The HAPEM User's Guide Hazardous Air Pollutant Exposure Model, Version 7

July 2015

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

Prepared by:
Chris Holder, Isaac Warren,
Jonathan Cohen, Casey Cavanagh,
Angelica McGee, Chris Stevens,
Arlene Rosenbaum, Kevin Wright,
Blanche Dean, and Courtney Skuce
ICF International
2635 Meridian Pkwy., Suite 200,
Durham, NC 27713

This page intentionally left blank.

Contents

Figures

Tables

This page intentionally left blank.

New Features in HAPEM7

The Hazardous Air Pollutant Exposure Model, version 7 (HAPEM7) includes a number of updated features. These updated features better reflect the residential locations, work locations, commuting habits, and activity patterns of the current (2010) U.S. population, and they also are designed to provide exposure estimates that better characterize the variability across the population. These updated features are summarized in the list below, and detailed in other portions of this User's Guide.

- Data on population, commuting patterns, and residential proximity to major roads have been updated based on information from the 2010 U.S. Census, including information for the U.S. Virgin Islands.
- Activity-pattern data have been updated based on the June 2014 version of CHAD-Master.
- As a result of disaggregating the previously-included microenvironments, four new commuting-related microenvironments are included in HAPEM7 for a total of 18 microenvironments.

This page intentionally left blank.

1. Introduction

The Hazardous Air Pollutant Exposure Model, version 7 (HAPEM7) User's Guide is designed to assist exposure analysts with running and interpreting results from HAPEM7. Throughout the User's Guide, for easier identification, the input filenames and file types are in italics (usually lowercase), model program names are uppercase underlined, and model variables are 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.

Most of the material in this HAPEM7 User's Guide was taken from the HAPEM6 User's Guide (Rosenbaum and Huang, 2007).

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 HAPEM7, 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 HAPEM7. 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 HAPEM7.

Chapter 1 Introduction. Provides a brief introduction to HAPEM7 modeling fundamentals, including a brief history of the development of HAPEM7.

Chapter 2 Getting Started—An Overview of HAPEM7. Provides an overview of the various components of HAPEM7 and basic information needed to run the model.

Chapter 3 HAPEM7 Input Files. Provides a description of the format, data, and options for each HAPEM7 input file.

Chapter 4 HAPEM7 Output Files. Provides a description of the format and data associated with each HAPEM7 output file.

Chapter 5 HAPEM7 Programs. Provides a description of the purpose, operations, inputs, and outputs, including a brief description of the computer code, for each HAPEM7 computer program.

Chapter 6 References.

Appendix A A 2015 technical memorandum from ICF International to Ted Palma, providing a description of the process of developing the activity-pattern clusters and cluster transitions for HAPEM7.

Appendix B A 2015 technical memorandum from ICF International to Ted Palma, providing a thorough description of the process of updating the default input files, microenvironments, and model source code for HAPEM7.

1.2. Background

The Hazardous Air Pollutant Exposure Model, version 7 (HAPEM7) is a screening-level exposure model appropriate for assessing average long-term inhalation exposures of the general population, or a specific sub-population, over spatial scales ranging from urban[1] to national. HAPEM7 provides a relatively transparent set of exposure assumptions and approximations, as is appropriate for a screening-level model.

HAPEM7 uses the general approach of tracking representatives of specified age groups as they move among indoor and outdoor microenvironments (MEs) and among geographic locations. The estimated HAP concentrations in each ME visited are combined into a time-weighted average concentration, which is assigned to members of the age group.

A microenvironment (**ME**) 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.

HAPEM7 uses four primary sources of information: population data from the U.S. Census Bureau (Census), population activity data from the U.S. Environmental Protection Agency (EPA) Consolidated Human Activity Database (CHAD), air quality data, and ME data. These data will be discussed briefly below, and in greater detail later in this User's Guide.

1.2.1. Population Data

The census is the primary source of most population demographic data. The census collects, among other things, information on where people live, their demographic makeup (e.g., age, gender, ethnic group; note that only age is used in HAPEM7), employment (which is not explicitly used in HAPEM7), and commuting behavior. The default population data for HAPEM7 for all areas except the U.S. Virgin Islands are derived from Table PCT12 in the 2010 Census data reported at the spatial resolution of census tracts. Census tracts are small, relatively permanent statistical subdivisions of a county and usually contain between 2,500 and 8,000 residents. Population data for the U.S. Virgin Islands were not available from the 2010 Census Table PCT12, and were instead gathered from other census surveys.

A second type of population data used in HAPEM7 is an estimate of the fraction of the population of each census tract that lives within certain distances of major roadways. These estimates were derived using geospatial software to perform proximity analyses on roadway location data from the 2013 Census TIGER/Line database (see Appendix B for more details on the HAPEM7 default input files). They are used, in conjunction with the *PROX* factors described below, to account for the enhanced outdoor concentrations of HAPs emitted from onroad vehicles at locations near major roadways, and the associated enhanced indoor concentrations.

1.2.2. Activity Data

HAPEM7 uses four types of population activity data: activity-pattern data, commuting-flow data, commuting-time data, and commuting-fraction data. Human activity pattern data are used to determine the frequency and duration of exposure for specific groups within various MEs.

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

Activity-pattern data are taken from demographic surveys compiled in CHAD 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 race/ethnicity 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. For the purposes of the HAPEM7 default files, age is the only CHAD demographic, respondent attribute information used, along with location, although the activity input file includes gender and race/ethnicity.

The default population activity file for HAPEM7 is derived from a database of activity pattern surveys called CHAD (McCurdy et al. 2000). The version of CHAD current in June 2014 (i.e., the version updated in July 2013) was used in HAPEM7 is composed of over 45,000 person-days of activity pattern data, including 137 specific activities (out of a possible 142 specific activities to choose from) and 43 specific locations (out of a possible 113 specific locations to choose from), collected and organized from 21 human activity pattern surveys. The CHAD contains the sequential patterns of activities for each individual, and each activity event has a corresponding location code so that the ME of each activity event is known. The default population activity file includes a commuting-status indicator which was calculated to define the daily commuting status as 1 if the daily activity pattern includes any minutes in a work activity. The ME categories currently incorporated into the default population activity file for HAPEM7 are presented in Table 1-1 (see Appendix B for more details on the HAPEM7 default input files, Appendix A for more details on the cluster analysis of the activity data).

Table 1-1. HAPEM7 MEs

ME Number	ME Description	Broader ME (for clustering)	Commuting?
1	Residential	1 Indoors Residence	No
2	School	2 Indoors Other	No
3	Hospital	2 Indoors Other	No
4	Office	2 Indoors Other	No
5	Public Access	2 Indoors Other	No
6	Bar/Restaurant	2 Indoors Other	No
7	Car/Truck	5 In-vehicle	Yes - Private Transit
8	Public Transit	5 In-vehicle	Yes - Public Transit
9	Air Travel	2 Indoors Other	No
10	Waiting Indoors for Public Transit	2 Indoors Other	Yes - Public Transit
11	Waiting Outdoors for Public Transit	3 Outdoors Near-roadway	Yes - Public Transit
12	Motorcycle/Bicycle	3 Outdoors Near-roadway	Yes - Private Transit
13	Ferryboat	4 Outdoors Other	Yes - Public Transit
14	Residential Garage	3 Outdoors Near-roadway	No
15	Outdoors, Near Roadway	3 Outdoors Near-roadway	No
16	Outdoors, Service Station	3 Outdoors Near-roadway	No
17	Outdoors, Parking Garage	3 Outdoors Near-roadway	No
18	Outdoors, Other	4 Outdoors Other	No

Because available activity data are not adequate to estimate the exposure of each individual in a population, HAPEM7 groups activity-pattern data together for people with similar demographic characteristics that are expected to influence exposure to HAPs (e.g., age and commuting status), and makes exposure estimates for these groups. The activity profiles for each person in an age group have an equal chance of being selected from the activity database (see Section 1.2.5 [Stochastic Elements]). The result is that HAPEM7 provides a distribution of exposure concentrations for each age group in each census tract.

HAPEM7 divides the population into six age groups. Activity pattern data are also separated into three day types (summer weekdays, other [non-summer] weekdays, and weekends), and commuting status (yes or no).

The commuting-flow data contained in the HAPEM7 default file were derived by the U.S. Department of Transportation Federal Highway Administration (FHWA) from the 2010 Census, as part of the Census Transportation Planning Package (CTPP) and commissioned by the American Association of State Highway and Transportation Officials (see the FHWA "Census Issues" web site http://www.fhwa.dot.gov/planning/census_issues/). The data files specify the number of residents of each census tract that work in that tract and every other tract (i.e., the population associated with each home-tract/work-tract pair). For HAPEM7, the distance between the centroids of the home and work tracts were calculated outside of the CTPP, using the 2010 Census Gazetteer spatial files and the Great Circle distance equation (see Appendix B for more details on the HAPEM7 default input files). HAPEM7 uses these data in coordination with the activity-pattern data to place an individual who commutes to work either in the home tract or the work tract at each time step.

For each census tract, the HAPEM7 default commuting-time file contains the proportion of commuting workers using public transit and the proportion using private transit, and it also contains the average commute time stratified by public or private transit, as derived from the 2010 Census (American Community Survey [ACS] tables B08301, C08134, and C08136; see Appendix B for more details on the HAPEM7 default input files). These data are combined with data on the centroid-to-centroid distances between tracts (see Section 5.2.5 [HAPEM]) to estimate the commuting time for each commuting replicate.

Data specifying the fraction of each age group in each census tract that commute to work, as contained in the HAPEM7 default commuting-fraction file, were derived from the 2010 Census (ACS tables B23001 and B08101; see Appendix B for more details on the HAPEM7 default input files).

1.2.3. Air-quality Data

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

HAPEM7 is able to estimate exposures over the entire US at spatial scales as small as a census tract. In order to preserve any characteristic diurnal patterns in ambient concentrations that might be important in the estimation of population exposure, HAPEM7 can treat annual-

[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.

average concentration estimates that are stratified by time of day in the air-quality input file. The time steps in the air-quality data must be an integral factor of the number of time steps in the activity input file (see Section 2.1.2 [The DURAV Program and the *Activity* and *Cluster* Files]). For example, the HAPEM7 default activity file contains data in (24) 1-hour time blocks, so an air-quality file used with the default activity file must contain data in (24) 1-hour time blocks, or (12) 2-hour time blocks, or (8) 3-hour time blocks, and so on. The air-quality data are combined in HAPEM7 with activity data to estimate exposure concentrations. The air-quality data can also be decomposed to reflect the contributions from various emission sources. The number of sources is a user-specified variable.

HAPEM7 is also able to incorporate spatial variability of air quality within each census tract. That is, the air quality within a tract is not limited to a single point estimate (diurnally- and source-stratified). Spatial variability may be incorporated in two different ways. One method is to characterize the air quality in a census tract by a set of up to 500 diurnally- and source-stratified values. How HAPEM7 handles this data set is explained below in Section 1.2.5 (Stochastic Elements).

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

1.2.4. ME Data

In order to calculate the exposure concentration for each demographic group (i.e., for each age group in HAPEM7), an estimate is required of the concentration in each ME specified by the activity pattern. In HAPEM7, these ME concentration estimates are derived from the outdoor concentration estimate for the census tract and a set of 3 ME factors: *PEN*, *PROX*, and *ADD*. These respectively account for penetration of outdoor air into the ME, concentration enhancement due to proximity of the ME to the emission source, and emission sources within the ME (note that *ADD* factors are currently set to zero, as discussed in Section 2.1.6 [The HAPEM Program, the ME *Factors* and *Mobiles* Files, and the Activity *Cluster-transition* File]).

The ME factors are entered into the model as data from input files that contain estimates of distributions for *PEN*, *PROX, and ADD* for three phases of HAPs: gases, particles, and HAPs that might be either phase depending on various conditions. The HAPEM7 default *PEN* distributions were obtained from an extensive review of literature and databases on indoor/outdoor ratios of HAPs. The HAPEM7 default *PROX* distributions for onroad-mobile sources were derived from modeling studies of the concentration gradients of HAPs near major roadways.[3] How the distributions are utilized in HAPEM7 is discussed below in Section 1.2.5 (Stochastic Elements).

As is the case with all other HAPEM7 input files, these data can be modified by the user. The ME factors should be updated as needed to reflect current knowledge, as available.

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

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. HAPEM7 incorporates six stochastic elements.

1.2.5.1. Commuting Status

The first stochastic element in the construction of a replicate is the determination of the commuting status (yes or no), according to the probabilities specific to census tracts and demographic groups (e.g., age groups in HAPEM7).

1.2.5.2. Activity Patterns

The second stochastic element is the selection of daily activity patterns to represent the demographic group (e.g., age group in HAPEM7) and commuting status of the replicate. HAPEM7 estimates long-term average concentrations, but the available sequences of population activity data are specified for 24-hour periods only. The general approach used by HAPEM7 for constructing long-term average activity sequences from short-term records is composed of several steps (see Appendix A for a detailed discussion, which is briefly summarized here). 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 the three specified HAPEM7 day types. 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., 65 for summer weekdays, 196 for other weekdays, and 104 for weekends). This process is repeated for several replicates[4] for each combination of census tract and demographic group (e.g., age group in HAPEM7), to create a set of annual exposure-concentration estimates for each group in each census tract.

To implement this approach, first all the activity pattern data are grouped according to demographic group (e.g., age group in HAPEM7), day type, and commuting status. Then, for each group/commuting-status/day-type combination, the activity patterns are stratified into from one to three categories, based on similarity of time spent in the various MEs, as determined by cluster analysis (see Appendix A for a detailed discussion on clustering).

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.

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

The HAPEM7 algorithms construct an average activity pattern for each replicate 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.

1.2.5.3. Work Tract

Another stochastic process is applied in HAPEM7 for replicates that commute to work. For those groups, a work census tract is selected at random from the set of work tracts specified for that home tract, using the proportion of workers commuting to each work tract for its selection probability.

1.2.5.4. ME Factors

Another stochastic feature of HAPEM7 is the ability to characterize ME factors as variable, instead of uniform over the population. That is, three of the four ME factors (*PEN*, *PROX*, and *ADD*) are represented by probability distributions rather than point estimates.[5] Several distribution types may be used, as discussed in Section 3.10 (ME *Factors* and *Mobiles* Files). For each replicate, a different set of ME factors is randomly selected.

1.2.5.5. Air Quality—General

HAPEM7 has the ability to characterize outdoor air concentrations as spatially variable within a census tract. It can do this in two different ways. One approach is to characterize the air quality for each tract as a data set with up to 500 sets of values (i.e., diurnally- and source-stratified). Then, for each replicate, a different set of ambient air concentrations is selected for the home (and work) tract to reflect the spatial variability in air quality within the tract.

1.2.5.6. Air Quality—Onroad Vehicle Related

When air quality is characterized by a single point estimate (diurnally- and source-stratified), another approach is used to account for enhanced onroad-vehicle-related HAP concentrations in the vicinity of major roadways. To implement this approach, the distance of the replicate's home (and workplace) from a major roadway is randomly selected based on probabilities specific to census tracts and demographic groups (e.g., age groups in HAPEM7). A *PROX*

[5] As noted above, in practice the default *PROX* values for emission source categories other than onroad vehicles are point values of 1.0 (i.e., no concentration enhancement due to proximity), and the default *ADD* values are point values of 0.0 (i.e., no indoor emission sources). However, HAPEM7 contains the structure to characterize these as distributions if appropriate data are available.

factor is then selected from a distribution and applied to the census-tract air-quality values for onroad-mobile sources, with the distribution dependent on the selected distance.

1.3. Strengths and Limitations of HAPEM7

All models have strengths and limitations. Therefore, for each application, it is important to carefully select the model that has the desired attributes. The following sections provide a summary of the strengths and potential limitations of HAPEM7. However, this is not an exhaustive list and may not address features important for specific applications of an exposure model.

1.3.1. Strengths

One strength of HAPEM7 is the ability to use air-concentration estimates from modeling, allowing exposure to population groups to be simulated at the census-tract level rather than relying solely data from the limited (in both areal extent and HAPs measured) nationwide network of fixed-site monitors.

Another important feature of HAPEM7 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., ME factors) used by the model, as they become available.

Another strength of HAPEM7 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 HAP concentrations in these locations may be very different.

Another important feature of HAPEM7 is the incorporation of stochastic processes for the selection activity patterns, work census tracts, ambient air quality among locations within a tract, and 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 HAPEM7 has also been facilitated by development of default input files derived from the databases discussed above: national census population and commuting information, CHAD activity data, and variable ME factors for gases, particles and those HAPs that might be either gaseous or particulate depending on conditions.

1.3.2. Limitations

HAPEM7 calculates long-term average exposure concentrations in order to address exposures to HAPs with carcinogenic and other long-term effects. Thus, HAPEM7 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 HAPs due to activities that are related in time is not preserved.

HAPEM7 only estimates exposures experienced through inhalation. For certain HAPs, inhalation might not be the major route of exposure, and, therefore, HAPEM7 may underestimate exposures in these instances. Also, although HAPEM7 is an inhalation-exposure model, it does not include any measures of the ventilation rate associated with an activity, so there is no ability to calculate the potential dose received when engaging in various activities.

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

- The population activity pattern data are limited. Only three of the 21 studies in the version of CHAD used for HAPEM7 were national in scope (with two other studies covering multiple metropolitan areas); therefore, the combined data set does not constitute a representative sample, at least with respect to geographic region.
- Commuting pattern data addresses only home-to-work travel. The population not employed outside the home is assumed to always remain in the residential census tract. Further, although several of the HAPEM7 MEs account for time spent in travel, the travel is assumed to always occur either in the home or work tract. No provision is made for the possibility of passing through other tracts during travel.
- The ME *PEN* factor distributions incorporated into HAPEM7 were derived from reported measurement studies. The data available were quite limited. As a result, most factors were not derived from a representative sample of measurements, and many were inferred on the basis of measurements of different HAPs and/or MEs that would be expected to be similar. In addition, the derivation of the *PEN* factors was based on the assumption that measured I/O ratios of 1.0 or less indicate the absence of indoor emission sources. Because this assumption is unlikely to be uniformly valid, *PEN* factors are likely to overestimate penetration by some unknown amount.
- The ME *PROX* factor distributions incorporated into HAPEM7 for the onroad-vehicle source category were derived from modeling studies for Portland, Oregon. They are subject to the standard uncertainties of air-dispersion modeling. They are also subject to the uncertainties of extrapolating from the traffic patterns of Portland to other locations.
- Air-quality data from modeling studies are uncertain, due to simplifications incorporated into modeling algorithms and limitations of input data (e.g., emissions, meteorology). Air-quality measurements are also uncertain due to limitations of measurement technology (e.g., minimum detection limits) and unknown representativeness of monitoring locations.

1.4. Applicability

HAPEM7 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, HAPEM7 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 HAP concentrations and/or activities that, by their nature, could result in potentially high exposures (e.g., occupational exposures). Furthermore, HAPEM7 cannot address cumulative exposure from multiple HAPs nor HAP mixtures.

1.5. Brief History of the Hazardous Air Pollutant Exposure Model

In 1985, the EPA's Office of Mobile Sources (OMS)[6] developed a model for estimating human exposure to nonreactive pollutants emitted by mobile sources. This model was similar to the probabilistic National Ambient Air Quality Standards Exposure Model (pNEM) in that both simulated the movements of population groups between home and work locations and through various MEs. They differed, however, in several respects. The pNEM provided minute-by-minute exposure estimates, which could be averaged over longer time periods, whereas the model now known as 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.

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 thereafter, 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 MEs 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 Census, and developing an algorithm for estimating ambient impacts in residences with attached garages.

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

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

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 US at the census-tract level. With these improvements, the model was able to directly estimate exposures to HAPs, and hence the model was again renamed by dropping the mobile source (-MS) acronym.

An earlier version of the model, HAPEM4, had other enhancements as well. These included broader flexibility in defining the study area (this can range from a single census tract up to the entire contiguous US), 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 MEs.

EPA used HAPEM4 in its National-scale Air Toxics Assessment (NATA) for 1996, an assessment that is updated periodically, is designed to help assess the prevalence of air toxics in the US, and is an important part of EPA’s Integrated Urban Air Toxics Strategy.

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

HAPEM6 included the ability to account for enhanced onroad-vehicle-related HAP concentrations in the vicinity of major roadways, a more accurate characterization of the fraction of the population of each census tract that commutes to work, and a more accurate estimate of the duration of commuting to work.

