAPEM5 programs. Therefore, if any of these programs is run alone, the user must ensure that the required HAPEM5 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. For this example these files are *census.da* and *census_direct.ind*, with filenames derived from *census.txt*. Therefore, the *parameter* file for running AIRQUAL must specify the full path name of the *population* file, and the two intermediate files must exist and reside in the subdirectory specified for *population* file.

3. HAPEM5 Input Files

The HAPEM5 programs use eight 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 HAPEM5 are discussed in Chapter 2. The reader is referred to that chapter for an overview of HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5-generated files are required for each of the five programs that comprise HAPEM5 is discussed in Chapter 2 and presented in Table 2-1.

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

- The *activity* file;
- The *cluster* file;
- The *population* file (national scope);
- The *commuting* file (national scope);
- The *factors* files (one each for gaseous, particulate, and semi-volatile 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 eighth 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 HAPEM5 runs:

3. HAPEM5 Input Files

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. Option for selecting commuting (YES or NO);
7. Seven internal parameter settings; and
8. 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 HAPEM5 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 HAPEM5 programs is presented in Figures 2-2a and 2-2b in Chapter 2.

The HAPEM5 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 HAPEM5 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, HAPEM5 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 is 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. Selecting the Commuting Option

To include commuting in the HAPEM5 simulations the value of the **comm** parameter must be set to YES. The default mode for this option is not to include commuting in the simulations.

3.1.4. Setting the Internal Parameters

The seven internal parameter settings (**nmicro**, **nblock**, **hblock**, **ntype**, **ngroup**, **nsource**, and **nreplic**) 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.

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 **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.

Internal Parameters	
nmicro	number of microenvironments in the <i>activity</i> and <i>factors</i> files
nblock	number of time blocks per day in the <i>activity</i> 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 <i>population</i> file
nsource	number of emission source categories in the <i>air quality</i> file
nreplic	number of replicates to be simulated for each demographic group in each census tract

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=37 NBLOCK=24 HW=2 IN FORTRAN ORDER)
--

Although most of the header record of the *activity* file is not used by the HAPEM5 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 as follows.

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 37 microenvironments (listed in Table 1-1), 24 time blocks per day, and two locations has a total of 1776 duration values. These values are sequenced so that the 37 microenvironment durations for the first time block in the home location come first, followed by the 37 microenvironment durations for the second time block in the home location, and so on, until all the 888 values for the home location are specified. These are followed by the 888 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.

3. HAPEM5 Input Files

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

DAYTYPE	integer indicator of day type for classification, with values as follows: <div> 1 = summer weekday 2 = non-summer weekday 3 = weekend </div>
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: <div> "1" = female "2" = male "9" = unknown </div>
RACE	1-character string, indicating race/ethnic group, used to assign the respondent to a demographic group, with values as follows: <div> "1" = White (non-Hispanic) "2" = Black (non-Hispanic) "3" = Hispanic (any race) "4" = Asian or Other (non-Hispanic) "9" = unknown </div> <p>This information is not used by the current version of the DURAV program.</p>
EMPLOYED	1-character string, indicating employment status of respondent, with values as follows <div> "Y" = Yes "N" = No "X" = missing </div> <p>This information is not used by the current version of the DURAV program.</p>

(continued)

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.
------------------	---

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

AGE	integer indicator of the age group, used to assign the respondent to a demographic group with values as follows 1 = 0-4 years old 2 = 5-11 years old 3 = 12-17 years old 4 = 18-64 years old 5 = 65 years or older 9 = unknown
COMMUTE	integer indicator of whether the respondent is a commuter, with values as follows: 0 = no commuting 1 = commuting

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

[illegible]

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.

3. HAPEM5 Input Files

- 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, GENDER, 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. 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* file (see Section 3.6). 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 HAPEM5 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 group (uncombined)	Day Type	CLUSTER	Ncluster
--------	--------------------------------	----------	---------	----------

Although the header record of the *cluster* file is not used by the HAPEM5 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 combination.

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

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, a gender designation of 1 or 2, and an age group designation of 1 –5, and a day type designation of 1-3). 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, and cluster.

3. HAPEM5 Input Files

Extract from Default *cluster* File

CHADID	Demographic	group	(uncombined)	Day Type	CLUSTER	Ncluster
CAC01170A	1	1	1	3		
CAC01174A	1	1	1	3		
CAC01251A	1	1	1	3		
CAC01489A	1	1	1	3		
CAC01562A	1	1	1	3		
CAC01569A	1	1	1	3		
CAC01830A	1	1	1	3		
CAC01982A	1	1	1	3		
CAC02036A	1	1	1	3		
CAC02084A	1	1	1	3		
CAC02111A	1	1	1	3		
CAC02213A	1	1	1	3		
CAC02219A	1	1	1	3		
CAC02321A	1	1	1	3		
CAC02349A	1	1	1	3		
CAC02354A	1	1	1	3		
CAC02411A	1	1	1	3		
CAC02416A	1	1	1	3		
CAC02419A	1	1	1	3		
CAC02555A	1	1	1	3		
CAC02568A	1	1	1	3		
CAC02581A	1	1	1	3		
CAC02812A	1	1	1	3		
CAC02846A	1	1	1	3		
CAC02867A	1	1	1	3		
CAC02987A	1	1	1	3		
CAC03171A	1	1	1	3		
CAC03247A	1	1	1	3		
CAC03259A	1	1	1	3		
CAC03323A	1	1	1	3		
CAC03392A	1	1	1	3		
CAC03531A	1	1	1	3		
CAC03539A	1	1	1	3		
CAC03547A	1	1	1	3		
CAC03709A	1	1	1	3		
CAC03733A	1	1	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 is 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	M_00	M_05	M_12	M_18	M_65	F_00	F_05	F_12	F_18	F_65
-------	------	------	------	------	------	------	------	------	------	------

Although the header records of the *population* file are not used by the HAPEM5 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”,

3. HAPEM5 Input Files

“*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 HAPEM5 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.																					
X_YY	<p>Integer specifying the number of 2000 tract residents with gender X and age group YY.</p> <p>The gender definitions are:</p> <table><tr><td>M</td><td>=</td><td>male</td></tr><tr><td>F</td><td>=</td><td>female</td></tr></table> <p>The age group definitions are:</p> <table><tr><td>00</td><td>=</td><td>0-4 years old</td></tr><tr><td>05</td><td>=</td><td>5-11 years old</td></tr><tr><td>12</td><td>=</td><td>12-17 years old</td></tr><tr><td>18</td><td>=</td><td>18-64 years old</td></tr><tr><td>65</td><td>=</td><td>65 years or older</td></tr></table>	M	=	male	F	=	female	00	=	0-4 years old	05	=	5-11 years old	12	=	12-17 years old	18	=	18-64 years old	65	=	65 years or older
M	=	male																				
F	=	female																				
00	=	0-4 years old																				
05	=	5-11 years old																				
12	=	12-17 years old																				
18	=	18-64 years old																				
65	=	65 years or older																				

3. HAPEM5 Input Files

Extract from Default *population* File

TRACT	M_00	M_05	M_12	M_18	M_65	F_00	F_05	F_12	F_18	F_65
	COM	COM	COM	COM	COM	COM	COM	COM	COM	COM
01001020100	50	118	101	622	82	44	110	96	603	95
01001020200	75	100	89	522	79	77	107	82	626	135
01001020300	115	210	164	899	151	130	183	158	1025	304
01001020400	120	230	219	1308	305	129	231	194	1422	398
01001020500	207	461	300	1837	167	239	393	267	1914	255
01001020600	106	244	162	993	135	119	201	193	1052	173
01001020700	98	143	130	847	109	116	147	138	968	204
01001020800	354	568	485	2947	328	303	528	476	2992	376
01001020900	173	287	274	1384	184	156	265	218	1403	285
01001021000	95	153	132	791	135	92	143	145	793	198
01001021100	109	172	157	853	142	116	149	148	925	211
01003010100	144	202	216	1231	261	137	217	174	1271	308
01003010200	85	137	143	816	127	81	144	115	821	165
01003010300	239	335	311	1973	376	224	330	277	2078	368
01003010400	183	291	233	1402	191	126	227	205	1357	227
01003010500	123	202	162	1540	282	110	154	174	1304	515
01003010600	212	252	200	963	132	197	241	195	1144	188
01003010701	197	357	308	1855	418	192	348	326	1893	616
01003010703	198	406	277	1582	207	195	323	261	1663	193
01003010704	171	226	231	1495	220	177	247	238	1595	232
01003010705	186	309	236	1983	259	202	275	225	2029	299
01003010800	163	328	328	1791	358	157	311	312	1965	541
01003010901	307	523	477	3315	439	323	487	418	2857	500
01003010902	333	506	495	2637	500	340	503	456	2857	581
01003011000	173	227	186	1218	240	160	223	186	1247	235
01003011100	239	393	357	2032	535	254	366	346	2263	739
01003011201	71	171	167	1070	426	83	152	177	1206	663
01003011202	151	218	222	1425	457	123	207	214	1627	761
01003011300	107	176	151	1119	259	107	167	143	1159	277
01003011401	223	357	262	2083	441	217	370	310	2182	46