HAPEM7 is not fundamentally different from HAPEM6. HAPEM7 includes updates to all census- and CHAD-related data in the default input files, and it now includes 18 default microenvironments (up from 14 in HAPEM6). HAPEM7 was applied as part of the NATA for 2011.

NOTE: HAPEM currently contains enhanced algorithms for estimating exposure concentrations from indoor emission sources. However, the algorithms have undergone only limited testing, and the development is not complete of the databases required to implement these algorithms. Therefore, we do not recommend the use of these algorithms at the present time.

This page intentionally left blank.

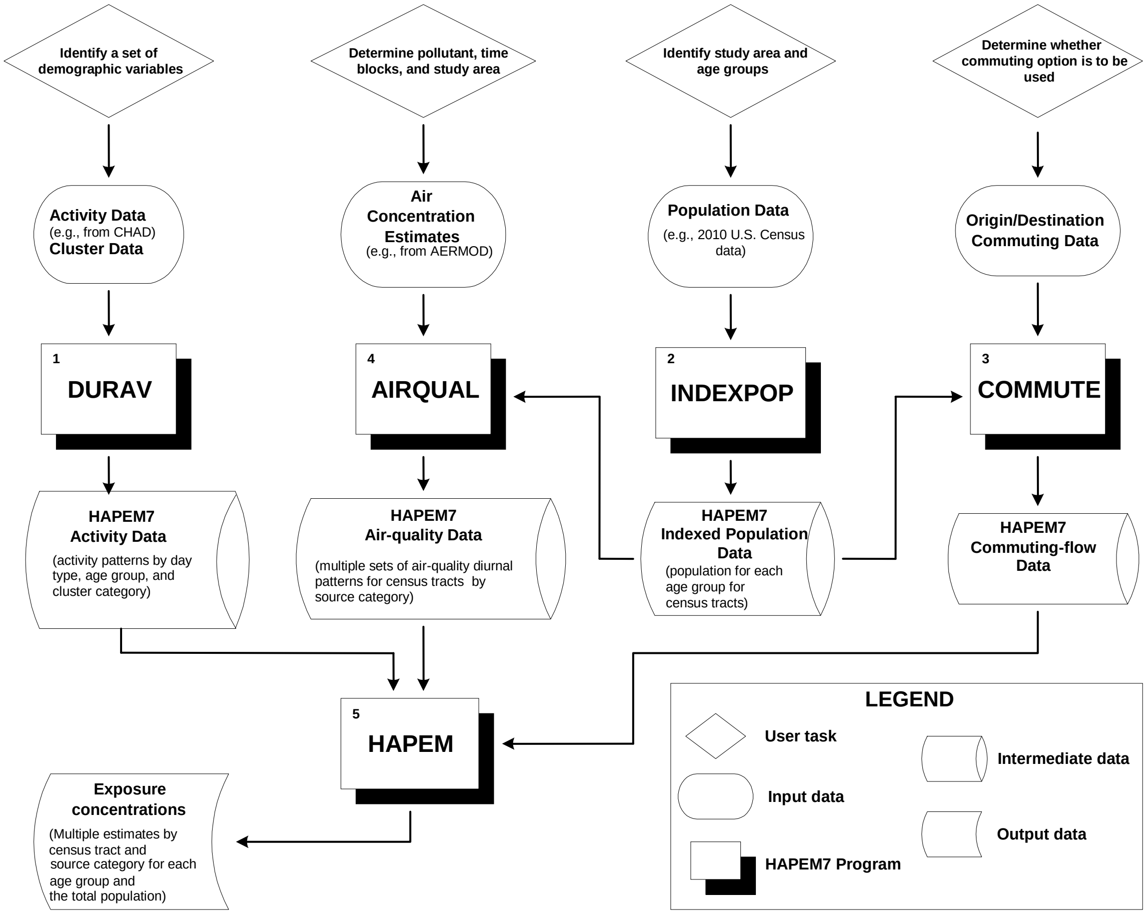
2. Getting Started—An Overview of HAPEM7

This chapter provides the user the basic information needed to run HAPEM7. The topics addressed in this chapter include the functions of the programs that are contained in HAPEM7, the contents of the various input and output files, and the meanings of parameter values. The chapter has been separated into the following sections.

Section 2.1	Model Structure. Describes the general structure of HAPEM7, the input and output files, and the parameter settings.
Section 2.2	Changing the Parameter Settings. Discusses considerations for changing parameter settings.
Section 2.3	Setting Up a HAPEM7 Run. Provides instructions for setting up and running HAPEM7.

Figure 2-1 presents a graphical overview of HAPEM7, 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.

Figure 2-1.
Overview of HAPEM7


Identify a set of demographic variables
Determine pollutant, time blocks, and study area
Identify study area and age groups
Determine whether commuting option is to be used
Activity Data
(e.g., from CHAD)
Cluster Data
Air Concentration Estimates
(e.g., from AERMOD)
Population Data
(e.g., 2010 U.S. Census data)
Origin/Destination Commuting Data
1
DURAV
4
AIRQUAL
2
INDEXPOP
3
COMMUTE
HAPEM7 Activity Data
(activity patterns by day type, age group, and cluster category)
HAPEM7 Air-quality Data
(multiple sets of air-quality diurnal patterns for census tracts by source category)
HAPEM7 Indexed Population Data
(population for each age group for census tracts)
HAPEM7 Commuting-flow Data
5
HAPEM
Exposure concentrations
(Multiple estimates by census tract and source category for each age group and the total population)
LEGEND
User task
Input data
HAPEM7 Program
Intermediate data
Output data

2.1. Model Structure

HAPEM7 contains five programs. These are listed below.

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

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

For a given modeling domain (e.g., a state, a set of states, the entire US), programs 1–3 need to be executed only once, even if several different air-quality scenarios/HAPs are evaluated. Programs 4–5 need to be executed one time each for each air-quality scenario/HAP. The modeling domain for running programs 4–5 must be included in the modeling domain used for running Programs 1–3, but it 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 4–5 for evaluating a single state or set of states.

The model programs use 12 groups of user-supplied input data files, and two or more *parameter* files. All are in American Standard Code for Information Exchange (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 information required in the *parameter* files is presented in Table 2-1 in a way that shows what information is supplied by user-defined files, what is supplied by user-defined parameters, and which model program requires the information. With one exception, noted below, any information in the *parameter* files that is not 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 model program, 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., for 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 programs

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.

1–3 so that they will stop if the keyword ***nreplic*** (required by the AIRQUAL and HAPEM programs) is encountered in the *parameter* file.

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

Table 2-1.
Keywords for *parameter* files and example filenames

User/Model Defined	Inputs	Outputs
DURAV.f90		
User-defined files	*activity* file (e.g., *activity_CHAD_v7.txt*) *cluster* file (e.g., *activity_cluster_v7.txt*)	*log_file.txt* *counter.dat*
User-defined parameters	***nmicro*** ***hblock*** ***ngroup*** ***nblock*** ***ntype***	
Model-defined files		*activity_CHAD_v7.wrong_chad* *activity_CHAD_v7.da* *activity_CHAD_v7.nonzero*
INDEXPOP.f90		
User-defined files	*population* file (e.g., *population_v7.txt*) *distance-to-road* file (e.g., *proximity_road_v7.txt*) *commuting-time* file (e.g., *commute_time_v7.txt*) *commuting-fraction* file (e.g., *commute_fraction_v7.txt*) *statefip* file (e.g., *FIPS_StateCrosswalk_v7.DAT*)	*log_file.txt* *counter.dat*
User-defined parameters	***region1*** ***region2*** ***ngroup***	
Model -defined files		*population_v7.da* *population_v7_direct.ind* *population_v7.county_tract_pop_range* *population_v7.state_county_pop_range* *proximity_road_v7.STIDX* *proximity_road_v7.dat* *commute_time_v7.STIDX* *commute_time_v7.dat* *commute_fraction_v7.STIDX* *commute_fraction_v7.dat*
COMMUTE.f90		
User-defined files	*commuting* file (e.g., *commute_flow_v7.txt*) *population* file (e.g., *population_v7.txt*) *distance-to-road* file (e.g., *proximity_road_v7.txt*) *commuting-time* file (e.g., *commute_time_v7.txt*) *commuting-fraction* file (e.g., *commute_fraction_v7.txt*) *statefip* file (e.g., *FIPS_StateCrosswalk_v7.DAT*)	*log_file.txt* *counter.dat* *mistract.dat*
User-defined parameters	***region1*** ***region2*** ***keep***	
Model-defined files	*population_v7_direct.ind* *population_v7.county_tract_pop_range* *population_v7.state_county_pop_range* *proximity_road_v7.STIDX* *proximity_road_v7.dat* *commute_time_v7.STIDX* *commute_time_v7.dat* *commute_fraction_v7.STIDX* *commute_fraction_v7.dat*	*commute_flow_v7st_comm1_ fip_ range* *commute_flow_v7.da* *commute_flow_v7.ind*

User/Model Defined	Inputs	Outputs
	AIRQUAL.f90	
User-defined files	*air quality* file *population* file (e.g., *population_v7.txt*) *distance-to-road* file (e.g., *proximity_road_v7.txt*) *statefip* file (e.g., *FIPS_StateCrosswalk_v7.DAT*)	*log_file.txt* *counter.dat* *mistract.dat*
User-defined parameters	***hblock*** ***ngroup*** ***region2*** ***nsource*** ***region1*** ***nreplic***	
Model-generated files	*population_v7.da* *proximity_road_v7.STIDX* *population_v7_direct.ind* *proximity_road_v7.dat*	*HAP.state_air_fip_range* *HAP.state_air1_fip_range* *HAP.state_air2_fip_range* *HAP.da* *HAP.pop_air_da* *HAP.air_da*
	HAPEM.f90	
User-defined files	*factors* file (e.g., *factors_*_v7.txt*) *mobiles* file (e.g., *factors_OnroadMobile_*_v7.txt*) *population* file (e.g., *population_v7.txt*) *air quality* file *commuting* file (e.g., *commute_flow_v7.txt*) *activity* file (e.g., *activity_CHAD_v7.txt*) *cluster-transition file* (e.g., *activity_ClusterTransition_v7.txt)* *Product* files[1] (specify path only) *AutoPduct* file[1]	*log_file.txt* *mistract.dat* *counter.dat* *afile* file (path only)
User-defined parameters	***pollutant*** ***backg*** ***Rseed1*** ***CAS***[2] ***sarod*** ***Rseed2*** ***unit*** ***nmicro*** ***Rseed3*** ***EPA*** ***hblock*** ***B_00*** ***nmobiles*** ***ntype*** ***B_02*** ***nemicro*** ***ngroup*** ***B_05*** ***nbmicro*** ***nsource*** ***B_16*** ***nvehicles*** ***nreplic*** ***B_18*** ***npublict*** ***region1*** ***B_65*** ***year*** ***region2***	
Model-generated files	*commute_flow_v7st_comm1_fip_ range* *commute_flow_v7.da* *HAP.air_da* *commute_flow_v7.ind* *HAP.pop_air_da* *activity_CHAD_v7.da* *HAP.state_air_fip_range* *activity_CHAD_v7.nonzero* *HAP.state_air1_fip_range* *HAP.da*	

Note: some entries in the above table are presented side-by-side instead of down the page, in order to save space

[1] A path to one or more indoor emission source inputs for the indoor source algorithms is specified in these statements (with the AutoPduct statement including a filename). These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword **CAS** to 99999, and specify any existing path (and file for AutoPduct; other than those otherwise specified for input or output for the HAPEM program) since no indoor source files will then actually be utilized by the HAPEM program.

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

In order for a record in the *parameter* file to be processed by the program, it must contain an equal sign (“=”). 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 is specified a full file pathname (all files except the *final exposure* output files and the indoor source files), a pathname (the *final exposure* output files and the indoor source files[7]), or a parameter value. As currently configured, HAPEM7 creates an exposure output file for each state/HAP combination. The names of these files are constructed by the program based on the HAP’s SAROAD code and the state’s Federal Information Processing Standard (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 HAP in the *parameter* file of the HAPEM program as the value for the parameter ***sarod***.

The names of the other user-defined input and output files should consist of two parts, separated by a dot (“.”). The part of the name preceding the dot, including the path, is the root and the part following the dot is the extension. Note that the program will not process a record in the *parameter* file that is longer than 120 characters, including the key word/phrase, the equal sign, and the filename/path or parameter value. The number of spaces between the keywords and the “=” signs and between the “=” signs and the filenames are not fixed and therefore can be any reasonable number. Figure 2-2a and 2-2b present example *parameter* files that can be used to run programs 1–3 and 4–5, respectively. Note that the input and output filenames must be listed before the parameter settings.

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

```
INPUT FILES:
  activity file    = input/activity pattern/activity_CHAD_v7.txt
  cluster file     = input/activity pattern/activity_cluster_v7.txt
  population file  = input/population/population_v7.txt
  commuting file   = input/commute/commute_flow_v7.txt
  CommutTime file  = input/others/commute_time_v7.txt
  CommutFrac file  = input/others/commute_fraction_v7.txt
  DistToRoad file  = input/others/proximity_road_v7.txt
  statefip file    = input/FIPS_StateCrosswalk_v7.dat
OUTPUT FILES:
  log file         = output/log_file.txt
  counter file     = output/counter.dat
  mistract file    = output/mistract.dat
PARAMETER SETTINGS:
  keep       = YES
  region1    = 1
  region2    = 53
  nmicro   = 18     ! Number of MEs
  nblock   = 24     ! Number of time blocks/day in CHAD file
  hblock   = 8      ! Number of time blocks/day for analysis
  ntype    = 3      ! Number of day types
  ngroup   = 6      ! Number of demographic (age) groups
```

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

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

```
INPUT FILES:
  activity file    = input/activity pattern/activity_CHAD_v7.txt
  ClusTrans file   = input/Activity Pattern/activity_ClusterTransition_v7.txt
  population file  = input/population/population_v7.txt
  commuting file   = input/commute/commute_flow_v7.txt
  air quality file = input/airqual/ALL_benzene_HAPEM7-2.txt
  factors file     = input/factor/factors_gas_v7.txt
  mobiles file     = input/factor/factors_OnroadMobile_Benzene_v7.txt
  CommutTime file  = input/others/commute_time_v7.txt
  DistToRoad file  = input/others/proximity_road_v7.txt
  statefip file    = input/FIPS_StateCrosswalk_v7.dat
  product file Pathname     = input/Add/
  AutoPduct file            = input/Add/AutoGarage.txt
Demographic Groups:
  B_00  =  Ages 0-1
  B_02  =  Ages 2-4
  B_05  =  Ages 5-15
  B_16  =  Ages 16-17
  B_18  =  Ages 18-64
  B_65  =  Ages >= 65
OUTPUT FILES:
  log file                 = output/log_file.txt
  counter file             = output/counter.dat
  mistract file            = output/mistract.dat
  afile                    = output/
  bfile                    = output/
  Keep intermediate files  = YES
PARAMETER SETTINGS:
  pollutant  = Benzene
  CAS        = 99999
  units      = ug/m3
  year       = 2011
  region1    = 1
  region2    = 53
  EPA Region = 1
  sarod      = 45201
  Rseed1   = -10    ! Random seed (negative) for selecting activity pattern data
  Rseed2   = -1     ! Random seed (negative) for selecting micro factors
  Rseed3   = -1     ! Random seed (negative) for selecting AQ dataset
  backg    = 0.0
  nmicro   = 18     ! Number of MEs
  nblock   = 24     ! Number of time blocks/day in CHAD file
  hblock   = 8      ! Number of time blocks/day for analysis
  ntype    = 3      ! Number of day types
  ngroup   = 6      ! Number of demographic (age) groups
  nsource  = 4      ! Number of source categories
  nmobiles = 3      ! Sequence # of onroad mobile source categories
  nbmicro  = 1      ! Beginning sequence # of MEs which are indoor environments
  nemicro  = 10     ! Ending sequence # of MEs which are indoor environments
  nvehicles= 7 12   ! Sequence # of MEs which are for private-transit commuting
  npublict = 8 10 11 13 ! Sequence # of MEs which are for public-transit commuting
  nreplic  = 30     ! Number of replicates to model per demographic (age) group
```

The model programs also create several intermediate output files that are used as input to other programs in the model set, but are not directly useful for the user. The model programs generate the names of the intermediate output files by changing the filename extensions (i.e., the text after the dot) of the input filenames. 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 (*population_v7.county_tract_pop_range* and *population_v7.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 model 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 census 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 the model 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.

2.1.2. The DURAV Program and the *Activity* and *Cluster* Files

The DURAV program performs the three main functions listed below.

- It categorizes and groups population activity data extracted from CHAD into demographic groups (e.g., age groups in HAPEM7), day types, commuting status, 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, currently *activity_CHAD_v7.txt*, contains data extracted from CHAD describing the amount of time spent in various MEs by individuals. Each record in the *activity* file consists of one person-day (i.e., 1,440 minutes for an individual) of activity data. This information is not an activity sequence; rather, it is the total number of minutes spent in each ME during each block of time throughout the day (i.e., the time increments used per 24-hour period).

For example, in the HAPEM7 default *activity* file, *activity_CHAD_v7.txt*, there are 18 MEs, (24) 1-hour time blocks, and 2 exposure districts (home and work), resulting in a total of 864 duration values. The duration in each of the 18 MEs for the first hour comes first in the *activity* file, followed by the 18 durations for the second hour, etc. This pattern is repeated for all 24 hours for the home exposure district, and then for the 24 hours and 18 MEs of the work district see Appendix B for more details on the HAPEM7 default input files).

The number of time blocks in the *activity* file is specified by the user in the *parameter* file of the DURAV program as ***nblock***. The number of MEs in both the *activity* file and the *factors* and *mobiles* files (discussed below) must be the same and is specified in the *parameter* files of the DURAV and HAPEM programs as ***nmicro***.[8] 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 a whole number (i.e., no decimals; this number can be zero) of minutes in each ME.

The number of time blocks for the analysis is specified in the *parameter* files of the DURAV, AIRQUAL, and HAPEM programs 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, the DURAV program will combine the (24) 1-hour activity time blocks into (8) 3-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 program source code, so that in order to change the definitions of the groups (e.g., the age groups in HAPEM7), 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 program source code. The number of groups is specified as ***ngroup*** in the *parameter* files of the DURAV, INDEXPOP, AIRQUAL, and HAPEM programs. The number of day types, ***ntype***, is specified on the *parameter* files of the DURAV and HAPEM programs.

The cluster category for each CHAD record, identified by CHAD identification code, is specified in the *cluster* file. The current version of the DURAV program divides the activity data into 12 person groups, based on demographic (e.g., age in HAPEM7; six categories) and commuting status (yes or no). Activity-pattern data are also separated into three day types: summer weekdays, other weekdays, and weekends. The number of clusters, derived from a statistical cluster analysis procedure, ranges from one to three, depending on the person group and day type (see Appendix A for a detailed discussion on clustering, Appendix B for a detailed discussion of all HAPEM7 input files).

2.1.3. The INDEXPOP Program and the *Population*, *Distance-to-road*, *Commuting-time*, and *Commuting-fraction* Files

The INDEXPOP program performs the three main functions listed below.

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

[8] As explained in Section 2.1.6 (The HAPEM Program, the ME *Factors* and *Mobiles* files, and the Activity *Custer-transition* File), there must be **nmicro** records for each onroad-mobile source category in the mobiles file.

- It creates direct-access files and associated index files of the data in the *distance-to-road*, *commuting-time*, and *commuting-fraction* files, to be used in the COMMUTE and AIRQUAL programs.

The main input file to the INDEXPOP program is the *population* file, which provides the number of people in each demographic group (e.g., age group in HAPEM7; defined in the DURAV program source code) for each census tract in the study area. The data must be sorted according to the state, county, and tract FIPS codes. These data are typically obtained from the census surveys (see Appendix B for more details on the HAPEM7 default input files).

Other input files with census-tract-specific information about the population, such as the *distance-to-road*, *commuting-time*, and *commuting-fraction* files, are also first processed in this program. The *distance-to-road* file provides information on the fraction of each demographic group (e.g., age group in HAPEM7) in each tract that resides within three different distance categories of major roadways, as well as the fraction of the tract area that is within each distance category. The *commuting-time file* provides information on the average commuting time for commuters residing in each tract. The *commuting-fraction* file provides information on the fraction of workers in each group that resides in each tract and that commutes to work (see Appendix B for more details on the HAPEM7 default input files).