3.5. 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 1990 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 HAPEM5 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	0.05
01001020100	01001020200	1.43	6.98e-03
01001020100	01001020300	2.77	0.04
01001020100	01001020500	6.10	0.23
01001020100	01001020700	5.52	0.17
01001020100	01001020800	3.55	0.03
01001020100	01011952200	81.32	5.81998e-03
01001020100	01051031100	13.10	0.01
01001020100	01073012903	109.14	8.14998e-03
01001020100	01085981000	31.69	5.82004e-03
01001020100	01101000100	19.64	0.04
01001020100	01101000200	20.07	0.07
01001020100	01101000300	18.08	0.01
01001020100	01101000600	21.51	0.04
01001020100	01101000700	20.69	0.01
01001020100	01101001000	17.80	0.01
01001020100	01101001400	22.40	4.66001e-03
01001020100	01101001800	22.68	0.02
01001020100	01101002000	24.97	4.66001e-03
01001020100	01101002201	22.50	5.76603e-03
01001020100	01101002202	23.39	1.22398e-03
01001020100	01101002400	20.16	2.03e-06
01001020100	01101002800	26.41	6.98e-03
01001020100	01101002900	27.80	0.02
01001020100	01101003000	17.68	0.02
01001020100	01101003200	27.76	8.14998e-03
01001020100	01101003301	24.95	5.81998e-03
01001020100	01101005301	23.58	6.98e-03
01001020100	01101005401	32.21	0.03
01001020100	01101005402	28.26	0.01
01001020100	01101005403	28.82	6.99002e-03
01001020100	01101005902	22.87	6.97798e-03
01001020100	01101006000	13.53	0.03

3.6. 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*

3. HAPEM5 Input Files

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	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_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- last	space-delimited concentration contribution values for each emission source category/time block combination plus one spatially variable background at the end(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 HAPEM5 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.

3.7. Microenvironmental Factors File

The microenvironmental (ME) *factors* file provides probability distributions for the factors used to calculate an estimated microenvironmental concentration from an outdoor concentration. The file contains 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.

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

$$ME(m, t, i) = ADD(m, t) \quad (3-1b)$$

$$ME(m, c, t, b) = PROX(m, s) \times PEN(m, t) \times [bckgd_u + bckgd_v(c)] \quad (3-1c)$$

$$ME(m, c, t) = \sum_s ME(m, c, t, s) + ME(m, t, i) + ME(m, c, t, b) \quad (3-1d)$$

where:

$ME(m, c, t, s)$: concentration in microenvironment m located in census tract c at time t due to source category s ,

$PROX(m, s)$: proximity factor for microenvironment m and source category s (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
- $ME(m,c,t)$: total concentration in microenvironment m located in census tract c at time t .

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,

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

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,

$$PROX_i = \frac{\text{outdoor concentration in immediate vicinity of indoor or in-vehicle ME}}{\text{air quality file concentration}} \quad (3-3a)$$

$$PROX_o = \frac{\text{outdoor ME concentration}}{\text{air quality file concentration}} \quad (3-3b)$$

As explained elsewhere, the air quality data from ASPEN represent a spatial average over some portion 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 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) 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

3. HAPEM5 Input Files

roadways tends to be relatively steep. This condition for motor vehicle emissions is reflected in the default *factors* file.

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.⁹

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* file has no header record, and contains one data record for each ME, with the same number, definition and order as the MEs in the *activity* file. The file is read in free format, once for each ME, with fields as specified in Table 3-3. All values are decimal numbers.

Distributions for *PROX* factors are specified for each of 4 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 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.

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.]

⁹ 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 HAPEM5 application until the database and algorithms have been tested and reviewed. To disable the indoor source algorithms, set keyword **CAS** to 99999.

3. HAPEM5 Input Files

- 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

For HAPEM5 default *factors* file are provided for each of three categories of HAPs: gaseous, particulate, and semi-volatile. 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. An excerpt from the default *factors* file for gaseous HAPs is presented below.

As with other HAPEM5 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	10	0.678	0.773	0.884	0.914	0.947	0.971	0.982	0.984	1.000	1.000
	5	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
	4	1	14.36	1.933	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
2	5	10	0.678	0.773	0.884	0.914	0.947	0.971	0.982	0.984	1.000	1.000
	5	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
	4	1	14.36	1.933	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
3	5	10	0.678	0.773	0.884	0.914	0.947	0.971	0.982	0.984	1.000	1.000
	5	1	0	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0
	4	1	14.36	1.933	0	0	0	0	0	0	0	0
	5	1	1	0	0	0	0	0	0	0	0	0

3. HAPEM5 Input Files

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

ME Factor	Data Fields	Parameter	
	Field 1	Number of Microenvironment (1-37)	
PEN	Field 2	<u>Distribution Type</u> 1 - Normal 2 - Log-normal 3 - Uniform 4 - Triangular 5 - Data Set	
	Field 3	<u>Distribution Type</u> Normal Lognormal Uniform Triangular Dataset	<u>Parameter</u> Mean Mean Minimum Minimum Number of data points
	Field 4	<u>Distribution Type</u> Normal Lognormal Uniform Triangular Dataset	<u>Parameter</u> Standard deviation Standard deviation Maximum Maximum First data point in the set
	Field 5	<u>Distribution Type</u> Normal Lognormal Triangular Dataset	<u>Parameter</u> 0 (always) 0 (always) Mode Second data point in the set
PEN (continued)	Field 6	<u>Distribution Type</u> Normal Lognormal Dataset	<u>Parameter</u> Lower bound (optional) Lower bound (optional) Third data point in the set
	Field 7	<u>Distribution Type</u> Normal Lognormal Dataset	<u>Parameter</u> Upper bound (optional) Upper bound (optional) Fourth data point in the set

(continued)

3. HAPEM5 Input Files

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

<i>PEN</i> (concluded)	Field 8	<u>Distribution Type</u> Dataset	<u>Parameter</u> Fifth data point in the set
	Field 9	<u>Distribution Type</u> Dataset	<u>Parameter</u> Sixth data point in the set
	Field 10	<u>Distribution Type</u> Dataset	<u>Parameter</u> Seventh data point in the set
	Field 11	<u>Distribution Type</u> Dataset	<u>Parameter</u> Eighth data point in the set
	Field 12	<u>Distribution Type</u> Dataset	<u>Parameter</u> Ninth data point in the set
	Field 13	<u>Distribution Type</u> Dataset	<u>Parameter</u> Tenth data point in the set
<i>ADD</i>	Fields 14-25	Repeat fields 2-13 for additive factor	
<i>PROX</i>			
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	
<i>LAG</i>			
Source 1	Field 38	hours	
Source 2	Field 51	hours	
Source 3	Field 64	hours	
Source 4	Field 77	hours	

(concluded)

3.8. ClusTrans File

The *ClusTrans* 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.

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	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 HAPEM5 programs, it provides documentation to inform the user of the meaning of the data fields.

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

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

The default *ClusTrans* file is presented below.