2.1.4. The COMMUTE Program and the *Commuting*, *Distance-to-road*, *Commuting-time*, and *Commuting-fraction* Files

The COMMUTE program performs the three main functions listed below.

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

- It creates a sequential index file to facilitate file searching in the HAPEM program.

- It adds the census-tract-specific information from the *distance-to-road*, *commuting-time*, and *commuting-fraction* direct-access files (created in the INDEXPOP program) to the commuting index file.

The *commuting* file is the main input file to the COMMUTE program. It specifies the number of residents of each home census tract that work in that tract and every other tract (i.e., the population associated with each home-tract/work-tract pair), which is typically derived from census data (see Appendix B for more details on the HAPEM7 default input files). While there are hundreds of 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 US and Alaska. To address this issue, the commuting flows were examined as a function of distance. To examine how the decline in commuting flow is affected by distance, researchers plotted the natural log of the natural log of the total flow versus distance. This plot revealed that the ln(ln(total flow)) is nearly linear for distances ranging from 0 to about 100 km. For distances greater than 100 km, the graph exhibits a decreasingly negative slope with distance (i.e., the curve "flattens out"). These findings suggest that people's "commuting behavior" is fairly consistent, on an aggregate basis, to a distance of approximately 100 km. Then, at greater distances, factors other than daily commuting may become increasingly important. Therefore, in constructing the commuting distance distributions for each census tract, commuting distances greater than 120 km are assumed to be atypical for a daily commuter and the COMMUTE program ignores these longer commutes.

2.1.5. The AIRQUAL Program and the *Air Quality* and *Distance-to-road* Files

The AIRQUAL program performs the four main functions listed below.

- It creates a sequential file of air-quality data to be used in the HAPEM program.
- 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 the HAPEM program.
- It adds the tract-specific information from the distance-to-road direct-access file (created in the INDEXPOP program) to the air-quality index files.

The *air quality* file contains the ambient air concentrations that are used by the AIRQUAL program. The file records can 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. The AIRQUAL program requires a separate *air quality* file for each HAP being evaluated. Details about the format of the *air quality* file can be found in Section 3.9 (*Air Quality* File).

The number of outdoor emission source categories is specified in the *parameter* files of the AIRQUAL and HAPEM programs as ***nsource***, and it must match the number in the *factors* file (see Section 2.1.6 [The HAPEM Program, the ME *Factors* and *Mobiles* Files, and the Activity *Cluster-transition* File]). The user specifies the number of time blocks for the analysis in the *parameter* files of the DURAV, AIRQUAL, and HAPEM programs 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 eight 3-hour time blocks per day; if ***hblock*** is set to 24, the AIRQUAL program will create 24 air-quality time blocks with three replicates of each of the eight air-quality values.

2.1.6. The HAPEM Program, the ME *Factors* and *Mobiles* Files, and the Activity *Cluster-transition* File

The HAPEM program performs the six main functions listed below.

- For each demographic group (e.g., each age group in HAPEM7) in each census tract, it randomly selects ***nreplic*** sets of ME factors based on the distribution data provided in the *factors* and *mobiles* files. Each set contains a subset of ME factors randomly selected for each of the time blocks (for the *PEN* and *ADD* factors) or each of the sources (for the *PROX* and *LAG* factors). Each subset contains randomly selected ME factors for each of ***nmicro*** MEs.
- For each demographic group (e.g., each age group in HAPEM7) in each census tract, it randomly selects ***nreplic*** sets of air-quality data from the data sets available for a tract.
- For each demographic group (e.g., each age group in HAPEM7) 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 group/day-type/commuting-status combination. The weights are determined by the relative frequencies of cluster types randomly selected in a one-stage Markov process,[9] based on the cluster transition probabilities provided in the *cluster-transition* file.
- For each activity pattern for a commuting demographic group (e.g., a commuting age group in HAPEM7), 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 ME based on ME factors and outdoor concentrations.
- It combines activity patterns, commuting status, and estimates of ME concentration to calculate ***nreplic*** annual-average exposure concentrations for each demographic group (e.g., each age group in HAPEM7) in each census tract.

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

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

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

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

[9] 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.

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 tract equal to the tract-average concentration contribution). However, when assessing exposure to motor vehicle emissions, for MEs near roadways (e.g., in-vehicle, residences near major roadways) the HAP concentration contribution in the immediate vicinity of the ME is expected to be higher than the average HAP concentration contribution over the census tract (i.e., *PROX* is expected to be greater than 1.0), and this is reflected in the default *factors* and *mobiles* files.

ADD is an additive factor that accounts for emission sources within or near to a ME (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.[7]

A fourth factor, *LAG*, is used to account for the possibility of very slow HAP diffusion and penetration, so that the relevant air-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-concentration value; otherwise, the previous time-block value is used).

The *factors* file includes distributions for each of these factors for each ME/emission-source-category combination, with the exception of *PROX* and *LAG* factors for onroad-mobile-source emissions, which are contained in the *mobiles* file with separate distributions specified for three distance-from-roadway categories. As noted above, the number of MEs in the *factors* and *mobiles* files 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. The *mobiles* file must contain ***nmicro*** records for each onroad-mobile source category specified with ***nmobiles***.

There are three default *factors* files: one each for gaseous HAPs, particulate HAPs, and HAPs which could be either phase depending on conditions. There are four default *mobiles* files for onroad-mobile sources: one each for benzene, 1,3-butadiene, diesel particulate matter (PM), as well as one for non-specific HAPs (see Appendix B for more details on the HAPEM7 default input files).

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

The *cluster-transition* file specifies for each combination of demographic group (e.g., age group in HAPEM7) and day type the number of activity patterns in each of two to three clusters (derived from cluster analysis on the activity pattern data from CHAD), along with the cluster-to-cluster transition probabilities (derived from the transition frequencies for multiple-day activity pattern records from CHAD; see Appendix B for more details on the HAPEM7 default input files). These values are used to create weights for averaging selected activity patterns, one from each cluster, to represent an individual within the demographic group (e.g., age group in HAPEM7) for that day type.

2.1.7. The *Statefip* File

The *statefip* file cross-references the two-character state FIPS codes for each U.S. state (plus Puerto Rico, the U.S. Virgin Islands, and Washington, DC) to its numerical ranking on the list. The numerical rankings range from 1 to 53 in the default file, although the FIPS codes 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*** (used in the *parameter* files of the INDEXPOP, COMMUTE, AIRQUAL, and HAPEM programs to specify the group of states to be included in the analysis according to numerical ranking). For example, setting ***region1*** to 1 and ***region2*** to 53 results in assessment of all the states, districts, and territories in the default *statefip* file (assuming the input files contain all the necessary data). Alternatively, 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 model 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, the INDEXPOP and COMMUTE programs could be run for region range 1 to 53, while the AIRQUAL and HAPEM programs 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 census tracts within a state, the AIRQUAL and HAPEM programs will simply make estimates for those tracts, as long as the state or states are specified within the ***region1*** and ***region2*** range.

2.1.8. Background Concentration

In addition to estimating exposure-concentration contributions for each emission-source category for which data are provided in the *air quality* file, the HAPEM program 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 composed 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 the HAPEM program. 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, the model creates an exposure output file for each state/HAP combination. The names of these files are constructed by the model based on the HAP SAROAD code (specified by ***sarod*** in the *parameter* file) and the state FIPS code (as SAROAD.FIPS.dat).

These output files contain ***nreplic*** records for each combination of census tract and demographic group (e.g., age group in HAPEM7). Each record identifies the census tract, the group, the number of people to which the exposure estimates apply (i.e., 1/***nreplic*** of the population of the group in the 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

HAPEM 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 HAPEM7 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 MEs

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

The model programs should be able to accommodate anywhere from one up to at least 100 MEs. However, large numbers of MEs 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 the DURAV, INDEXPOP, AIRQUAL and HAPEM programs, must be consistently represented in

- the definitions of demographic groups (e.g., age groups in HAPEM7) in the source code for the DURAV program,
- the number of columns in the *population* file,
- the number of columns in the *commuting-fraction* file,

- the number of columns in the *distance-to-road* file, and
- the number of demographic groups (e.g., age groups in HAPEM7) specified in the *cluster* and *cluster-transition* files.

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

The six current age groups are as follows, in years.

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

The number of demographic groups (e.g., age groups in HAPEM7) is unlimited. However, the user is cautioned that for narrowly-defined groups, there might not be enough activity pattern data to calculate a reliable group average or create meaningful activity-pattern clusters. An extreme example of this is where no activity patterns fit a 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) might differ from those in winter (or cold weather), especially for children or other non-workers. Currently, season and day of week are used to determine three day types as

- weekdays in summer (June–August),
- other weekdays, or
- weekends.

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

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

The traditional method for running HAPEM has been to use one-hour time increments (referred to as time blocks). However, the current model 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 24 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 HAPEM7 Run

This section shows how to set-up and conduct a simple HAPEM7 model run. Subsequent sections and chapters provide more detailed explanation about HAPEM7's input and output files and the model's programs.

The example shown in this section is for a hypothetical HAPEM7 analysis of benzene.

The most important consideration for making a HAPEM7 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 11–12 data input files (depending on the HAP and source category), the *parameter* files defining the run, and the five executable files for the five programs that are contained in the model. The programs can 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. As the first three programs in the model sequence (DURAV, INDEXPOP, and COMMUTE) require different inputs from the final two programs (AIRQUAL and HAPEM), it is suggested that two separate *parameter* files be generated for the model sequence of a given simulation or set of simulations. The first *parameter* file (Figure 2-2a) should be used for the first three programs and the second *parameter* file (Figure 2-2b) should be used for the final two programs.

Input/output Files

As seen in Figures 2-2a and 2-2b, the input files (including full pathnames) 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 pathname 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 pathname should be specified for the *product* files, and the full pathname of any existing file (except other model input or output files) must be specified as the *AutoPduct* file in the *parameter* file used with the model. In the future, these files will be part of the input for evaluating indoor sources, but for now the file will not actually be utilized by the HAPEM program. To disable the indoor source algorithms, set keyword ***CAS*** to "99999".

Parameter Settings

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

The last group of information in the *parameter* file shows that there are 18 MEs to be modeled (***nmicro***). This number of MEs must be consistent with the number of ME factors specified in the *factors* and *mobiles* files (i.e., *factors_*_v7.txt* and *factors_OnroadMobile_*_v7.txt*) and the number of duration values specified in the *activity* file (i.e., *activity_CHAD_v7.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 4 (***nsource***). The data in the *air quality* file (for this example the file is *ALL_benzene_HAPEM7-2.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***) (e.g., age groups in HAPEM7) must be consistent with the groups specified in the DURAV source code and in the *population*, *commuting-fraction*, *distance-to-road*, *cluster*, *and cluster-transition* files (i.e., *population_v7.txt*, *commute_fraction_v7.txt*, *proximity_road_v7.txt*, *activity_cluster_v7.txt*, *activity_ClusterTransition_v7.txt*). The number of replicates to be simulated for each group in each tract is 30 (***nreplic***).

In addition, there are five parameter settings that specify the sequence numbers of particular emission source categories and ME types that are subject to special treatment in HAPEM7. In the example, the sequence number for the onroad-mobile source category in the *air quality* file is 3 (***nmobiles***). The sequence numbers of the indoor MEs (including in-vehicle) in the *factors* and *mobiles* files are 1–10 (***nbmicro*** through ***nemicro***). There are two MEs for private commuting, with sequence numbers 7 and 12 (***nvehicles***), and there are four MEs for public-transit commuting, with sequence numbers 8, 10, 11, and 13 (***npublict***). (There may be up to 10 values each for ***nmobiles***, ***nvehicles***, and ***npublict***.)

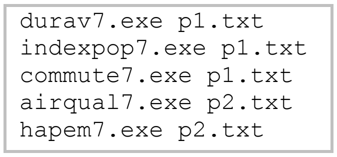
The HAP name (***pollutant***), measurement units (***units***), target year for the analysis (***year***), and the definitions of demographic groups (e.g., age groups in HAPEM7) are listed here by the user so that they can be repeated at the beginning of the final output file for tracking. They have no effect on the exposure results.

2.3.1. Running HAPEM7 as a "Batch" Job

When running HAPEM7 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 model programs sequentially, with all five programs residing in the same directory as the batch file. The batch file is shown in Figure 2-3.

Figure 2-3.
Example “batch” file for running the five model programs



```
durav7.exe p1.txt
indexpop7.exe p1.txt
commute7.exe p1.txt
airqual7.exe p2.txt
hapem7.exe p2.txt
```

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 HAPEM7 executable programs will be run.

2.3.2. Running HAPEM7 Programs Individually

Any of model 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 command line:

```
durav7.exe p1.txt
```

The other model programs are run similarly.

As indicated in Table 2-1, COMMUTE, AIRQUAL and HAPEM all require input files that are generated from running other model programs. Therefore, if any of these programs is run alone, the user must ensure that the required model-generated input files exist and are in the same subdirectory as the original input file from which their filenames were derived (see Table 2-1). For example, running AIRQUAL requires two files with filenames derived from the *population* file and two files with names derived from the *distance-to-road* file. For this example, these files are *population_v7.da*, *population_v7_direct.ind*, *proximity_road_v7.STIDX*, and *proximity_road_v7.dat*, with filenames derived from *population_v7.txt* and *proximity_road_v7.txt*. Therefore, the *parameter* file for running AIRQUAL must specify the full pathname of the *population* and *distance-to-road* files, and the four intermediate files must exist and reside in the subdirectories specified for *population* and *distance-to-road* files, respectively.

If the user wishes to run the model for multiple pollutants using the same regions and settings, it should be noted that DURAV, INDEXPOP, and COMMUTE need only be run in sequence one time. AIRQUAL and HAPEM may then be run for multiple pollutants without rerunning DURAV, INDEXPOP, and COMMUTE. This may be accomplished by either running the programs individually, as directed above, or by creating one batch file for the execution of DURAV, INDEXPOP, and COMMUTE and then a batch file for each successive run of AIRQUAL and HAPEM. If the user chooses to do this, it is suggested that upon completing the execution of DURAV, INDEXPOP, and COMMUTE that the user save the *log* and *mistract* files, if needed, both before and after the execution of AIRQUAL and HAPEM, as each successive run will overwrite these files. The *log* and *mistract* files saved before the execution of AIRQUAL and HAPEM apply to each successive run, as they include the information from the first three programs in the sequence.

3. HAPEM7 Input Files

The model programs use 11–12 user-supplied data input files (depending on the HAP and source category), 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 HAPEM7 are discussed in Chapter 2. The reader is referred to that chapter for an overview of HAPEM7 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 HAPEM7 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 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 filenames used by the model programs are input from the *parameter* file. Thus, the user must check that the *parameter* file refers to the correct filenames before conducting a simulation. Which of the user-supplied files and model-generated files are required for each of the five programs that HAPEM7 contains is discussed in Chapter 2 and presented in Table 2-1.

As explained in Chapter 2, there are default files available for 11 of the 12 user-supplied input files. They are listed below.

Default input files available for HAPEM7

population	(i.e., *population_v7.txt*; national scope)
activity	(i.e., *activity_CHAD_v7.txt*)
cluster	(i.e., *activity_cluster_v7.txt*)
cluster-transition	(i.e., *activity_ClusterTransition_v7.txt*)
commuting	(i.e., *commute_flow_v7.txt*; national scope)
commuting-time	(i.e., *commute_time_v7.txt*; national scope)
commuting-fraction	(i.e., *commute_fraction_v7.txt*; national scope)
distance-to-road	(i.e., *proximity_road_v7.txt*; national scope)
factors	(one each gaseous HAPs, particulate HAPs, and HAPs which could be either phase depending on conditions; i.e., *factors_gas_v7.txt*, *factors_particulate_v7.txt*, and *factors_mixed_v7.txt*, respectively)
mobiles	(one each for benzene, 1,3-butadiene, diesel PM, and non-specific HAPs; i.e., *factors_OnroadMobile_Benzene_v7.txt*, *factors_OnroadMobile_Butadiene_v7.txt*, *factors_OnroadMobile_Diesel_v7.txt*, and *factors_OnroadMobile_Other_v7.txt*, respectively)
statefip	(i.e., *FIPS_StateCrosswalk_v7.DAT*; national scope)

See Appendix B for more details on the HAPEM7 default input files. The user may provide his or her own files as replacements for any or all of these files, using the file formats described in this chapter.

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

3.1. *Parameter* Files

The *parameter* files contain the seven types of information listed below for use in HAPEM7 runs.

- Paths and filenames for the input data files (except the indoor source files, which are not currently used) and the diagnostic-type output files.
- Pathnames for the *final exposure*-output files.
- Identification of the set of states (optionally including the District of Columbia, Puerto Rico, and the U.S. Virgin Islands) for the simulation.
- Identification of the HAP, the units of measurement, the target year of the analysis, and the definitions of demographic groups (e.g., age groups in HAPEM7).
- A spatially-uniform background concentration.
- Internal parameter settings.
- 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 model 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 model 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 model programs is presented in Figures 2-2a and 2-2b in Chapter 2.

The model programs only scan lines containing an equal ("=") sign. The word or words to the left of the equal sign identify which variable is being set and thus should not be changed. The data to the right of the equal sign are the values or settings that the user selects for the model 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 equal sign, which could cause them to be parsed accidentally by the model. 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 equal signs.

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 pathnames for input and output files (except the indoor source files [not currently used] and the *final exposure* output files). The names can be up to 100 characters in length and should not use quotation marks to enclose the filenames. If the full pathnames exceed 100 characters, the user may use abbreviated paths (location of the files relative to the *parameter* file's directory) but must always update these abbreviated paths if the *parameter* file is moved. Some PC systems might require backslashes ("\") in pathnames, rather than the forward slashes ("/") used in UNIX and other 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, district, or territory included in the simulation) for which full pathnames must be specified. The diagnostic output files are *log*, *counter*, and *mistract*.

As explained in Chapter 2, HAPEM7 creates an exposure output file for each state/HAP combination. The names of these files are constructed by the program based on the HAP SAROAD code (specified as the value of the ***sarod*** parameter) and the state FIPS code. Thus, the pathnames, but not the filenames, for these files must be specified in the *parameter* file.

3.1.2. Identifying the Uniform Component of the Background Concentration

In addition to estimating exposure-concentration contributions for each emission source category for which data are provided in the *air quality* file, HAPEM also estimates the exposure-concentration contribution from the background outdoor concentration. This background exposure contribution, of which there are two possible components, is added together with the contributions from the source categories 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, in the same units 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, and/or global transport. The 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 released outside of the modeling domain or a combination of those emissions and the outdoor concentration that would occur in the absence of all anthropogenic emissions. In the latter case, the value of ***backg*** would be set to 0.0, since its constituents would be included in the location-specific background value.

3.1.3. Setting the Internal Parameters

The 12 internal parameter settings (***nmicro***, ***nblock***, ***hblock*****,** ***ntype***, ***ngroup***, ***nsource***, ***nreplic***, ***nmobiles*****,** ***nbmicro***, ***nemicro***, ***nvehicles***, and ***npublict***) are specified by the user in one or more of the *parameter* files and must be consistent with the structure of the other input data files. Each of these parameters is defined in the adjacent text box. Thus, if the user wishes to change the number of MEs, for example, the input files that specify MEs must also be altered in a consistent manner.

Internal Parameters

nmicro number of MEs in the *activity*, *factors*, and *mobiles* files

nblock number of time blocks per day in the *activity* file

hblock number of time blocks per day for the analysis

ntype number of day types in the DURAV source code

ngroup number of demographic groups (e.g., age groups in HAPEM7) in the DURAV source code and the *population* file

nsource number of emission source categories in the *air quality* file

nreplic number of replicates to be simulated for each demographic group (e.g., age group in HAPEM7) in each census tract

nmobiles sequence numbers of up to 10 onroad-mobile emission source categories in the *air quality* file

nbmicro sequence number of the first indoor ME (including in-vehicle) in the *activity*, *factors*, and *mobiles* files

nemicro sequence number of the last indoor ME (including in-vehicle) in the *activity*, *factors*, and *mobiles* files

nvehicles sequence numbers of up to 10 MEs for private commuting in the *activity*, *factors*, and *mobiles* files (e.g., cars, trucks)

npublict sequence numbers of up to 10 MEs for public-transit commuting in the *activity*, *factors*, and *mobiles* files (e.g., buses, trains)

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

- The value of ***hblock*** must be an integral factor of ***nblock*** (the number of time blocks per day in the *activity* file) so that the activity time blocks can be combined if necessary to match to match ***hblock.***
- The value of ***hblock*** must be an integral multiple of the number of time blocks per day in the *air quality* file, so that the air quality values can be replicated if necessary to create ***hblock*** air quality values.