Default *ClusTrans* file

1	1 3	0.67045	0.90909	1.00000	0.92647	0.97059	1.00000	0.10000	0.90000	1.00000	0.12500	0.25000	1.00000
1	2 3	0.36705	0.82370	1.00000	0.52885	0.98077	1.00000	0.09524	0.90476	1.00000	0.06250	0.15625	1.00000
1	3 2	0.61739	1.00000	0.00000	0.52941	1.00000	0.00000	0.66667	1.00000	0.00000	0.00000	0.00000	0.00000
2	1 3	0.25856	0.76426	1.00000	0.45455	0.90909	1.00000	0.05714	0.88571	1.00000	0.08333	0.37500	1.00000
2	2 3	0.20129	0.69437	1.00000	0.79167	0.91667	1.00000	0.11111	0.83333	1.00000	0.16667	0.75000	1.00000
2	3 3	0.66440	0.79183	1.00000	0.71429	0.78571	1.00000	0.37500	0.62500	1.00000	0.58333	0.75000	1.00000
3	1 3	0.50000	0.65574	1.00000	0.85714	0.92857	1.00000	0.58333	0.83333	1.00000	0.05357	0.08929	1.00000
3	2 3	0.13559	0.44310	1.00000	0.50000	0.83333	1.00000	0.08333	0.50000	1.00000	0.02500	0.46250	1.00000
3	3 3	0.68358	0.75821	1.00000	0.75000	0.81250	1.00000	0.50000	0.75000	1.00000	0.62500	0.70833	1.00000
4	1 3	0.35900	0.89573	1.00000	0.67647	0.94118	1.00000	0.13158	0.97368	1.00000	0.58333	0.75000	1.00000
4	2 3	0.24031	0.85513	1.00000	0.63636	0.96364	1.00000	0.07759	0.95690	1.00000	0.09524	0.33333	1.00000
4	3 3	0.62996	0.75513	1.00000	0.83721	0.88372	1.00000	0.55556	0.88889	1.00000	0.47368	0.57895	1.00000
5	1 3	0.22628	0.89051	1.00000	0.85000	0.96667	1.00000	0.10714	0.92857	1.00000	0.16667	0.41667	1.00000
5	2 3	0.48026	0.90132	1.00000	0.90909	0.98485	1.00000	0.31618	0.98529	1.00000	0.10000	0.45000	1.00000
5	3 2	0.50000	1.00000	0.00000	0.75000	1.00000	0.00000	0.23077	1.00000	0.00000	0.00000	0.00000	0.00000
6	1 3	0.70370	0.88889	1.00000	0.92647	0.97059	1.00000	0.10000	0.90000	1.00000	0.12500	0.25000	1.00000
6	2 3	0.42653	0.87447	1.00000	0.52885	0.98077	1.00000	0.09524	0.90476	1.00000	0.06250	0.15625	1.00000
6	3 2	0.63100	1.00000	0.00000	0.52941	1.00000	0.00000	0.66667	1.00000	0.00000	0.00000	0.00000	0.00000
7	1 2	0.73113	1.00000	0.00000	0.90000	1.00000	0.00000	0.14286	1.00000	0.00000	0.00000	0.00000	0.00000
7	2 3	0.15509	0.53172	1.00000	0.66667	0.83333	1.00000	0.05000	0.82500	1.00000	0.40000	0.80000	1.00000
7	3 3	0.60524	0.77341	1.00000	0.50000	0.87500	1.00000	0.62500	0.87500	1.00000	0.58333	0.83333	1.00000
8	1 3	0.63128	0.91620	1.00000	0.90714	0.98571	1.00000	0.16667	0.91667	1.00000	0.58333	0.75000	1.00000
8	2 3	0.57931	0.74253	1.00000	0.83333	0.85417	1.00000	0.05000	0.92500	1.00000	0.38333	0.41667	1.00000
8	3 3	0.51117	0.66201	1.00000	0.69118	0.72059	1.00000	0.41667	0.83333	1.00000	0.18750	0.31250	1.00000
9	1 3	0.41362	0.67745	1.00000	0.84112	0.90654	1.00000	0.12308	0.70769	1.00000	0.08081	0.24242	1.00000
9	2 3	0.43735	0.82064	1.00000	0.89316	0.94444	1.00000	0.06584	0.93827	1.00000	0.11429	0.26667	1.00000
9	3 3	0.26555	0.41173	1.00000	0.59184	0.65306	1.00000	0.02703	0.86486	1.00000	0.08276	0.12414	1.00000
10	1 3	0.52093	0.92558	1.00000	0.85000	0.96667	1.00000	0.10714	0.92857	1.00000	0.16667	0.41667	1.00000
10	2 3	0.64344	0.93841	1.00000	0.90909	0.98485	1.00000	0.31618	0.98529	1.00000	0.10000	0.45000	1.00000
10	3 2	0.64962	1.00000	0.00000	0.75000	1.00000	0.00000	0.23077	1.00000	0.00000	0.00000	0.00000	0.00000

3.9. 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

4. HAPEM5 Output Files

The HAPEM5 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 HAPEM5 program. The final exposure output files contain all of the exposure estimates from a HAPEM5 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 HAPEM5 analysis. Three of the HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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:\HAPEM5\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:\HAPEM5\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 HAPEM5 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 HAPEM5 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
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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 <i>activity</i> file (<i>durhw.txt</i>)
		- number of <i>activity</i> file data records with 1440 minute total
INDEXPOP	record 2	- number of data records (tracts) in the <i>population</i> file (<i>census.txt</i>)
		- number of counties in the <i>population</i> file (<i>census.txt</i>)
COMMUTE	record 3	- number of records in the population index file (<i>census_direct.ind</i>)
		- number of records in the <i>commuting</i> file (<i>comm.txt</i>)
	record 4	- number of records in the work tract file (<i>comm.da</i>)
		- number of records in the commuting index file (<i>comm.ind</i>)

AIRQUAL	record 5	- number of matching tracts in the <i>air quality</i> (e.g., <i>benzene.txt</i>) and population index (<i>census_direct.ind</i>) files
		- number of counties with matching tracts in the <i>air quality</i> (e.g., <i>benzene.txt</i>) and population index (<i>census_direct.ind</i>) files
AIRQUAL	record 6	- number of tracts in the <i>air quality</i> (e.g., <i>benzene.txt</i>)
		- number of data records in the air quality file (e.g., <i>benzene.txt</i>)

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 HAPEM5 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 HAPEM5, since both these pieces of information about a tract (population and air quality) are needed to make an exposure estimate.
- The HAPEM program records information on the *mistract* file only if commuting is included in the simulation. 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, HAPEM5 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 module, 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		

4. HAPEM5 Output Files

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 <i>population</i> 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 <i>air quality</i> 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 $3.767 = 113/30$.

4. HAPEM5 Output Files

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 0.1388E+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. HAPEM5 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 preprocessors 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 HAPEM5

The source code for each of the five HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 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 HAPEM5 program are presented in Table 2.1 as user-defined files.

The HAPEM5 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, HAPEM5 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 HAPEM5 *parameter* files

Keyword	Definition
activity	name of the <i>activity</i> file (input)
cluster	name of the <i>cluster</i> file (input)
ClusTrans	name of the <i>cluster transition probability</i> file (input)
populat	name of the <i>population</i> file (input)
commut	name of the <i>commuting</i> file (input)
quality	name of the <i>air quality</i> file (input)
factors	name of the <i>factors</i> file(input)
statefip	name of the <i>statefip</i> file (input)
Log	name of the <i>log</i> file (output)
counter	name of the <i>counter</i> file (output)
mistract	name of the <i>mistract</i> file(output)
afile	path of final exposure file (output)
product ¹⁰	path of indoor source files (input)
AutoPduct ¹¹	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.

- It categorizes and groups population activity data extracted from CHAD into demographic groups, day types (season, day-of-week), 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.

¹⁰ A path to one or more indoor emission source inputs for the HAPEM5 indoor source algorithms is specified in this statement. These algorithms are included in the HAPEM5 module, 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.

¹¹ The full path name of an existing file must be specified as the *AutoPduct* file in HAPEM5, although its only function is as input to the indoor source algorithms. These algorithms are included in the HAPEM5 module, 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 module.

- It creates a sequential file of the activity pattern records for use by the HAPEM program.

The 10 demographic groups are defined by combinations of the following age and gender categories.