3.2. *Activity* File

The *activity* file, the primary input to the DURAV program, contains information on the time individuals spent in various MEs. This information is not presented as an activity sequence; rather, it is presented in the *activity* file as the total time spent in each ME 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, *activity_CHAD_v7.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=18 NBLOCK=24 HW=2 IN FORTRAN ORDER)
```

Although most of the header record of the *activity* file is not used by the model 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 fixed-width, space-delimited record following the header record consists of one person-day (1,440 minutes) of activity data. The variables in the default *activity* file, extracted from CHAD, are defined in Table 3-1. See Appendix B for more details on the HAPEM7 default input files.

Following the commuting indicator is a series of duration values. The values specify the integral number of minutes (possibly zero) spent in each ME/time block/location combination, where locations are at home or at work. The default *activity* file, *activity_CHAD_v7.txt*, with 18 MEs (listed in Table 1-1), 24 time blocks per day, and two locations has a total of 864 duration values. These values are sequenced so that the 18 ME durations for the first time block in the home location come first, followed by the 18 ME durations for the second time block in the home location, and so on, until all the 432 values for the home location are specified. These are followed by the 432 values for the work location. An example of a record from *activity_CHAD_v7.txt* is presented below.

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

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

Variable	Description
EMPLOYED	1-character string, indicating employment status of respondent, with values as follows: "Y" = Yes "N" = No "X" = missing This information is not used by the current version of the DURAV program.
YEAR, MONTH, DAY	Numeric variables (four-digit year) that identify the date when the activities actually took place. This information is not used by the current version of the DURAV program.
AGE	Integer indicator of the age of the subject (missing = -999.00)
COMMUTE	Integer indicator of whether the respondent is a commuter, with values as follows: 0 = no commuting 1 = commuting
DURATION(MICRO,BLOCK,HW)	Duration of event (minutes). There are 864 of these fields, cycling through each of the 18 MEs; within each ME, cycling through each of the 24 hours of the day; within each ME-hour, cycling through whether the respondent is at work or at home.

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

```
CHADID ZIP DAYTYPE STATE COUNTY GENDER RACE EMPLOYED YEAR MONTH DAY AGE COMMUTE
DURATION(MICRO,BLOCK,HW) (NMICRO=14 NBLOCK=24 HW=2 IN FORTRAN ORDER)
BAL97001A 21204  2  24 000 1 1 N 1997   1 21   77.00 0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  30   0   0   0   0  30   0   0   0   0   0   0   0   0
45   0   0   0   0  15   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0  60   0   0   0   0   0   0
0   0   0   0   0   0   0  60   0   0   0   0   0   0   0   0   0   0   0   0   0
60   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

3.2.2. Replacing or Modifying the Default File

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

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

The variables used by the DURAV program for classifying activity records (i.e., daytype, demographic [e.g., age in HAPEM7], 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 (e.g., age groups in HAPEM7; presented in Section 2.2.2 [Changing the Number and/or Definitions of the Demographic Groups]) are part of the DURAV source code, so that in order to change the 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 groups and day types is unlimited. However, the user is cautioned that for narrowly-defined groups and day types, there might not be enough activity-pattern data to calculate a reliable group average or create meaningful activity-pattern clusters. An extreme example of this is where no activity patterns fit a group's definition, resulting in incorrect exposure calculations (i.e., exposure concentrations equal to zero) for that group.

The number, definition, and order of MEs must be the same in both the *activity* file and the *factors* and *mobiles* files (see Section 3.10 [ME *Factors* and *Mobiles* Files]). 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 filename, 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 model input files, the user can add comments or other information after the last data record in the file. To prevent the program from 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 group (e.g., age group in HAPEM7), day type, cluster type of each complete (i.e., with 1,440 minutes per day) CHAD record in *activity* file. The file is used in DURAV to group CHAD records according to cluster. See Appendix A for more details on the HAPEM7 cluster file, and Appendix B for more details on the HAPEM7 default input files.

3.3.1. Variables and Format of the Default File

The first line of the fixed-width, space-delimited *cluster* file (*activity_cluster_v7.txt*) is a text header that indicates the order of the variables in each record. The header in the default *cluster* file is as follows.

Header of default *cluster* file

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

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

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

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 1,440 minutes, a demographic (e.g., age in HAPEM7) specification, and a day-type designation of 1–3, and a commuting-status specification). Any record in the *activity* file without a corresponding record in the *cluster* file will not be used.

- The records should be sorted by demographic group (e.g., age group in HAPEM7), day type, commuting status, and cluster.

Extract from default *cluster* file

```
CHADID Demographic DayType "Comtype(1=non-commute," CLUSTER Ncluster
CAC01166A      1      1      1      1     1
CAC01251A      1      1      1      1     1
CAC01489A      1      1      1      1     1
CAC01562A      1      1      1      1     1
CAC01568A      1      1      1      1     1
CAC01809A      1      1      1      1     1
CAC01830A      1      1      1      1     1
CAC01982A      1      1      1      1     1
CAC02036A      1      1      1      1     1
CAC02132A      1      1      1      1     1
```

3.4. *Population* File

The *population* file, the primary input to the INDEXPOP program, provides the number of people in each demographic group (e.g., age group in HAPEM7) residing in each census tract of the study area. The data must be sorted according to state FIPS, county FIPS, and tract FIPS. The data are typically derived from the U.S. Census data. The group definitions are defined in the DURAV source code, and presented in Section 2.2.2 (Changing the Number and/or Definitions of the Demographic Groups). See Appendix B for more details on the HAPEM7 default input files.

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 and second header records of the default *population* file are as follows.

Header of default *population* file

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

Although the header of the *population* file is not used by the model programs, it provides documentation to inform the user of the meaning of the data fields. Each fixed-width, space-delimited data record following the header consists of a census-tract identifier and a population value for each of the indicated demographic groups (e.g., age groups in HAPEM7) in that tract. The definitions of the data fields in the default *population* file are presented in Table 3-2.

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

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

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

Extract from default *population* file

```
TRACT           B_00     B_02     B_05     B_16     B_18     B_65
                COM      COM      COM      COM      COM      COM
01001020100     40       78       303      86       1184     221
01001020200     45       82       391      88       1350     214
01001020300     92       151      537      114      2040     439
01001020400     88       146      634      147      2467     904
01001020500     274      455      2097     336      6478     1126
01001020600     113      158      603      134      2249     411
01001020700     84       108      411      83       1845     360
01001020801     70       113      521      111      1925     341
01001020802     288      439      1808     374      6466     1060
01001020900     147      227      952      185      3534     630
01001021000     70       106      470      104      1797     347
```

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 (e.g., age groups in HAPEM7) 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 [Changing the Number and/or Definitions of the Demographic Groups]) are part of the DURAV source code, so that in order to change the definitions of demographic groups (e.g., age groups in HAPEM7) 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 filename, but with different filename extensions. Thus, the user should NOT name a *population* file with any of the following filename extensions: *.da*, *.county_tract_pop_range*, and *.state_county_pop_range*. There is also an intermediate file with the characters *_direct.ind* attached to the *population* file root name.

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

3.5. *Commuting-time* File

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

The *commuting-time* file has no header records, only data records. Each tab-delimited data record contains the following five variables, derived from U.S. Census data for the default file.

Variables in the *commuting-time* file

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

The default *commuting-time* file is sorted by tract ID, smallest to largest in numerical order. Several example data records from the default *commuting-time* file are presented below. See Appendix B for more details on the HAPEM7 default input files.

Extract from the default *commuting-time* file

```
01001020100  0.0472      0.9528      14.4946     35.9467
01001020200  0.0000      1.0000      0.0000      51.6934
01001020300  0.0092      0.9908      14.4946     35.9467
01001020400  0.0000      1.0000      0.0000      40.9754
01001020500  0.0000      1.0000      0.0000      38.9659
01001020600  0.0000      1.0000      0.0000      47.1820
01001020700  0.0000      1.0000      0.0000      50.3055
01001020801  0.0000      1.0000      0.0000      35.9467
01001020802  0.0000      1.0000      0.0000      54.3099
01001020900  0.0000      1.0000      0.0000      67.2094
```

3.6. *Commuting-fraction* File

The *commuting-fraction* file provides data for each tract on the proportion of workers in each demographic group (e.g., age group in HAPEM7) that commutes to work. This information is used by the HAPEM program to determine for each replicate in each group whether they commute to work, and therefore, which set of activity patterns should be sampled to represent that replicate. The data in the default *commuting-fraction* file are derived from U.S. Census data.

The *commuting-fraction* file has no header records, only data records. Each tab-delimited data record contains 13 variables, as follows.

Tract ID	(11-character string: state FIPS, county FIPS, and tract FIPS)
Proportion of workers in demographic-group 1 (e.g., age-group 1 in HAPEM7) that does not commute to work	(decimal number)
Proportion of workers in demographic-group 1 (e.g., age-group 1 in HAPEM7) that commutes to work	(decimal number)
Repeat the latter two above for groups 2–6	

The default *commuting-fraction* file is sorted by tract ID, smallest to largest in numerical order. Several example data records from the default *commuting-fraction* file are presented below. See Appendix B for more details on the HAPEM7 default input files.

Extract from the default *commuting-fraction* file
(in “wrapped” view)

```
01001020100   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5500 0.4500 0.2742 0.7258
       0.7883 0.2117
01001020200   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.3671 0.6329
       0.8627 0.1373
01001020300   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5126 0.4874 0.3107 0.6893
       0.8809 0.1191
01001020400   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6906 0.3094 0.2840 0.7160
       0.8335 0.1665
01001020500   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6108 0.3892 0.2375 0.7625
       0.8590 0.1410
01001020600   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6727 0.3273 0.3108 0.6892
       0.8679 0.1321
01001020700   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.8545 0.1455 0.2739 0.7261
       0.9306 0.0694
01001020801   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.7731 0.2269 0.2645 0.7355
       0.8790 0.1210
01001020802   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5788 0.4212 0.3168 0.6832
       0.7830 0.2170
01001020900   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.9576 0.0424 0.4008 0.5992
       0.9027 0.0973
```

3.7. *Distance-to-road* File

The *distance-to-road* file provides data for each tract on the proportion of tract area and the proportion of each demographic group (e.g., age group in HAPEM7) that resides within three distance categories from a major roadway: 0–75 meters, 75–200 meters, and greater than 200 meters. This information is used by the HAPEM program to determine, for each replicate for each ME, the distance from a major roadway and, therefore, which *PROX* factor distributions in the *mobiles* file, described below, to sample from for the onroad-mobile source categories.

The *distance-to-road* file has no header records, only data records. Each tab-delimited data record contains the 22 variables listed below, derived in the default file using the U.S. Census data as well as census data processed by a third party.

Variables in the *distance-to-road* file

Tract ID	(11-character string: state FIPS, county FIPS, and tract FIPS)
Fractions of tract area within each of three distance categories from a major roadway: 0–75 meters, 75–200 meters, greater than 200 meters	(4-decimal numbers)
Fractions of demographic-group 1 (e.g., age-group 1 in HAPEM7) that reside within each of three distance categories from a major roadway: 0–75 meters, 75–200 meters, greater than 200 meters	(4-decimal numbers)
Repeat the latter one for groups 2–6	

The default *distance-to-road* file is sorted by tract ID, smallest to largest in numerical order. Several example data records from the default *distance-to-road* file are presented below. See Appendix B for more details on the HAPEM7 default input files.

Extract from the default *distance-to-road* file (in "wrapped" view)

```
01001020100   0.0643  0.0970  0.8387  0.0582  0.0882  0.8535  0.0797  0.1179
              0.8023  0.0598  0.0913  0.8489  0.0488  0.0794  0.8717  0.0634
              0.0977  0.8389  0.0767  0.1022  0.8211
01001020200   0.0326  0.0622  0.9052  0.0031  0.0158  0.9811  0.0284  0.0627
              0.9089  0.0273  0.0512  0.9215  0.0202  0.0435  0.9363  0.0287
              0.0655  0.9058  0.0803  0.0789  0.8408
01001020300   0.0586  0.1039  0.8375  0.0484  0.0824  0.8692  0.0528  0.0818
              0.8654  0.0551  0.0910  0.8539  0.0567  0.0925  0.8508  0.0586
              0.1029  0.8384  0.0671  0.1395  0.7934
01001020400   0.1293  0.2003  0.6704  0.1742  0.2039  0.6219  0.1382  0.2498
              0.6121  0.1270  0.2046  0.6684  0.1247  0.1873  0.6880  0.1331
              0.1992  0.6678  0.1156  0.1942  0.6902
01001020500   0.0214  0.0426  0.9361  0.0189  0.0380  0.9431  0.0215  0.0407
              0.9378  0.0196  0.0399  0.9405  0.0185  0.0386  0.9429  0.0213
              0.0430  0.9357  0.0265  0.0484  0.9252
01001020600   0.1399  0.2283  0.6319  0.1447  0.2417  0.6136  0.1346  0.2135
              0.6518  0.1386  0.2257  0.6358  0.1244  0.2010  0.6746  0.1409
              0.2319  0.6272  0.1419  0.2231  0.6350
01001020700   0.0898  0.1699  0.7404  0.0833  0.1573  0.7594  0.0796  0.1509
              0.7694  0.0825  0.1540  0.7634  0.0877  0.1669  0.7454  0.0945
              0.1810  0.7245  0.0788  0.1400  0.7812
01001020801   0.0609  0.0826  0.8565  0.0586  0.0823  0.8591  0.0661  0.0774
              0.8565  0.0499  0.0700  0.8801  0.0559  0.0762  0.8679  0.0610
              0.0824  0.8566  0.0775  0.1068  0.8156
01001020802   0.0597  0.0836  0.8567  0.0586  0.0909  0.8505  0.0681  0.0910
              0.8409  0.0605  0.0867  0.8528  0.0568  0.0788  0.8644  0.0598
              0.0831  0.8571  0.0554  0.0782  0.8665
01001020900   0.1019  0.1052  0.7929  0.1195  0.1237  0.7567  0.0985  0.1091
              0.7924  0.1032  0.1067  0.7901  0.0957  0.1121  0.7922  0.1019
              0.1041  0.7941  0.0991  0.1014  0.7995
```

3.8. *Commuting* 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 U.S. Census data identifying the tract of work and tract of residence for individuals in all 50 states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands. Only those home-tract/work-tract pairs with non-zero flows are included in the default *commuting* file.

The *commuting* file has data records with no header records. Each fixed-width, space-delimited data record contains five variables (the first being empty), as follows.

Variables in the *commuting* file

Leading space in file	
Home tract ID	(11-character string: state FIPS, county FIPS, and tract FIPS)
Work tract ID	(11-character string: state FIPS, county FIPS, and tract FIPS)
Distance apart in kilometers	(2-decimal number)
Fraction of workers in the commuting flow	(8-decimal number; 8 decimals; sums to 1 across all instances of a home tract)

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. See Appendix B for more details on the HAPEM7 default input files.

The *commuting* file is read by the COMMUTE program, which creates several intermediate output files with the same path and root filename, 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 model 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 01051031100 13.13   0.01156069
 01001020100 01101000900 15.91   0.10982659
 01001020100 01101005902 23.67   0.01156069
 01001020100 01101005901 33.87   0.01156069
 01001020100 01101005406 33.9    0.01156069
 01001020100 01101003100 25.88   0.01156069
 01001020100 01101002900 28.94   0.01156069
 01001020100 01101001600 23.35   0.01156069
 01001020100 01101000700 21.28   0.01156069
 01001020100 01101005409 29.97   0.01734104
```

3.9. *Air Quality* File

The *air quality* file contains the ambient air concentrations that are used by the AIRQUAL program. AIRQUAL requires a separate *air quality* file for each HAP 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 an AERMOD-derived *air quality* file used for the recent NATA analysis, which indicates the order of the variables in each of the data records, is as follows.

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

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

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.

Variables in the *air quality* file

Leading spaces in file	
Home tract ID	(11-character string: state FIPS, county FIPS, and tract FIPS)
Space-delimited air concentrations for spatially-variable (but temporally constant) background concentration	(decimal number, optionally in exponential format)
Space-delimited air concentrations for each combination of emission source category and time block	(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)

```
   01001  020100 0.00E+00 2.00E-03 2.00E-03 1.34E-03 6.44E-04 5.30E-04 1.16E-03 1.89E-
03 1.93E-03 1.52E-02 1.75E-02 1.85E-02 1.23E-02 9.36E-03 2.79E-02 4.21E-02 2.34E-02
2.68E-03 4.29E-03 1.10E-02 6.29E-03 3.84E-03 9.17E-03 9.80E-03 5.06E-03 1.56E-02
1.76E-02 1.87E-02 1.14E-02 9.25E-03 2.17E-02 2.71E-02 1.98E-02
```

The *air quality* file is read by the AIRQUAL program, which creates several intermediate output files with the same path and root filename, 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 model input files, the user can add comments or other information after the last data record in the file. To prevent the program reading these comments as data, a blank line must be inserted after the last data record and before any comments.

3.10. ME *Factors* and *Mobiles* Files

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

$$ME(m,c,t,s,d) = PROX(m,s,d) \times PEN(m,t) \times AMB(c,t_{LAG(m)},s)$$

$$ME(m,t,i) = ADD(m,t)$$

$$ME(m,c,t,b) = PROX(m.,s) \times PEN(m,t) \times [bckgd\_u + bckgd\_v(c)]$$

$$ME(m,c,t.d) = \sum_{s} ME(m,c,t,s,d) + ME(m,t,i) + ME(m,c,t,b)$$

where:

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

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

PEN(m,t): penetration factor for ME *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 ME *m* at time *t* due to indoor sources,

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

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

bckgd_u: uniform component of ambient background concentration,

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

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

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

$$PEN = \frac{\text{indoor or in - vehicle ME concentration}}{\text{outdoor concentration in immediate vicinity of indoor or in - vehicle ME}}$$

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,

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

$$PROX_O = \frac{\text{outdoor ME concentration}}{\text{air quality file concentration}}$$

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

ADD is an additive factor that accounts for emission sources within or near to a ME (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.[10]

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

The *factors* and *mobiles* files have no header records. The *factors* file contains a set of records (one for each ME) for each outdoor-source category being modeled (the number identified as **nsource**), in the same order as the source categories are specified in the *air quality* file. The *mobiles* file contains a set of records (one for each ME) for the onroad-mobile source category identified with ***nmobiles*** and for each distance-from-road category.[11] The MEs must be the same number, definition, and order as the MEs in the *activity* file. The files are read in free format, once for each ME, with fields as specified in Tables 3-3a and 3-3b. All values are decimal numbers.

[10] A database 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 database. Therefore, it is currently recommended that indoor sources be omitted from HAPEM7 applications until the database and algorithms have been tested and reviewed. To disable the indoor-source algorithms, set keyword **CAS** to 99999.

[11] Note that a *PROX* factor distribution is specified in the *factors* file for the onroad-mobile source category as a place-holder and such values should be set to 1. The *PROX*-factor distributions in the *mobiles* file are then multiplied by the distributions from the *factors* file.

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

ME Factor	Field Num.	Parameter	
(N/A)	1	Number of ME (1–18)	
PEN	2	Distribution Type 1 - Normal 2 - Log-normal 3 - Uniform 4 - Triangular 5 - Data Set	
	3	Distribution Type Normal Lognormal Uniform Triangular Dataset	Parameter Mean Mean Minimum Minimum Number of data points
	4	Distribution Type Normal Lognormal Uniform Triangular Dataset	Parameter Standard deviation Standard deviation Maximum Maximum First data point in the set
	5	Distribution Type Normal Lognormal Triangular Dataset	Parameter 0 (always) 0 (always) Mode Second data point in the set
	6	Distribution Type Normal Lognormal Dataset	Parameter Lower bound (optional) Lower bound (optional) Third data point in the set
	7	Distribution Type Normal Lognormal Dataset	Parameter Upper bound (optional) Upper bound (optional) Fourth data point in the set
	8	Distribution Type Dataset	Parameter Fifth data point in the set
	9	Distribution Type Dataset	Parameter Sixth data point in the set
	10	Distribution Type Dataset	Parameter Seventh data point in the set
	11	Distribution Type Dataset	Parameter Eighth data point in the set
	12	Distribution Type Dataset	Parameter Ninth data point in the set
	13	Distribution Type Dataset	Parameter Tenth data point in the set
ADD	14-25	Repeat fields 2-13 for additive factor	
PROX			
Source 1	26-37	Repeat fields 2-13 for proximity factor	
Source 2	39-50	Repeat fields 2-13 for proximity factor	
Source 3	52-63	Repeat fields 2-13 for proximity factor	
Source 4	65-76	Repeat fields 2-13 for proximity factor	

ME Factor	Field Num.	Parameter
LAG		
Source 1	38	hours
Source 2	51	hours
Source 3	64	hours
Source 4	77	hours

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

ME Factor	Field Num.	Parameter	
(N/A)	1	Number of ME (1–18)	
PROX for Onroad-mobile Source Category: **Distance-from-source Category 1**	2	Distribution Type 1 - Normal 2 - Log-normal 3 - Uniform 4 - Triangular 5 - Data Set	
	3	Distribution Type Normal Lognormal Uniform Triangular Dataset	Parameter Mean Mean Minimum Minimum Number of data points
	4	Distribution Type Normal Lognormal Uniform Triangular Dataset	Parameter Standard deviation Standard deviation Maximum Maximum First data point in the set
	5	Distribution Type Normal Lognormal Triangular Dataset	Parameter 0 (always) 0 (always) Mode Second data point in the set
	6	Distribution Type Normal Lognormal Dataset	Parameter Lower bound (optional) Lower bound (optional) Third data point in the set
	7	Distribution Type Normal Lognormal Dataset	Parameter Upper bound (optional) Upper bound (optional) Fourth data point in the set
	8	Distribution Type Dataset	Parameter Fifth data point in the set
	9	Distribution Type Dataset	Parameter Sixth data point in the set
	10	Distribution Type Dataset	Parameter Seventh data point in the set
	11	Distribution Type Dataset	Parameter Eighth data point in the set
	12	Distribution Type Dataset	Parameter Ninth data point in the set
	13	Distribution Type Dataset	Parameter Tenth data point in the set

ME Factor	Field Num.	Parameter
LAG for Onroad-mobile Source Category: **Distance-from-source Category 1**	14	Hours
Distance-from-source Category 2	15-27	Repeat fields 2-14
Distance-from-source Category 3	28-40	Repeat fields 2-14

The fields in the *factors* file include *PROX* distributions (one per ME and source category), *PEN* distributions (one per ME), *ADD* distributions (one per ME), and *LAG* factors (one per ME; *LAG* factors are either 0 or 1).

The fields in the *mobiles* file include distributions of *PROX* and *LAG* factors for onroad-mobile source category identified with ***nmobiles***. Distributions of *PROX* factors in the *mobiles* files are stratified for each of three distance-from-source categories: 0–75 meters, 75–200 meters, and beyond 200 meters, and this information is combined with the data in the *distance-to-road* file in the HAPEM program to determine from which probability distribution the *PROX* factor should be selected for a given tract/ME combination (see Section 5.2.5 [HAPEM] for more details). The distributions in the *mobiles* file override those in the *factors* file override those in the *factors* file for the onroad-mobile source category identified with ***nmobiles***.

Distributions can take any of 5 different forms: normal, lognormal, uniform, triangular, or data set. The data set is composed 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 listed below.

Distribution types used in the *factors* and *mobiles* files

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

For HAPEM7, default *factors* files are provided for each of three phases of HAPs: gaseous, particulate, and HAPs that might be either phase depending on various conditions. Default *mobiles* files are provided for benzene 1,3-butadiene, diesel PM, and non-specific HAPs (formatted for a single onroad-mobile source category). As noted above, because a new approach to evaluating indoor sources is in development, the default *ADD* factors are uniformly set to zero. Due to lack of data, default *LAG* factors are uniformly set to zero. Excerpts from the default *factors* and *mobiles* files for gaseous HAPs and non-specific HAPs, respectively, are presented below. See Appendix B for more details on the HAPEM7 default input files.

As with other model 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 *factors* file (in “wrapped” view)

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

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

```
1     2    2.477  2.0477  0        1      8.0532 0       0       0       0      0       0       0
      2    1.61131.9292  0        1      4.7492 0       0       0       0      0       0       0
      5    1      1       0        0      0      0       0       0       0      0       0       0
2     2    2.477  2.0477  0        1      8.0532 0       0       0       0      0       0       0
      2    1.61131.9292  0        1      4.7492 0       0       0       0      0       0       0
      5    1      1       0        0      0      0       0       0       0      0       0       0
3     2    2.477  2.0477  0        1      8.0532 0       0       0       0      0       0       0
      2    1.61131.9292  0        1      4.7492 0       0       0       0      0       0       0
      5    1      1       0        0      0      0       0       0       0      0       0       0
```

3.11. *Cluster-transition* File

The *cluster-transition* file specifies, for each combination of demographic group (e.g., age group in HAPEM7), day type and commuting status, the number of activity patterns in each of 1–3 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 group for that day type.

The *cluster-transition* file begins with a text header record, followed by one data record for each combination of day type and demographic group (e.g., age group in HAPEM7). The header record indicates the order of the variables in each of the data records. Although the header record of the *cluster-transition* file is not used by the model programs, it provides documentation to inform the user of the meaning of the data fields. The header record of the default *cluster-transition* file is as follows.

Header record from the default *cluster-transition* file

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

The *cluster-transition* file is read in free format with the variables defined in Table 3-4 for each combination of day type and demographic group (e.g., age group in HAPEM7).

Table 3-4.
Variables in the *cluster-transition* file

Variable	Description
Demographic	demographic group (e.g., age group in HAPEM7)
DayType	day type
Comtype (1=non-commute 2=commuting)	commuting status of subject
Ncluster	number of clusters for the group/day type (1–3)
cluster1	cumulative fraction of group/day type in cluster #1
cluster2	cumulative fraction of group/day type in clusters #1–2
cluster3	cumulative fraction of group/day type in clusters #1–3
prob11	cumulative transition probability from cluster #1 to #1
prob12	cumulative transition probability from cluster #1 to clusters #1–2
prob13	cumulative transition probability from cluster #1 to clusters #1–3
prob21	cumulative transition probability from cluster #2 to #1
prob22	cumulative transition probability from cluster #2 to clusters #1–2
prob23	cumulative transition probability from cluster #2 to clusters #1–3
prob31	cumulative transition probability from cluster #3 to #1
prob32	cumulative transition probability from cluster #3 to clusters #1–2
prob33	cumulative transition probability from cluster #3 to clusters #1–3

The default *cluster-transition* file is presented below. See Appendix A for more details on the HAPEM7 cluster file, and Appendix B for more details on the HAPEM7 default input files.

Default *cluster-transition* file (in “wrapped” view)

```
Demographic DayType "Comtype(1=non-commute,2=commuting)" Ncluster cluster1 cluster2 cluster3 prob11 prob12 prob13 prob21
prob22 prob23 prob31 prob32 prob33
1      1      1      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
1      1      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
1      2      1      2      0.79132 1.00000 0.00000 0.96552 1.00000 0.00000 0.50000 1.00000 0.00000 0.00000 0.00000 0.00000
1      2      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
1      3      1      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
1      3      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
2      1      1      3      0.56255 0.88783 1.00000 0.60392 0.91373 1.00000 0.37383 0.84579 1.00000 0.32000 0.69333 1.00000
2      1      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
2      2      1      3      0.48133 0.69467 1.00000 0.66667 0.82456 1.00000 0.07692 0.71795 1.00000 0.19608 0.39216 1.00000
2      2      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
2      3      1      3      0.51881 0.77713 1.00000 0.60417 0.77083 1.00000 0.41176 0.70588 1.00000 0.53333 0.80000 1.00000
2      3      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
3      1      1      3      0.22707 0.45280 1.00000 0.44091 0.58182 1.00000 0.13725 0.46667 1.00000 0.14408 0.31098 1.00000
3      1      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
3      2      1      3      0.18265 0.82500 1.00000 0.51304 0.96522 1.00000 0.14741 0.82470 1.00000 0.17647 0.76471 1.00000
3      2      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
3      3      1      3      0.73402 0.85778 1.00000 0.79412 0.94118 1.00000 0.51852 0.96296 1.00000 0.61905 0.76190 1.00000
3      3      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
4      1      1      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
4      1      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
4      2      1      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
4      2      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
4      3      1      2      0.74422 1.00000 0.00000 0.90909 1.00000 0.00000 0.20000 1.00000 0.00000 0.00000 0.00000 0.00000
4      3      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
5      1      1      3      0.36570 0.93345 1.00000 0.56215 0.98145 1.00000 0.27810 0.94902 1.00000 0.20513 0.76923 1.00000
5      1      2      3      0.52284 0.66370 1.00000 0.79745 0.83668 1.00000 0.14331 0.76752 1.00000 0.25067 0.35600 1.00000
5      2      1      3      0.38219 0.85709 1.00000 0.55385 0.96154 1.00000 0.31098 0.94512 1.00000 0.17391 0.52174 1.00000
5      2      2      3      0.55211 0.76894 1.00000 0.71816 0.79958 1.00000 0.22297 0.92568 1.00000 0.54331 0.60630 1.00000
5      3      1      3      0.76220 0.89489 1.00000 0.85837 0.95279 1.00000 0.76190 0.97619 1.00000 0.55556 0.66667 1.00000
5      3      2      3      0.46041 0.79277 1.00000 0.69231 0.84615 1.00000 0.16000 0.84000 1.00000 0.16667 0.41667 1.00000
6      1      1      3      0.44116 0.86413 1.00000 0.67627 0.94900 1.00000 0.27088 0.84114 1.00000 0.18243 0.66892 1.00000
6      1      2      3      0.31198 0.76240 1.00000 0.71667 0.86667 1.00000 0.10569 0.88618 1.00000 0.11594 0.39130 1.00000
6      2      1      3      0.39745 0.89898 1.00000 0.58377 0.94241 1.00000 0.30841 0.96262 1.00000 0.27778 0.44444 1.00000
6      2      2      3      0.67063 0.92262 1.00000 0.93789 0.98758 1.00000 0.52174 0.91304 1.00000 0.36364 0.45455 1.00000
6      3      1      3      0.57117 0.93835 1.00000 0.71141 0.97315 1.00000 0.43878 0.94898 1.00000 0.23529 0.58824 1.00000
6      3      2      1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
```

3.12. *Statefip* File

The *statefip* file cross-references the two-character FIPS code for each U.S. state (and District of Columbia, Puerto Rico, and the U.S. Virgin Islands, totaling 53 areas) to its numerical ranking on the list. The format of each record is as follows.

Variables in the *statefip* file

Numerical Rank	(integer)
State (or district or territory) FIPS code	(2-character string)

As discussed in Section 2.1.7 (The *Statefip* File), the *statefip* file is used in conjunction with ***region1*** and ***region2*** specified in the *parameter* files of the INDEXPOP, COMMUTE, AIRQUAL, and HAPEM programs to specify the areas 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	U.S. Virgin Islands

This page intentionally left blank.

4. HAPEM7 Output Files

HAPEM7 creates three diagnostic output files and a set of *final exposure* output files. The diagnostic files record error messages and information about the parameters of the simulations. The names for these three files are specified by the user in the *parameter* files of each model program. The *final exposure* output files contain all of the exposure estimates from a model run. The pathnames 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 model analysis. Three of the model programs (INDEXPOP, COMMUTE, and HAPEM) will append records onto an existing *log* file, as specified their *parameter* file, without overwriting previous records. The DURAV and AIRQUAL programs will overwrite any records on an existing *log* file. Therefore, if a single *log* filename is used to run all the model programs, a running record will be written for the DURAV, INDEXPOP, and COMMUTE programs, but then the AIRQUAL program will erase those records and begin a new record, and the HAPEM program will add to it. To maintain a complete *log* file record of a HAPEM7 simulation, the two alternatives below can be used.

- If a single *parameter* file is for a complete simulation, so that the *log* filename is the same for all five programs, manually rename the *log* file created by the first three 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 model 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 start time consists of three numbers—the current time, the size of the time increment equivalent to one second, and 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* record: program and start time

```
DURAV Start time=   34862630 1000 86399999
```

The last two records written to the *log* by each model 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 types of records is presented below.

Example *log* records: program, stop time, and run time

```
DURAV End time =  34880980
DURAV Job time =  18.3500004
```

If an error occurs that HAPEM7 considers to be fatal, a diagnostic message will be written to the *log* file and the program will stop. 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* record: error message

```
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 model programs, each program writes some specialized information to the *log* file. The DURAV program writes the names of the input *activity* file and the intermediate file used to sort activity patterns for each group. An example of these types of records is presented below.

Example *log* records: input and intermediate files

```
Data read from file=input/activity pattern/activity_CHAD_v7.txt
 direct access in file (deleted) =
 input/activity pattern/activity_CHAD_v7.draft
```

The DURAV program also records the number of records (person-days) extracted from the *activity* file, and it produces a table of frequency counts for each combination of demographic group (e.g., age group in HAPEM7), commute status, 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* records: file matrix

```
Total number of person-days processed=       45628
 Frequency table for person-days:
 By demographic group (rows) & day type (cols)
   178   599   596
  1159  1500  1382
  2246  6143  5414
    64   740   692
  2945  2470  3444
  1759  1960  1103
     0     5     2
     6    11     8
    36   106    83
    40   137   131
  4160  4423  1023
   484   504    75
```

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

Example *log* record: output file

```
Data written to file=
 input/activity pattern/activity_CHAD_v7.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 model programs, the INDEXPOP program writes two 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* records: opened files and tract counts

```
Finished opening files
 total number of tracts is       74034
```

4.1.3. COMMUTE Output to the *Log* File

The COMMUTE program writes 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 model programs, the AIRQUAL program writes several other records to the *log* file. First, a summary of the *air quality* file is written to document the number of census tracts and distinct counties found in the file. 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 *population* file is recorded in the line containing the phrase “unpaired air tracts”. This is followed by the list (if any) of unpaired tracts. Then, the tracts in the *population* file are compared to the tracts in the *air quality* file—the number of tracts in the *population* file but not in the *air quality* file is reported, along with the number of matching tracts as well as the number of *population* tracts with multiple *air quality* tracts. Next, similar statistics are given comparing counties in the *population* file to counties in the *air quality* file. Any tract in the *population* file but not the *air quality* file will not be modeled; any tract in the *air quality* file but not in the *population* file will not be modeled. An example of the log output produced by the AIRQUAL program is presented below.

Example *log* records: AIRQUAL statistics

```
 # air tracts =       74859      # of air records =       74859
 # counties on air file =        3224
 There were         2  unpaired air tracts.
 01003274048
 01003277049
Overall, there were:
         202  unpaired census tracts.
       73832  census tracts with a matching air tract.
           0  census tracts with 2 or more air tracts.
 Within the counties on the air file, ther were:
         202  unpaired census tracts.
       73832  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 model programs, the HAPEM program writes two other records to the *log* file. It reports the time when dynamic array allocation is complete 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 the HAPEM program is presented below.

Example *log* record: HAPEM array allocation and tracts

```
HAPEM Allocation = 35921930
There were        74034  tracts in the study area.
```

4.2. *Counter* File

A second diagnostic file created by HAPEM7 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 troubleshooting and keeping track of which files were used in the simulation.

It is important to use same *counter* file for all the model programs in a simulation—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 model 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 (Model Structure; though running the COMMUTE program is optional). For example, the INDEXPOP program reads records generated by the DURAV program, and then it begins its own recording. If the INDEXPOP program is run a second time using the same *counter* file, the second run will overwrite the previous records generated by the INDEXPOP program.

The specific information recorded in the *counter* file is provided in Table 4-1, using the names of the HAPEM7 default input files. An example counter file is also shown below.

Table 4-1.
Variables in the *counter* file

HAPEM Program	Record Number	Description
DURAV	1	-number of data records (person-days) in the *activity* file (*activity_CHAD_v7.txt*) -number of *activity*-file data records (person-days) with 1,440 total minutes
INDEXPOP	2	-number of data records (tracts) in the *population* file (*population_v7.txt*) -number of counties in the *population* file (*population_v7.txt*)
COMMUTE	3	-number of data records in the population index file (*population_v7_direct.ind*) -number of data records (home-tract/work-tract pairs) in the *commuting* file (*commute_flow_v7.txt*)
	4	-number of data records in the work-tract file (*commute_flow_v7.da*) -number of records in the commuting index file *(commute_flow_v7.ind*)

HAPEM Program	Record Number	Description
AIRQUAL	5	-number of matching tracts in the *air quality* (e.g., *benzene.txt*) and population index (*population_v7.ind*) files -number of counties with matching tracts in the *air quality* (e.g., *benzene.txt*) and population index (*population_v7.ind*) files
	6	-number of tracts in the *air quality* file (e.g., *benzene.txt*) -number of data records in the *air quality* file (e.g., *benzene*.txt)

Example *counter* file

```
      45628       45628
      74034        3224
      74034     4156458
    4017682       74034
      73832        3224
      74859       74859
```

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

- The number of records in the *population* file (*population_v7.txt*), the population index file (*population_v7.ind*), and the commuting index file *(commute_flow_v7.ind*) should all be the same.
- The number of records in the work-tract file (*commute_flow_v7.da*) may be larger or smaller than the number of records in the *commuting* file (*commute_flow_v7.txt*). It may be larger because the COMMUTE program will create a "commuting" flow for a tract that is in the *population* file but is not a home tract in the *commuting* file (using the *population* tract as both the home and work tract). It may also be smaller if the study area in the *population* file is smaller than the study area in the *commuting* file (which is all U.S. states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands in the default *commuting* file).

4.3. *Mistract* File

A third diagnostic file created by the COMMUTE, AIRQUAL, and HAPEM programs is the *mistract* file. If the same *mistract* filename is used for the COMMUTE and AIRQUAL programs, the COMMUTE program's information will be overwritten by that of the AIRQUAL program. The HAPEM program will then append records onto an existing *mistract* file. To maintain a complete record of this information for a HAPEM7 simulation, either different *mistract* filenames should be used for the COMMUTE and AIRQUAL programs (requiring different *parameter* files), or the *mistract* file should be manually renamed 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, as detailed in the list below. Below the list are example excerpts from each program's *mistract* file.

- The COMMUTE program's *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 the COMMUTE program, as explained in the previous section, by creating a “commuting” flow using the *population* tract as both the home and work tract.

- The AIRQUAL program’s *mistract* file records the record number and the state, county, and tract FIPS codes 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 HAPEM7, since both these pieces of information about a tract (population and air quality) are needed to make an exposure estimate.
- The HAPEM program’s *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 air-quality 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 program’s *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 values of the home tract.

Example excerpt from the COMMUTE program’s *mistract* file

```
MISSING TRACTS OF COMMUTE & AIRQUAL in COMMUTE
       44        2        1 01003990000  0
      109        8        1 01015981903  0
      139       13        1 01025957601  0
[etc.]
```

Example excerpt from the AIRQUAL program’s *mistract* file

```
MISSING TRACTS for AIRQUAL & POPULATION DATA in AIRQUAL
     2375 04013980500
     5785 06037320000
     7049 06037980001
[etc.]
```

Example excerpt from the HAPEM program’s *mistract* file

```
MISSING TRACTS of AIRQUAL & COMMUTE IN HAPEM
 airtract match with worktract not found
home      3   2375 04013980500  0
 airtract match with worktract not found
[etc.]
```

4.4. *Final Exposure* File

As explained in Section 2.1.9 (Exposure Output Files), HAPEM7 creates an exposure output file for each combination of state and HAP. The names of these files are constructed by the model based on the HAP SAROAD code (specified by ***sarod*** in the *parameter* file) and the state FIPS code (as SAROAD.FIPS.dat).