Age:

- 0 - 4
- 5 - 11
- 12 - 17
- 18 - 64
- 65+

Gender:

- Male
- Female

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 and day type. There are 2 or 3 cluster types for each demographic group/day type 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 in section 2.1.2, ***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 3.2.1. 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 day type (indicated as part of the activity record) and demographic group, as defined in the DURAV source code. If any records cannot be classified, an error message is written to the *log* file and the program is stopped. The serial record 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 day type/demographic group category with and without commuting (i.e., time in work location) is determined. Day type/demographic group combinations with more than 10% commuting records are classified as “commuting” activity sets.
- 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/cluster combination, the number of clusters in each demographic group/day type combination, and the commuting classifications for each demographic group/day type combination are recorded in an intermediate file with filename extension “.*nonzero*”.
- The total number of compressed activity records processed and their allocation among day types and demographic groups 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, and cluster type, and the filename is recorded in the *log* file.

5.2.2. INDEXPOP

As explained in section 2.1.3, the INDEXPOP program performs two main functions:

- 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.

INDEXPOP Processing Operations

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

- 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), and the fraction of home tract workers in each work tract.
- It creates sequential index files to facilitate file searching in HAPEM.

COMMUTE Processing Operations

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

- 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 serial record 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. Each work tract ID and its adjusted fraction 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 sequential file with filename extension “.ind”. The fields in the record are the population tract ID, the serial record 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 serial record numbers of the first and last data record for tracts in that state in the “.ind” file.

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 37 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 *ClusTrans* 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 37 microenvironments are read from the *factors* file and saved into arrays.
- For each day type/demographic group combination the number of activity patterns for each cluster, and the commuting status of the set is read from the “.nonzero” file created in DURAV.

- For each day type/demographic group, the frequency of each cluster, and the cluster-to-cluster transition probabilities are read from the *ClusTrans* file.
- For each day type/demographic group/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.
- If commuting is included in the simulation, for each day type/demographic group combination with non-commuting status (10% or less of *activity* file records with activity time at work location), any activity time at a work location in an annual average activity pattern is transferred to the home location.
- 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 is 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.
- If commuting is included in the simulation, indices are read from the “.st_comm1_fip_range” and “.ind” files created by COMMUTE, and data from the “.da” file created by COMMUTE. Otherwise, indices are read from the “.state_air2_fip_range” file created by AIRQUAL.
- If commuting is included in the simulation, 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.
- If commuting is not included, every air quality record from the “.air_da” file created by AIRQUAL is processed.
- 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* file, 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). Each subset contains randomly selected ME factors for each of 37 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.
- 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, **nreplic** sets of average activity patterns are calculated as weighted averages of activity patterns randomly selected from each cluster in a demographic group/day type 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.
- If commuting is included in the simulation, for each day type/demographic group combination with commuting status, 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 is 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 is used. The air quality data is 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.
- 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 if commuting is included) to estimate **nreplic** exposure concentrations for each day type/demographic group combination. A separate set of estimates is made for each emission source category. The algorithm for each day type/demographic group combination in the tract is as follows.

$$ExpConc = \frac{\sum_{TimeBlocks} \sum_{Microenvironments} Conc_{t,m} \times Duration_{t,m}}{\sum_{TimeBlocks} \sum_{Microenvironment} Duration_{t,m}}$$

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. Tippet, 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: Development of Algorithm for Creating Longitudinal Activity Patterns with Cluster Analysis

This page intentionally left blank.



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-t² 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.¹² 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 ix	=	number of pairs of consecutive days for the same individual where the first day is in cluster i .
	=	trans $i1$ + trans $i2$ + trans $i3$
prob ij	=	trans ij / trans ix

There are trans ix days where an individual is in cluster i on one day and where the next day is in the database. Of those trans ix 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 .

¹² 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 $0.5 / \text{trans } i$, 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:

$$\text{steady}_j = \text{limiting probability for cluster } j = \sum_i \text{steady}_i \times \text{prob}_{ij}$$

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:

$$\text{expdays}_i = \text{steady state expected number of days for cluster } i$$

$$\text{low}_i = 5^{\text{th}} \text{ percentile of the simulated number of days for cluster } i$$

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_{*ijj*} 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