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

Information at the top of the *final exposure* output file

State FIPs code	
HAP SAROAD code	(sarod)
HAP name	(pollutant)
HAP CAS number	(CAS)
Air-quality data units	(units)
Year of air-quality data	(year)
Number of outdoor-air-emission source categories	(nsource)
Random number seed for activity pattern selection	(Rseed1)
Random number seed for ME factors selection	(Rseed2)
Random number seed for air quality data selection	(Rseed3)
Number of indoor-product emission-sources types	(Footnote 7)
Number of indoor-material emission-source types	(Footnote 7)
Number of indoor-combustion emission-source types	(Footnote 7)
Number of vehicle-in-residential-garage emission-source types	(Footnote 7)
EPA Region of indoor-emission-source data	(Footnote 7)
Number of demographic groups (e.g., age groups in HAPEM7)	(ngroup)
Number of replicates for each demographic group (e.g., age group in HAPEM7)	(nreplic)
Definition of each demographic group (e.g., age group in HAPEM7), ordered as in the *population* file	(under "Demographic Groups:" heading in the parameter file for the AIRQUAL and HAPEM programs)

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
(in "wrapped" view)

```
ST CTY CENSUS GRUP POPUL   SOURCE01   SOURCE02   SOURCE03   SOURCE04  BackgConc
IndCon_Pro IndCon_Mat IndCon_Com IndCon_Veh Total Conc
```

The header record is then followed by ***nreplic*** data records for each combination of group and tract combination. The format of each data record, assuming ***nsource*** = 4, is provided in Table 4-2.

Table 4-2.
Variables in the *final exposure* output file

Field Numbers	Description	Format
1	leading space	
2–3	state FIPS code	(2-character string)
5–7	county FIPS code	(3-character string)
9–14	tract FIPS code	(6-character string)
16–17	demographic-group (e.g., age-group in HAPEM7) indicator	(integer 1–10, ordered as in the *population* input file)
19–25	number of people to which the exposure estimates in the data record apply	(decimal number) (equal to the population of the group/tract combination divided by ***nreplic***)
27–36	estimated exposure-concentration contribution from emission-source-category 1	(decimal number in scientific notation; units of measurement as in the *air quality* file)
38–47	estimated exposure-concentration contribution from emission-source-category 2	(decimal number in scientific notation; units of measurement as in the *air quality* file)
49–58	estimated exposure-concentration contribution from emission-source-category 3	(decimal number in scientific notation; units of measurement as in the *air quality* file)
60–69	estimated exposure-concentration contribution from emission-source-category 4	(decimal number in scientific notation; units of measurement as in the *air quality* file)
71–80	estimated exposure-concentration contribution from background	(decimal number in scientific notation; units of measurement as in the *air quality* file) (derived from the sum of the uniform background—***backg***—and the variable background concentrations)
82–91	estimated exposure-concentration contribution from indoor-product emission sources	(decimal number in scientific notation; units of measurement as in the *air quality* file)
93–102	estimated exposure-concentration contribution from building-materials indoor emissions	(decimal number in scientific notation; units of measurement as in the *air quality* file)
104–113	estimated exposure-concentration contribution from indoor-combustion emission sources	(decimal number in scientific notation; units of measurement as in the *air quality* file)
115–124	estimated exposure-concentration contribution from vehicles in attached garages	(decimal number in scientific notation; units of measurement as in the *air quality* file)
123–135	estimated total-exposure concentration	(decimal number in scientific notation; units of measurement as in the *air quality* file) (the sum of the preceding nine values)

An example of a set HAPEM7 exposure output records (for 30 replicates of one demographic group [e.g., age group in HAPEM7] in one tract) is presented below. The total population for group 1 in this tract is 36 and ***nreplic*** = 30, so that the number of people to which the exposure estimates in each record apply is 36/30 = 1.200.

Example set of exposure output records (for 30 replicates of one demographic group [e.g., age group in HAPEM7] in one tract)

```
ST CTY CENSUS GRUP POPUL   SOURCE01   SOURCE02  BackgConc IndCon_Pro IndCon_Mat IndCon_Com IndCon_Veh Total Conc
78 010 970100  1   1.200 0.4119E+00 0.1527E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4121E+00
78 010 970100  1   1.200 0.3939E+00 0.1699E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3941E+00
78 010 970100  1   1.200 0.3360E+00 0.1102E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3361E+00
78 010 970100  1   1.200 0.4791E+00 0.1654E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4793E+00
78 010 970100  1   1.200 0.4057E+00 0.1228E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4058E+00
78 010 970100  1   1.200 0.4652E+00 0.1805E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4654E+00
78 010 970100  1   1.200 0.4535E+00 0.1430E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4536E+00
78 010 970100  1   1.200 0.4412E+00 0.1815E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4414E+00
78 010 970100  1   1.200 0.3891E+00 0.1564E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3893E+00
78 010 970100  1   1.200 0.3771E+00 0.1278E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3772E+00
78 010 970100  1   1.200 0.3785E+00 0.1366E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3786E+00
78 010 970100  1   1.200 0.4700E+00 0.1662E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4702E+00
78 010 970100  1   1.200 0.3918E+00 0.1333E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3919E+00
78 010 970100  1   1.200 0.3923E+00 0.1632E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3925E+00
78 010 970100  1   1.200 0.3228E+00 0.8932E-04 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3229E+00
78 010 970100  1   1.200 0.2741E+00 0.8380E-04 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2742E+00
78 010 970100  1   1.200 0.5163E+00 0.2219E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.5165E+00
78 010 970100  1   1.200 0.3965E+00 0.1391E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3966E+00
78 010 970100  1   1.200 0.3831E+00 0.1792E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3833E+00
78 010 970100  1   1.200 0.3829E+00 0.1528E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3831E+00
78 010 970100  1   1.200 0.3717E+00 0.1006E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3718E+00
78 010 970100  1   1.200 0.4407E+00 0.1510E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4409E+00
78 010 970100  1   1.200 0.3463E+00 0.1926E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3465E+00
78 010 970100  1   1.200 0.5035E+00 0.1562E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.5037E+00
78 010 970100  1   1.200 0.4628E+00 0.1433E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4629E+00
78 010 970100  1   1.200 0.3847E+00 0.1175E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3848E+00
78 010 970100  1   1.200 0.3232E+00 0.3503E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3236E+00
78 010 970100  1   1.200 0.3479E+00 0.1098E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.3480E+00
78 010 970100  1   1.200 0.4049E+00 0.3598E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4053E+00
78 010 970100  1   1.200 0.4263E+00 0.1639E-03 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.4265E+00
```

This page intentionally left blank.

5. HAPEM7 Programs

This section contains detailed descriptions of the five programs that are contained in the model: DURAV, INDEXPOP, COMMUTE, AIRQUAL, and HAPEM. The first four programs (DURAV, INDEXPOP, COMMUTE, and AIRQUAL) are pre-processers that convert input-data files into the form required for efficient exposure calculations. The final program (HAPEM) performs the exposure calculations and summarizes the results.

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

5.1. Programming Guidelines Used to Develop HAPEM7

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

The model programs incorporate a structured programming style as summarized by the attributes listed below.

- No “GO TO” statements or line numbers are 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. Instead, this information is specified in the *parameter* file, which is read in from the command line.
- Most parameter values are input from the *parameter* file so that the programs themselves only allocate space for carrying out as many calculations as are necessary.
- Most 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 model 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 filenames and the parameter settings. This section is similar in all the model 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 provided in Table 5-1. Which filenames and paths are required by each model program are shown in Table 2.1 as user-defined files.

Table 5-1. The filename keywords in the *parameter* files recognized by the model programs

Keyword	Definition
activity	name of the *activity* file (input)
cluster	name of the *cluster* file (input)
ClusTrans	name of the *cluster-transition* probability file (input)
populat	name of the *population* file (input)
CommutTime	name of the *commuting-time* file (input)
CommutFrac	name of the *commuting-fraction* file (input)
DistToRoad	name of the *distance-to-road* file (input)
commut	name of the *commuting* file (input)
quality	name of the *air quality* file (input)
factors	name of the *factors* file (input)
mobiles	name of the *mobiles* file (input)
statefip	name of the *statefip* file (input)
log	name of the *log* file (output)
counter	name of the *counter* file (output)
mistract	name of the *mistract* file (output)
afile	path of *final exposure* file (output)
Product[1]	path of indoor source files (input)
AutoPdcut[1]	name of file for automobile-related consumer products (input)

[1] A path to one or more indoor emission source inputs for the indoor source algorithms is specified in these statements (with the AutoPduct statement including a filename). These algorithms are included in the HAPEM program, but have not yet been tested and reviewed. Therefore, they are currently not recommended for use, and instructions for their use are omitted from this document. To disable the indoor source algorithms, set keyword **CAS** to 99999, and specify any existing path (and file for AutoPduct; other than those otherwise specified for input or output for the HAPEM program) since no indoor source files will then actually be utilized by the HAPEM program.

The model 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 filenames 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 saves space and time by only using as much space as is necessary, allows for the parameters to be increased or decreased without recompiling the program, and 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

This section describes the purpose and structure of the processing section of each of the five model programs.

5.2.1. DURAV

As explained in Section 2.1.2 (The DURAV Program and the *Activity* and *Cluster* Files), the DURAV program performs the three main functions listed below.

- It categorizes and groups population activity data extracted from CHAD into demographic group (e.g., age group in HAPEM7), day types, commuting status, 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 six age groups in HAPEM7 are as follows, in years.

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

Currently, season and day of week are used to determine three day types as

- weekdays in summer (June–August),

- other weekdays, or
- weekends.

Cluster types are used to represent variations in activity pattern within each combination of demographic group (e.g., age group in HAPEM7), day type, and commuting status. There are 1–3 cluster types for each combination of group, day type, and commuting status. 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 operations described below.

- The *parameter* file, which stores input model variables and input file names, is read in from the command line.
- 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), specified in the *parameter* file, are checked for compatibility. As explained elsewhere in this guide, ***hblock*** must be an integral factor of ***nblock***, so that the activity time blocks can be combined if necessary to match ***hblock*.** If the check fails, then an error message is written to the *log* file and the program stops.

Further, the setup section of DURAV conducts the operation described below.

- The number of time blocks per day in the *activity* file is determined from the header record, as explained elsewhere in this guide. 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 stops.

Finally, the main processing section of DURAV conducts the several operations described below.

- 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 1,440 minutes. If the check fails, then the record is recorded in an intermediate output file with filename extension *.wrong_chad* and it is dropped from further processing.
- The ***nblock*** time blocks in each activity record are aggregated, if necessary, to create ***hblock*** time blocks.
- The aggregated activity records are checked to ensure the total activity time is still 1,440 minutes. If the check fails, an error message is written to the *log* file and the program stops.
- The aggregated activity records are written into a direct-access file with a filename extension of *.draft.*

- Each aggregated record is classified by demographic group (e.g., age group in HAPEM7), day type, and commuting status, as defined in the DURAV source code. If any records cannot be classified, and error message is written to the *log* file and the program stops. The sequence numbers of the aggregated 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 1,440 minutes, are recorded in the *counter* file.
- The number of aggregated records in each combination of demographic group (e.g., age group in HAPEM7), day type, and commuting status is determined.
- The aggregated activity records are read from the direct-access *.*draft* file and each is additionally classified by cluster within its combination of demographic group (e.g., age group in HAPEM7) and day type, according to the specifications in the *cluster* file.
- The number of aggregated activity records in each combination of demographic group (e.g., age group in HAPEM7), day type, commuting status, and cluster, and the number of clusters in each combination of group, day type, and commuting status, are recorded in an intermediate file with filename extension .*nonzero*.[12] This information is used in the HAPEM program, as described in Section 5.2.5 (HAPEM).
- The total number of aggregated activity records processed and their allocation among demographic group (e.g., age group in HAPEM7), day types, and commuting status is written into the *log* file.
- The activity patterns are written into a sequential file with filename extension *.da* sorted by demographic group (e.g., age group in HAPEM7), day type, commuting status, 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 and the *Population*, *Distance-to-road*, *Commuting-time*, and *Commuting-fraction* Files), the INDEXPOP program performs the two main functions listed below.

- It creates a direct-access file of population data to be used in the AIRQUAL program.
- It creates sequential ASCII index files for the population data census tracts, to facilitate file searching in the COMMUTE and AIRQUAL programs.
- It creates direct-access files and associated index files of the data in the *distance-to-road*, *commuting-time*, and *commuting-fraction* files, to be used in the COMMUTE and AIRQUAL programs.

[12] The *.*nonzero* file also records a flag for each combination of demographic group (e.g., age group in HAPEM7) and day type, indicating whether 10 percent of the activity patterns include commuting. This flag was used by an earlier version of the HAPEM program, but is not used in this version.

INDEXPOP Processing Operations

The specific operations performed in the main processing section of INDEXPOP are described below.

- The *parameter* file, which stores input model variables and input file names, is read in from the command line.
- Each record in the *distance-to-road* file is read and written into a direct-access file with filename extension *.dat*, and an associated index file is created with filename extension *.STIDX*.
- Each record in the *commuting-time* file is read and written into a direct-access file with filename extension *.dat*, and an associated index file is created with filename extension *.STIDX*.
- Each record in the *commuting-fraction* file is read and written into a direct-access file with filename extension*.dat*, and an associated index file is created with filename extension *.STIDX*.
- The number of data records in the *population* file is determined so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- Each data record in the *population* file is read. The population array is recorded in a direct-access file with the filename extension *.da*. The state FIPS, county FIPS, tract FIPS, 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 are in 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* file.
- 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 and the *Commuting*, *Distance-to-road*, *Commuting-time*, and *Commuting-fraction* Files), the COMMUTE program performs the three main functions described below.

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

COMMUTE Processing Operations

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

- The *parameter* file, which stores input model variables and input file names, is read in from the command line.
- The *distance-to-road* index file (filename extension *.STIDX*, created in INDEXPOP) is read twice: first to determine the number of records for array allocation, and then to populate the arrays with the data in the file.
- The *commuting-time* index file (filename extension *.STIDX*, created in INDEXPOP) is read twice: first to determine the number of records for array allocation, and then to populate the arrays with the data in the file.
- The *commuting-fraction* index file (filename extension *.STIDX*, created in INDEXPOP) is read twice: first to determine the number of records for array allocation, and then to populate the arrays with the data in the file.
- The number of data records in the *commuting* file is determined so that memory can be allocated for various arrays used to hold the input-data records and other data derived from them.
- The number of *commuting* file records with home tracts in each state is determined.
- For each state, the sequence numbers of the first and last data record indicating a home tract in that state are determined.
- The number of records in the *population* file is read from the *counter* file, so that memory can be allocated for various arrays used to hold the input data records and other data derived from them.
- All the tract FIPS 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 FIPS and number of work

tracts within 120 km are determined. For each home tract, the fractions of total commuting flow to work tracts, which are specified in the *commuting* file, are adjusted to the fractions of the total commuting flow within 120 km.

- For each home-tract/work-tract pair, the centroid-to-centroid distance from the *commuting* file is determined and a normalized distance is calculated as distance/(average distance).
- Each work-tract FIPS, its adjusted flow fraction, and its normalized distance are 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 tract FIPS and the indices of the tract in the **.state_county_pop_range*, **.county_tract_pop_range*, and the **_direct.ind* files.
- For *population* tracts with no matching *commuting* home tracts, a record is recorded in the **.da* file indicating the *population* tract as the work tract, with fractional commuting flow of 1.0 (i.e., all work takes place in the home tract).
- For each *population* tract, a record is written into a temporary index file. The fields in the record are the *population* tract FIPS, the sequence numbers of the first and last work tract record in the **.da* file, and a flag indicating whether the *population* tract was matched by a home tract in the *commuting* file (0=no; 1=yes).
- Two records are added to the *counter* file. The first record indicates the number of records found in the **_direct.ind* file (created by <u>INDEXPOP</u>) and the number of data records found in the *commuting* file. The second record records the number of records in the **.da* file and the number of records in the **.ind* file.
- A sequential index file is created with filename extension *.st_comm1_fip_range*. For each state, there is a record in this file indicating the sequence numbers of the first and last data record for tracts for that state in the temporary index file.
- The temporary index file is read from the beginning. Each record is matched by tract with a record in the *distance-to-road*, *commuting-time*, and *commuting-fraction* direct-access files (filename extensions of *.dat*, created in <u>INDEXPOP</u>). The combined data for each tract are written into a direct-access file with the root filename of the *commuting* file and the filename extension *.ind*.

5.2.4. <u>AIRQUAL</u>

As explained in Section 2.1.5 (The <u>AIRQUAL</u> Program and the *Air Quality* and *Distance-to-road* Files), the <u>AIRQUAL</u> program performs the four main functions listed below.

- It creates a sequential file of air-quality data to be used in the <u>HAPEM</u> program.
- 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 the <u>HAPEM</u> program.
- It adds the tract-specific information from the distance-to-road direct-access file (created in the <u>INDEXPOP</u> program) to the air-quality index files.

AIRQUAL Processing Operations

The specific operations performed in the main processing section of AIRQUAL are described below.

- The *parameter* file, which stores input model variables and input file names, is read in from the command line.
- 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 (The AIRQUAL Program and the *Air Quality* and *Distance-to-road* Files), ***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 check fails, an error message is written to the *log* file and the program stops.
- 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., *HAP.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., *population_v7.pop_air_da)*. The tract code (state FIPS, county FIPS, and tract FIPS) 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., *HAP.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 code 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. <u>HAPEM</u>

As explained in Section 2.1.6 (The <u>HAPEM</u> Program, the ME *Factors* and *Mobiles* Files, and the Activity *Cluster-transition* File), the <u>HAPEM</u> program performs the six main functions described below.

- For each demographic group (e.g., each age group in HAPEM7) in each census tract, it randomly selects ***nreplic*** sets of ME factors based on the distribution data provided in the *factors* and *mobiles* files. Each set contains a subset of ME factors randomly selected for each of the time blocks (for the *PEN* and *ADD* factors) or each of the sources (for the *PROX* and *LAG* factors). Each subset contains randomly selected ME factors for each of ***nmicro*** MEs.

- For each demographic group (e.g., each age group in HAPEM7) in each census tract, it randomly selects ***nreplic*** sets of air-quality data from the data sets available for a tract.

- For each demographic group (e.g., each age group in HAPEM7) 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 group/day-type/commuting-status combination. The weights are determined by the relative frequencies of cluster types randomly selected in a one-stage Markov process,[9] based on the cluster transition probabilities provided in the *cluster-transition* file.

- For each activity pattern for a commuting-demographic group (e.g., a commuting-age group in HAPEM7), 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 ME based on ME factors and outdoor concentrations.

- It combines activity patterns, commuting status, and estimates of ME concentration to calculate ***nreplic*** annual-average exposure concentrations for each demographic group (e.g., each age group in HAPEM7) in each census tract.

<u>HAPEM</u> Processing Operations

The specific operations performed in the main processing section of <u>HAPEM</u> are described below.

- The *parameter* file, which stores input model variables and input file names, is read in from the command line.
- The distribution data of ME factors for each of ***nmicro*** MEs is read from the *factors* and *mobiles* files (as appropriate) and saved into arrays.
- For the *PROX* distributions in the *mobiles* file for onroad-mobile sources, the average *PROX* factor for the second distance category (75–200 meters) over all the indoor MEs is calculated. (This value will be used later to calculate the ambient concentration for the third distance category [beyond 200 meters], as described below.)
- For each combination of demographic group (e.g., age group in HAPEM7), day type, and commuting status, the number of activity patterns for each cluster is read from the *.*nonzero* file created in DURAV.
- For each combination of demographic group (e.g., age group in HAPEM7), day type, and commuting status, the frequency of each cluster, and the cluster-to-cluster transition probabilities, are read from the *cluster-transition* file.
- For each combination of demographic group (e.g., age group in HAPEM7), day type, commuting status, and cluster with a positive number of activity records, the activity pattern records are read from the *.*da* file (created in DURAV) and the values saved into an array.
- Each activity pattern is checked to ensure a total activity time of 1,440 minutes. If this check fails, an error message is written to the *log* file and the program stops.
- Several values are read from the *counter* file to allocate memory for various arrays.
- Indices are read from the *.*state_air_fip_range* and *.*state_air1_fip_range* files (created by AIRQUAL).
- Data are read from the *.*pop_air_da* file and the index ranges for air records from the *.*air_da* file (created by AIRQUAL).
- Air-data records are read from **.da* files created by AIRQUAL.
- Indices are read from the *.*st_comm1_fip_range* and *.*ind* files created by COMMUTE, and data are read from the *.*da* file created by COMMUTE.
- For each tract in the *.*ind* file created by COMMUTE, an attempt is made to find a matching tract in the *.*state_air_fip_range* file created by AIRQUAL. If a match is not found, the *commuting* tract is recorded in the *mistract* file.
- For each demographic group (e.g., age group in HAPEM7) in each census tract, ***nreplic*** sets of ME factors are randomly selected based on the distribution data provided in the *factors* and *mobiles* files, using subroutines “DISTRIBUTION” and “DATASET”. Each set contains a subset of ME factors randomly selected for each time block (for the *PEN* and *ADD* factors) or each source (for the *PROX* factor). For onroad-mobile source categories, first a distance-from-source category is selected for each indoor ME based on the population fractions in each distance category that were taken from the *distance-*

to-road file and added to the *commuting* index file in COMMUTE.[13] Then, a *PROX* factor for each indoor ME is selected from the appropriate distribution. Each subset contains randomly selected ME factors for each of ***nmicro*** MEs.

- For each demographic group (e.g., age group in HAPEM7) in each census tract, ***nreplic*** sets of air-quality data are randomly selected from the data sets available for the census tract in the **.da*" file created by AIRQUAL.
- When a single set of ambient concentrations are provided for each tract in the *air quality* file (as is typically the case), they represent spatial averages over the tract, excluding locations very close to an emission source. For onroad-mobile source categories, it is assumed that the ambient concentrations in the *air quality* file represent spatial averages over the second and third distance categories (the distances 75–200 meters and beyond 200 meters) for the *distance-to-road* and *mobiles* files. Because HAPEM estimates the ambient concentration for the second distance category by applying a *PROX* factor to the "tract-average" ambient concentration, the ambient concentration for the third distance category is also adjusted to make the area-weighted average over these two distance categories equal to the "tract average". This is done as shown below.

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

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

$$CONC_{D3} = \frac{CONC_{AQ}}{\left(AREA_{D3} + AREA_{D2} \times PROX_{D2}\right)}$$

where:

$CONC_{AQ}$: the "tract-average" concentration from the *air quality* file,

$CONC_{D2}$: average ambient concentration in second distance category (75–200 meters),

$CONC_{D3}$: average ambient concentration in third distance category (beyond 200 meters),

$AREA_{D2}$: fraction of the tract area in the second distance category (from the *distance-to-road* file),

$AREA_{D3}$: fraction of the tract area in the third distance category (from the *distance-to-road* file), and

$PROX_{D2}$: average *PROX* factor for the second distance category over all the indoor-source categories (calculated above).[14]

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

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

- 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 (e.g., age group in HAPEM7) 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), the variable background-concentration values for each data record in **.da* file created by AIRQUAL, and the randomly selected ME factors.
- For each census-tract, demographic-group (e.g., age-group in HAPEM7), and day-type replicate, a commuting status is selected based on the data from the *commuting-fraction* file (that were added to the *commuting* index file in COMMUTE). If the replicate is a commuter, then a commuting mode (public or private transit) is randomly selected based on the data from the *commuting-time* file (that were added to the *commuting* index file in COMMUTE). This selection also determines an associated average commuting time for the tract.
- For each replicate that commutes, a work tract is randomly selected for each selected activity pattern, using the attached subroutine “RANDOMR”. The work tract is selected from the set of work tracts corresponding to that home tract, as specified in the **.da* file created by COMMUTE. The air-quality data for that work tract are randomly selected from the data sets available for the work tract in the **air_da* file created by AIRQUAL. If the work tract cannot be found in the **.air_da* file, the air-quality data for the home tract are used. The air-quality data are adjusted and combined with the ME factors randomly selected in the same way as the home tract, in order to estimate the concentrations for each ME/time-block combination for that work tract.
- For each replicate/day-type combination, an average activity pattern is calculated as the weighted average of activity patterns randomly selected from each cluster in a combination of demographic group (e.g., age group in HAPEM7), day type, and commuting status 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 *cluster-transition* file.
- The average activity pattern for the day-type is adjusted so that the commuting time for the replicate is equal to the product of the tract-average commuting time for the commuting mode selected above, and the normalized home-tract/work-tract distance calculated in COMMUTE and recorded in the *commuting* direct-access file (created in COMMUTE). The adjustments are made by uniform scaling of the time in each time block for commuting MEs (so that the sum matches the total calculated commuting time), and corresponding uniform scaling of the time in each time block for non-commuting MEs.
- The ME/time-block time durations of the weighted-averaged activity patterns are combined with the estimated ME/time-block concentrations for the home tract and the work tracts to estimate ***nreplic*** exposure concentrations for each combination of demographic group (e.g., age group in HAPEM7) and day type. A separate set of estimates is made for each emission-source category. The algorithm for each combination of group and day type in the tract is as follows.

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

where:

$Conc_{t,m}$: the emission-source-category concentration during time-block *t* in ME *m*, and

$Duration_{t,m}$: the duration of activity during time-block *t* in ME *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: 0.178 for summer weekday, 0.537 for other weekdays, and 0.285 for weekends.
- A total annual-average exposure concentration is calculated by summing the annual-average values for each emission-source category, the background contribution, and from the indoor-source *ADD* factor.
- The results are written into the *final exposure* output files, with ***nreplic*** records for each demographic group (e.g., age group in HAPEM7) in each tract. The format of the files is described in Section 4.4 (*Final Exposure* File).

6. References

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

Rosenbaum, A.S. and M. Huang, 2007: The HAPEM6 User's Guide, Hazardous Air Pollutant Exposure Model, Version 6. Prepared for Ted Palma, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, NC. http://www2.epa.gov/sites/production/files/2013-08/documents/hapem6_guide.pdf.

This page intentionally left blank.

Appendix A: Update to: Proposed modification of HAPEM algorithm for creating longitudinal activity patterns: Results of data analysis

This page intentionally left blank.

MEMORANDUM

To: Ted Palma and Terri Hollingsworth
U.S. EPA, Office of Air Quality Planning and Standards

From: Jonathan Cohen, Isaac Warren, and Chris Holder
ICF International

Date: 03/30/2015

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

1. Summary

In 2002, ICF (“we”) proposed and implemented a new algorithm for the Hazardous Air Pollutant Exposure Model (HAPEM) to create longitudinal activity patterns using Markov chains. We implemented the algorithm starting with HAPEM5 and described it in a 2002 memorandum,[1] which was also made Appendix A of the HAPEM6 User’s Guide.[2] This memorandum builds on that 2002 memorandum, updates that discussion to describe how, for HAPEM7, we refit the Markov chain model to a more-recent version of the U.S. Environmental Protection Agency (EPA) Consolidated Human Activity Database (CHAD) that now includes more activity pattern studies and thus more daily activity patterns.

For HAPEM5, the data analysis used the default grouping of daily activity patterns into 30 combinations of day type (i.e., summer weekday, non-summer weekday, and weekend) and demographic group (i.e., males or females; age groups: 0–4, 5–11, 12–17, 18–64, and 65 and older). For HAPEM6 and HAPEM7, an updated data analysis incorporated an expanded CHAD database and revised the approach to use a default grouping of daily activity patterns into 36 combinations of day type (i.e., summer weekday, non-summer weekday, and weekend), age group (i.e., 0–1, 2–4, 5–15, 16–17, 18–64, and 65 years and older), and commuter type (i.e., commutes or does not commute). Including grouping by commuter type better reflects the different activity patterns and exposures between commuters and non-commuters. We did not update the 2002 memorandum or the HAPEM6 Users’ Guide Appendix A at that time to reflect the final 36 day-age-commuter combinations.

[1] ICF and ECR Memorandum “Proposed modification of HAPEM algorithm for creating longitudinal activity patterns: Results of data analysis.” from July 23, 2002, addressed to Ted Palma at EPA’s Office of Air Quality Planning and Standards.

[2] The HAPEM6 User’s Guide is available as of January 20, 2015 at http://www2.epa.gov/fera/hazardous-air-pollutant-exposure-model-hapem-users-guides.

For HAPEM7, the analysis used the CHAD-Master database as of July 2013, which was the most current version as of June 2014. The HAPEM7 data analysis grouped the CHAD daily activity patterns into one, two, or three categories (or clusters) of similar patterns for each of the same 36 day-age-commuter combinations used for HAPEM6. In HAPEM, for each day-age-commuter combination, 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.

In the sections below, we discuss how and why we updated HAPEM's grouping of activity patterns into clusters and HAPEM's estimations of the day-to-day activities of a modeled subject. The approach applies to HAPEM starting with version 5, though some details are specific to HAPEM7. We also discuss below the accompanying Microsoft® Excel™ file "cluster.feb2015.r1a.xlsx."

2. Background, before HAPEM5

The original approach used for preliminary NATA simulations prior to 2002 selected with replacement (365) 24- hour activity patterns for each demographic 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 demographic-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 demographic-tract combination, rather than the variability of activity patterns among group members.

In response to comments from EPA's Science Advisory Board, this approach was modified to try to represent the variability among individuals within a demographic-tract combination. For each demographic-tract combination, three group-specific activity patterns were selected, one for each day-of-week and season combination. This approach implied that, for any individual, the activity pattern is identical for every day in a day-of-week and season category; that is, the probability of transition to a different pattern equals zero. This approach tends to maximize the differences between individuals, perhaps to an unrealistic extent.

3. Updated HAPEM Algorithm, HAPEM5 and Later Versions

To improve this approach, starting with HAPEM5 we revised the algorithm to treat transition probabilities in more detail. Information could be used on the probabilities of changes among daily activity patterns for a single individual, stratified by day type and commuter type. For example, for some demographic-day-commuter combination, suppose activity patterns could be grouped into two categories, A and B, based on differences among the times spent in various microenvironments (MEs).

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 and commuter type (e.g., non-summer weekdays for a commuter).

1. For day 1, randomly select an initial activity pattern for non-summer weekdays for a commuter. Call that a type-A pattern.
2. For day 2, either retain the same activity pattern with probability P_{AA}, or randomly select a type-B pattern with probability P_{AB}.
3. If a type-A activity pattern is selected for day 2, repeat Step 2 to find the activity pattern for day 3. If a type-B activity pattern is selected for day 2, either retain the same activity pattern for day 3 with probability P_{BB}, or use the same type-A pattern selected in Step 1 with probability P_{BA}. 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 the absence of sufficient 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.

In the sections below, we discuss how and why we implemented the above methods for HAPEM Version 5 and later, with some details specific to HAPEM7.

3.1. Grouping Days into Categories

The first data-analysis task was to use the CHAD data to group activity pattern days into categories for each day-age-commuter combination. First, we summarized each daily activity pattern by the total minutes in each of five broad MEs: Indoors Residence, Indoors Other, Outdoors Near-roadway, Outdoors Other, and In-vehicle. We assumed these five numbers represent the most important features of the activity pattern for their exposure impact (the five numbers are not independent since they sum to 1440). We analyzed separately each day-age-commuter combination. 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.” We used a cluster analysis to group the activity pattern days in each day-age-commuter combination into clusters of days with similar values of the minutes in each of the five MEs (i.e., 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.

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); that is, 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 (which we did not ultimately use) 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 (i.e., groupings of the vectors into k clusters) is huge, the k-means algorithm chooses an initial set of k cluster seeds (i.e., 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 one 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 using 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 (which we ultimately used), 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. We chose Ward’s method these analyses.

An important consideration is whether or not to rescale the measurement variables before applying the clustering algorithm (we did not ultimately rescale the measurement variables). 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 a mean of zero and a standard deviation of one. If all measurements are in the same units, as in the present case (i.e., minutes in an ME), then the statistical literature is less definitive on the need for rescaling. A classic textbook, Hartigan's *Clustering Algorithms*,[3] points out that rescaling to a constant variance often tends to down-weight variables that cluster well. For these analyses, we dud bit rescale the five measurement.

After applying Ward's method to the CHAD data, we calculated the number of clusters for each day-age-commuter combination. If the measurement variables are uncorrelated, then various statistical measures (e.g., pseudo-F statistic, pseudo-t2 statistic, cubic clustering criterion, and so on) 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 we could not apply the various statistical stopping rules.[4] 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. We decided to have at most three clusters and to require a minimum of five pairs of consecutive days for each initial cluster, where the first day is in the initial cluster and the second day is for the same individual on the next calendar day. On this basis, in HAPEM7, we chose three clusters for 17 of the 36 day-age-commuter combinations, two clusters for two combinations, and one cluster for the other 17 combinations.[5]

The result of the Ward method cluster analysis was an assignment of every CHAD activity pattern day to a cluster. In the accompanying Microsoft® Excel™ file "cluster.feb2015.r1a.xlsx", the worksheet FINALTREE contains the assigned cluster number (1, 2, or 3) for each CHAD event record, the numbers of minutes in each of the five MEs, and also includes the day type, age group, commuter type, and number of clusters for that day-age-commuter combination.

[3] Hartigan, J. A. *Clustering Algorithms*; John Wiley & Sons., Inc.: New York, NY, 1975.

[4] 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, using an earlier CHAD dataset, we repeated the cluster analysis using all but the time-in-residence, which is usually where the greatest exposure time occurs. We 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 second least-populous clusters.

[5] There were no activity patterns for the day-age-commuter = 1-1-2, but we included that case in cluster-transition file in order for HAPEM to execute properly.

3.2. Estimating Transition Probabilities

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

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.

= $\text{trans}_{i1} + \text{trans}_{i2} + \text{trans}_{i3}$

prob_{ij} = $\text{trans}_{ij} \div \text{trans}_{ix}$

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.

We provide the results of this analysis in Table 1 and Table 2 presented below and in the FINALTRANS worksheet in the accompanying Microsoft® Excel™ file "cluster.feb2015.r1a.xlsx",[6] including the transition counts and estimated transition probabilities for each day-age-commuter combination. For the two combinations with only two clusters, the values with i or j equal to 3 are missing or zero, since cluster 3 is undefined. For the 17 combinations with only 1 cluster, the values with i or j equal to 2 or 3 are missing or zero, since clusters 2 and 3 are undefined. We also include the variables cluster1, cluster2, and cluster3, giving the total numbers of CHAD activity patterns in clusters 1, 2, and 3, respectively and the associated probabilities for each cluster. Instead of the probabilities themselves, the HAPEM input files use cumulative probabilities, such as cprob_{ij}, which is the probability of a transition from cluster_i to a cluster_k <= j (e.g., cprob32 is the probability of a transition from cluster 3 either to cluster 1 or 2). We include the cumulative probabilities in the Excel™ file but not in Table 1 and Table 2.

There were seven day-age-commuter combinations where there were no individuals with consecutive days in CHAD. For those combinations, there are no transitions available to estimate transition probabilities. We assigned all those activity patterns to cluster 1 and assigned a transition probability of 1 for trans11.

4. 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 might be possible to simplify the algorithm considerably by applying well-established concepts from Markov chain theory (we ultimately rejected that

[6] We used the FINALTREE and FINALTRANS sheets of this Excel™ file to produce the HAPEM7 default cluster and cluster-transition input files.

simplification). 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 weekdays, 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.

Because there were no estimated transition probabilities between different clusters that were equal to zero, the chains for each day-age-commuter combination are irreducible, aperiodic, and recurrent, so that steady-state, limiting probabilities exist. These steady-state probabilities are found by solving the linear equations:

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

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, data analyses of a previous version of CHAD using simulated seasons for each day type and demographic group showed that the numbers of days per cluster varies significantly around the limiting value. This analysis leads to our recommendation that the HAPEM algorithm uses the transition probabilities to directly simulate the cluster transitions, instead of using the steady-state estimates of the number of days per cluster.

Table 1. HAPEM7 Transition Counts

Day Type	Demographic (Age) Group	Commuter Type (1 = no commuting, 2 = commuting)	Number of Clusters	Transition Counts Between Clusters												
				Trans All	Trans 1x	Trans 2x	Trans 3x	Trans 11	Trans 12	Trans 13	Trans 21	Trans 22	Trans 23	Trans 31	Trans 32	Trans 33
1	1	1	1	27	27	0	0	27	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	3	544	255	214	75	154	79	22	80	101	33	24	28	23
1	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	1	3	1,176	220	255	701	97	31	92	35	84	136	101	117	483
1	3	2	1	5	5	0	0	5	0	0	0	0	0	0	0	0
1	4	1	1	11	11	0	0	11	0	0	0	0	0	0	0	0
1	4	2	1	6	6	0	0	6	0	0	0	0	0	0	0	0
1	5	1	3	1,480	539	863	78	303	226	10	240	579	44	16	44	18
1	5	2	3	2,160	1,096	314	750	874	43	179	45	196	73	188	79	483
1	6	1	3	1,090	451	491	148	305	123	23	133	280	78	27	72	49
1	6	2	3	252	60	123	69	43	9	8	13	96	14	8	19	42
2	1	1	2	37	29	8	0	28	1	0	4	4	0	0	0	0
2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	1	3	147	57	39	51	38	9	10	3	25	11	10	10	31
2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	3	1	3	417	115	251	51	59	52	4	37	170	44	9	30	12
2	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	4	1	1	26	26	0	0	26	0	0	0	0	0	0	0	0
2	4	2	1	5	5	0	0	5	0	0	0	0	0	0	0	0
2	5	1	3	317	130	164	23	72	53	5	51	104	9	4	8	11
2	5	2	3	754	479	148	127	344	39	96	33	104	11	69	8	50
2	6	1	3	882	382	428	72	223	137	22	132	280	16	20	12	40
2	6	2	3	195	161	23	11	151	8	2	12	9	2	4	1	6

Table 1. HAPEM7 Transition Counts

Day Type	Demographic (Age) Group	Commuter Type (1 = no commuting, 2 = commuting)	Number of Clusters	Transition Counts Between Clusters												
				Trans All	Trans 1x	Trans 2x	Trans 3x	Trans 11	Trans 12	Trans 13	Trans 21	Trans 22	Trans 23	Trans 31	Trans 32	Trans 33
3	1	1	1	18	18	0	0	18	0	0	0	0	0	0	0	0
3	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	1	3	80	48	17	15	29	8	11	7	5	5	8	4	3
3	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3	1	3	116	68	27	21	54	10	4	14	12	1	13	3	5
3	3	2	1	2	2	0	0	2	0	0	0	0	0	0	0	0
3	4	1	2	16	11	5	0	10	1	0	1	4	0	0	0	0
3	4	2	1	2	2	0	0	2	0	0	0	0	0	0	0	0
3	5	1	3	293	233	42	18	200	22	11	32	9	1	10	2	6
3	5	2	3	63	26	25	12	18	4	4	4	17	4	2	3	7
3	6	1	3	264	149	98	17	106	39	4	43	50	5	4	6	7
3	6	2	1	3	3	0	0	3	0	0	0	0	0	0	0	0

Table 2. HAPEM7 Transition Probabilities

Day Type	Demographic (Age) Group	Commuter Type (1 = no commuting, 2 = commuting)	Number of Clusters	Transition Probabilities Between Clusters								
				Prob 11	Prob 12	Prob 13	Prob 21	Prob 22	Prob 23	Prob 31	Prob 32	Prob 33
1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	2	1	1	0	0	0	0	0	0	0	0
1	2	1	3	0.604	0.310	0.086	0.374	0.472	0.154	0.320	0.373	0.307
1	2	2	1	1	0	0	0	0	0	0	0	0
1	3	1	3	0.441	0.141	0.418	0.137	0.329	0.533	0.144	0.167	0.689
1	3	2	1	1	0	0	0	0	0	0	0	0
1	4	1	1	1	0	0	0	0	0	0	0	0
1	4	2	1	1	0	0	0	0	0	0	0	0
1	5	1	3	0.562	0.419	0.019	0.278	0.671	0.051	0.205	0.564	0.231
1	5	2	3	0.797	0.039	0.163	0.143	0.624	0.232	0.251	0.105	0.644
1	6	1	3	0.676	0.273	0.051	0.271	0.570	0.159	0.182	0.486	0.331
1	6	2	3	0.717	0.150	0.133	0.106	0.780	0.114	0.116	0.275	0.609
2	1	1	2	0.966	0.034	0	0.500	0.500	0	0	0	0
2	1	2	1	1	0	0	0	0	0	0	0	0
2	2	1	3	0.667	0.158	0.175	0.077	0.641	0.282	0.196	0.196	0.608
2	2	2	1	1	0	0	0	0	0	0	0	0
2	3	1	3	0.513	0.452	0.035	0.147	0.677	0.175	0.176	0.588	0.235
2	3	2	1	1	0	0	0	0	0	0	0	0
2	4	1	1	1	0	0	0	0	0	0	0	0
2	4	2	1	1	0	0	0	0	0	0	0	0
2	5	1	3	0.554	0.408	0.038	0.311	0.634	0.055	0.174	0.348	0.478
2	5	2	3	0.718	0.081	0.200	0.223	0.703	0.074	0.543	0.063	0.394
2	6	1	3	0.584	0.359	0.058	0.308	0.654	0.037	0.278	0.167	0.556
2	6	2	3	0.938	0.050	0.012	0.522	0.391	0.087	0.364	0.091	0.545

Table 2. HAPEM7 Transition Probabilities

Day Type	Demographic (Age) Group	Commuter Type (1 = no commuting, 2 = commuting)	Number of Clusters	Transition Probabilities Between Clusters								
				Prob 11	Prob 12	Prob 13	Prob 21	Prob 22	Prob 23	Prob 31	Prob 32	Prob 33
3	1	1	1	1	0	0	0	0	0	0	0	0
3	1	2	1	1	0	0	0	0	0	0	0	0
3	2	1	3	0.604	0.167	0.229	0.412	0.294	0.294	0.533	0.267	0.200
3	2	2	1	1	0	0	0	0	0	0	0	0
3	3	1	3	0.794	0.147	0.059	0.519	0.444	0.037	0.619	0.143	0.238
3	3	2	1	1	0	0	0	0	0	0	0	0
3	4	1	2	0.909	0.091	0	0.200	0.800	0	0	0	0
3	4	2	1	1	0	0	0	0	0	0	0	0
3	5	1	3	0.858	0.094	0.047	0.762	0.214	0.024	0.556	0.111	0.333
3	5	2	3	0.692	0.154	0.154	0.160	0.680	0.160	0.167	0.250	0.583
3	6	1	3	0.711	0.262	0.027	0.439	0.510	0.051	0.235	0.353	0.412
3	6	2	1	1	0	0	0	0	0	0	0	0

This page intentionally left blank.

Appendix B: Updating the Hazardous Air Pollutant Exposure Model (HAPEM) for Use in the 2011 National-scale Air Toxics Assessment (NATA)

This page intentionally left blank.

APPENDIX B

MEMORANDUM

To: Ted Palma and Terri Hollingsworth
U.S. EPA, Office of Air Quality Planning and Standards

From: Chris Holder, Isaac Warren, Jonathan Cohen, Casey Cavanagh, Angelica McGee, Chris Stevens, and Kevin Wright
ICF International

Date: 04/08/2015

Re: Updating the Hazardous Air Pollutant Exposure Model (HAPEM) for Use in the 2011 National-scale Air Toxics Assessment (NATA)

ICF ("we") updated the default, ancillary files accompanying the Hazardous Air Pollution Exposure Model (HAPEM), and we updated some of the HAPEM source code to accommodate the new default files. The resulting new version of HAPEM (i.e., HAPEM7), with its default files, models exposure concentrations for all populated census tracts according to the 2010 U.S. Census, using demographic and other behavior data from the 2010 Census and a recent version of the U.S. Environmental Protection Agency (EPA) Consolidated Human Activity Database (CHAD). In this technical memorandum, we describe how we updated the default files and source code, including the quality-assurance (QA) steps we used and the format of the final default files. There is generally one section or subsection on each topic (e.g., one subsection for the population file, one section for the commuting file, and so on). HAPEM7 and its updated default files will be available for download as EPA's latest, default version of HAPEM.[1] We modeled exposure concentrations using HAPEM7 for the 2011 National-scale Air Toxics Assessment (NATA), as described in a separate memorandum.[2]

[1] We anticipate HAPEM7 and its User's Guide will be made available by EPA online in Spring 2015.

[2] We describe the use of HAPEM7 in the 2011 NATA in the ICF Memorandum "Running HAPEM7 for the 2011 National-scale Air Toxics Assessment (NATA)." from April 8, 2015, to Ted Palma and Terri Hollingsworth of EPA's Office of Air Quality Planning and Standards (OAQPS).

1. Introduction to HAPEM and its Use in NATA

HAPEM is a model used by EPA to perform screening-level assessments of long-term inhalation exposures to hazardous air pollutants (HAPs). By using the default, ancillary files provided by EPA for download with HAPEM,[3] exposure concentrations output by HAPEM are stratified by location (i.e., U.S. Census tract), time of day, age group, and the individual source categories and HAPs being modeled. These default files cover all 50 states in the US, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands.[4]

NATA uses HAPEM with these default files, so exposure concentrations produced for NATA have the same stratifications discussed above, though NATA-specific post-processing includes accumulating exposure concentrations into a lifetime period of exposure. NATA is a nationwide modeling assessment of air concentrations, exposure concentrations, and potential, chronic human health cancer risks and hazards associated with HAP emissions from man-made and naturally occurring sources. These NATA results are spatially partitioned by county and by tract. EPA models air concentrations using AERMOD (the atmospheric dispersion model developed by the American Meteorological Society and the EPA Regulatory Model Improvement Committee) and CMAQ (EPA's Community Multiscale Air Quality model). Those modeled air concentrations are the "air quality" inputs for HAPEM. NATA is not an enforcement tool to determine compliance with various standards of emissions, air quality, or health impacts; rather, it is a screening-level tool used to rank HAPs based on potential health impacts (nationally and locally), estimate the numbers of people and demographics potentially subject to health risks above levels of concern, identify gaps in data, and prioritize locations, source categories, and HAPs to inform additional data collection and assessment.

Data on where people live, work, and otherwise spend their time are critical to the completeness of the exposure modeling conducted with HAPEM. The version of HAPEM currently available for download (HAPEM6) uses U.S. Census data from the year 2000 as well as activity patterns gleaned from a version of CHAD available in 2002.[5] Parallel with the 2011 NATA effort, we have updated the default files used by HAPEM to reflect 2010 Census data and the version of CHAD available in June 2014. We have also updated HAPEM source code as necessary, mostly to accommodate the array sizes of the updated data.

2. Updating Census-based Data

2.1. Population File – "population_v7.txt"

The HAPEM default population input file ("CENSUS2000.txt" in HAPEM6; "population_v7.txt" in HAPEM7) provides the number of people in each HAPEM age group residing in each tract in the 50 states plus the District of Columbia, Puerto Rico, and the U.S. Virgin Islands.[4] The HAPEM6 population

[3] As of January 20, 2015, HAPEM6 is available for download at http://www2.epa.gov/fera/download-hazardous-air-pollutant-exposure-model-hapem.

[4] The new HAPEM files discussed in this memorandum include data for the U.S. Virgin Islands, whereas the files they replace did not. See individual sections of this memorandum for information on the completeness of data available for the U.S. Virgin Islands.

[5] The content, functionality, and implementation of HAPEM6 are discussed in the HAPEM6 User's Guide, available as of January 20, 2015 at http://www2.epa.gov/fera/hazardous-air-pollutant-exposure-model-hapem-users-guides.

data were derived from the 2000 Census; the specific Census table was not named in the HAPEM6 User's Guide, though it was likely the main population-by-age table in the Census' Summary File 1, Table PCT12: "Sex by Age".

For HAPEM7, we used the 2010 Census' Table PCT12 to update the HAPEM population file for all areas except the U.S. Virgin Islands (see next paragraph for more information on the U.S. Virgin Islands). We downloaded Table PCT12 under a previous work assignment related to updating population files for EPA's Air Pollutants Exposure Model (APEX).[6] The APEX population files ("pop_m_all_2010.txt" and "pop_f_all_2010.txt") respectively contained male and female population counts by tract for all incremental ages 0 to "99+". The HAPEM7 default ages are binned into six groups: 0–1, 2–4, 5–15, 16–17, 18–64, and 65 and older. We processed these two APEX population files to produce the HAPEM7 population file for all areas except the U.S. Virgin Islands, summing across the two genders (in APEX) and across the six age groups (in HAPEM).

Population data for the U.S. Virgin Islands were not available from the 2010 Census' Table PCT12 and are not used in APEX. For updating the HAPEM population file, we gathered 2010 population data for the U.S. Virgin Islands from other Census surveys available by querying the Census' American FactFinder web page.[7] These data were not available for each incremental age or for the same age groups used in HAPEM. For the purposes of fitting the Census age groups to the HAPEM age groups, we assumed population counts were evenly distributed among the incremental years represented in the Census 0–4 group (i.e., two fifths being 0–1 and three fifths being 2–4) and in the 15–17 group (i.e., one third being 15 and two thirds being 16–17); all other Census age groups (e.g., 5–9, 10–14, 18–19,…,62–64, 65–66,…,85 and over) required no subdivision to fit into the HAPEM age groups. We completed the HAPEM7 population file by appending these data for the U.S. Virgin Islands to the data gathered for the 50 states, District of Columbia, and Puerto Rico.

2.1.1. Quality Assurance

We thoroughly checked the accuracy of each of the calculations we used to generate the HAPEM7 population file. We also compared our processed population counts in multiple tracts against data retrieved from web-based queries of 2010 Census data; one error was caught with this method, though it was due to one inaccurate value in the APEX population file, which we corrected both for HAPEM and for APEX. Finally, we compared the HAPEM7 population file against that of HAPEM6 to ensure proper formatting; one error was caught with this method, where table headers were repeated at the bottom of the file, which we removed.

2.1.2. Content and Format

The HAPEM7 population data are contained in a fixed-with, space-delimited text file with characteristics shown in Table 1. The file contains seven columns and a total of 74,034 rows of data (after two header

[6] Work Assignment 1-12, Amendment 2, Task 5, as described in the ICF Memorandum "Updating APEX Input Files for 2010 Census Population and Employment Probability" from January 30, 2014, to Stephen Graham of EPA-OAQPS.

[7] As of January 21, 2015, the U.S. Census American FactFinder is available at http://factfinder2.census.gov/.

rows). Each data row corresponds to a tract, where the first field identifies the tract using Census FIPS coding,[8] and fields 2–7 contain population counts per age group. Population counts are whole numbers (no commas separating thousands). The first header row labels the fields, where the age-group columns are identified by the youngest age within the group (i.e., B_00 for age group 0–1, B_02 for age group 2–4, and so on). The second header row serves an unknown purpose but was retained from the HAPEM6 population file.

Table 1. Characteristics of the HAPEM7 Population File

Variable	Description	Character Start Position on Data Row	Character Length on Data Row[a]
TRACT	Full Census FIPS Code for Home Tract	1	11
B_00	Total Population Ages 0–1	17	8
B_02	Total Population Ages 2–4	25	8
B_05	Total Population Ages 5–15	33	8
B_16	Total Population Ages 16–17	41	8
B_18	Total Population Ages 18–64	49	8
B_65	Total Population Ages 65 and older	57	8

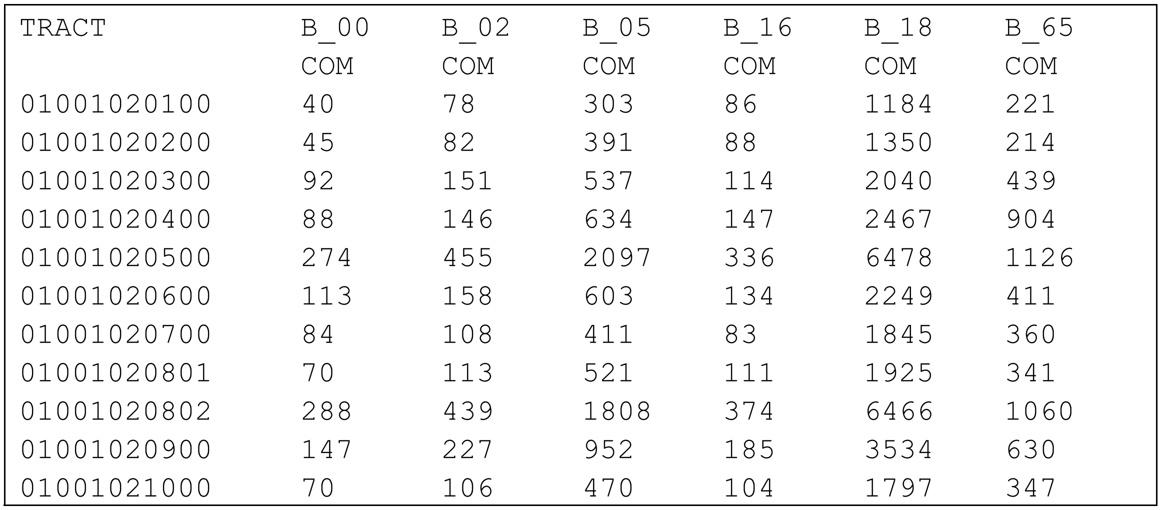
[a] Any unused character space after a number and/or between fields consists of blank spaces.

Figure 1 and Figure 2 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 population files.

```
TRACT               B_00     B_02     B_05     B_16     B_18     B_65
                    COM      COM      COM      COM      COM      COM
01001020100         44       50       357      68       1225     176
01001020200         58       94       321      57       1148     214
01001020300         98       147      609      106      1924     453
01001020400         104      145      740      134      2730     702
01001020500         166      280      1237     184      3751     422
01001020600         79       146      691      109      2045     308
01001020700         90       124      462      96       1815     313
01001020800         276      381      1733     324      5939     704
01001020900         124      205      873      171      2787     468
01001021000         78       109      484      89       1584     331
```

Figure 1. Excerpt from the HAPEM6 Population File (“CENSUS2000.txt”)

[8] The full tract identifier used by Census consists of a 2-digit state code, a 3-digit county code, and a 6-digit tract code, concatenated together to form an 11-digit code.



```
TRACT            B_00     B_02     B_05     B_16     B_18     B_65
                 COM      COM      COM      COM      COM      COM
01001020100      40       78       303      86       1184     221
01001020200      45       82       391      88       1350     214
01001020300      92       151      537      114      2040     439
01001020400      88       146      634      147      2467     904
01001020500      274      455      2097     336      6478     1126
01001020600      113      158      603      134      2249     411
01001020700      84       108      411      83       1845     360
01001020801      70       113      521      111      1925     341
01001020802      288      439      1808     374      6466     1060
01001020900      147      227      952      185      3534     630
01001021000      70       106      470      104      1797     347
```

Figure 2. Excerpt from the HAPEM7 Population File ("population_v7.txt")

In the HAPEM7 population file, total tract populations range from 0 (for 579 tracts across 43 states and territories, which is less than 1 percent of all tracts) to 37,452, with an average of 4,222. The total population in this file is 312,577,732.

2.2. Commuting File – "commute_flow_v7.txt"

In HAPEM, the tract where a person resides is their home tract, and the tract where a person works is their work tract. Some people work within their home tract (i.e., work tract is the home tract); the remaining employed people work outside their home tract. For the employed people in each home tract, the HAPEM default commuting input file ("comm2000.txt" in HAPEM6; "commute_flow_v7.txt" in HAPEM7) provides the fraction of those people who work within their home tract as well as the fraction that commute to work in each other tract. For each home tract, the file only contains the tract(s) where residents of the home tract work (i.e., there are no fractions of 0). These commuting data are provided for nearly all of the (home) tracts contained in the HAPEM population file, with exceptions noted in the discussion below.

The HAPEM6 commuting data were derived from the 2000 Census, as provided by the U.S. Department of Transportation (DOT) Bureau of Transportation Statistics—specifically, in their Census Transportation Planning Package (CTPP).[9] For the HAPEM7 commuting file, we identified equivalent CTPP data for the year 2010 from the DOT's Federal Highway Administration (FHWA)—specifically, their Microsoft® Access™-based CTPP 2006–2010 file,[10] based on 2006–2010 five-year summary data from the Census' American Community Survey (ACS) and commissioned by the American Association of State Highway and Transportation Officials (AASHTO). This CTPP Access™ database contains estimates of the total number of workers commuting within or between tracts.

[9] As of January 21, 2015, the data used to derive HAPEM6's commuting file are available at http://www.transtats.bts.gov/Tables.asp?DB_ID=630.

[10] As of January 21, 2015, the data used to derive HAPEM7's commuting file are available at http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm.

To produce the commuter fractions in the HAPEM7 commuting file, we divided the number of workers in each home-tract/work-tract pair by the total number of workers residing in the home tract. To produce the distance between each home-tract/work-tract pair, we used the 2010 Census' coordinates of tract centroids, available from the 2010 Census Gazetteer and downloaded for a previous work assignment.[6] More specifically, we used the following Great Circle calculation between the home-tract coordinates (i.e., Lat_1 and Lon_1) and the work-tract coordinates (i.e., Lat_2 and Lon_2):

$$Distance = Radius_{Earth} * \ ArcCos\begin{pmatrix} Cos(Lat_1) * Cos(Lon_1) * Cos(Lat_2) * Cos(Lon_2) + \\ Cos(Lat_1) * Sin(Lon_1) * Cos(Lat_2) * Sin(Lon_2) + \\ Sin(Lat_1) * Sin(Lat_2) \end{pmatrix}$$

A small number of tracts in nearly every state, the District of Columbia, and Puerto Rico were absent from the CTPP data (totaling 1,067 tracts, or 1 percent of all tracts). HAPEM will model each missing tract as if all its employed residents work within the tract (i.e., for the purposes of HAPEM modeling, they essentially do not commute), so we did not insert any data for these missing tracts. Additionally, the CTPP contained no data on all 32 tracts in the U.S. Virgin Islands. To prevent the Islands from being conspicuously missing from the commuting file, we inserted one record for each Island tract, where work tract equals home tract and the commute is 0 km, which is how HAPEM would model them if they remained missing from the file.

2.2.1. Quality Assurance

We thoroughly checked the accuracy of each of the queries used in Microsoft® Access™ to generate the HAPEM7 commuting file. We confirmed the numbers of home and work tracts at various stages of the analysis. We also spot-checked the commuting fractions, including a full check that the cumulative commuting fraction equaled 1 for each home tract (with an allowance for very small rounding errors). We used mapping software to spot-check commuting distances, and as a bounding check we compared the distribution of commuting distances in HAPEM7 against that of HAPEM6 and found similar distributions.

2.2.2. Content and Format

The HAPEM7 commuting data are contained in a fixed-with, space-delimited text file with characteristics shown in Table 2. The file contains five columns (the first being empty) and a total of 4,156,458 rows of data with no header rows. Each data row corresponds to a unique home-tract/work-tract pair, where the second and third fields respectively contain the home and work tract identifiers using Census FIPS coding,[8] and the fourth and fifth fields respectively contain the commuting distance (in kilometers) and the fraction of workers commuting between the home and work tracts. Distance values are presented to no more than two decimal places (i.e., hundredths of kilometers, which is tens of meters), while commuting fractions are presented to no more than eight decimal places.

Table 2. Characteristics of the HAPEM7 Commuting File

Field Number	Description	Character Start Position on Data Row	Character Length on Data Row[a]
1	Leading space in file	1	1
2	Full Census FIPS Code for Home Tract	2	11
3	Full Census FIPS Code for Work Tract	14	11
4	Distance in km between home and work tract	26	8
5	Fraction of workers in the home tract commuting to the work tract	34	10

[a]Any unused character space after a number and/or between fields consists of blank spaces.

Figure 3 and Figure 4 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 commuting files.

```
 01001020100 01001020100    0.00 0.02896871
 01001020100 01001020200    1.40 0.06952491
 01001020100 01001020300    2.80 0.04866744
 01001020100 01001020400    3.90 0.04287370
 01001020100 01001020500    6.10 0.07647740
 01001020100 01001020600    3.50 0.08342990
 01001020100 01001020700    5.50 0.09965237
 01001020100 01001020800    3.60 0.01853998
 01001020100 01001020900   19.30 0.00695249
 01001020100 01021060101   41.30 0.00347625
```

Figure 3. Excerpt from the HAPEM6 Commuting File ("comm2000.txt")

```
 01001020100 01051031100 13.13   0.01156069
 01001020100 01101000900 15.91   0.10982659
 01001020100 01101005902 23.67   0.01156069
 01001020100 01101005901 33.87   0.01156069
 01001020100 01101005406 33.9    0.01156069
 01001020100 01101003100 25.88   0.01156069
 01001020100 01101002900 28.94   0.01156069
 01001020100 01101001600 23.35   0.01156069
 01001020100 01101000700 21.28   0.01156069
 01001020100 01101005409 29.97   0.01734104
```

Figure 4. Excerpt from the HAPEM7 Commuting File ("commute_flow_v7.txt")

In the HAPEM7 commuting file, on average there are 56 work tracts per home tract (not counting records where the home and work tracts are the same), up to a maximum of 296 work tracts. In 583 home tracts (which is less than 1 percent of home tracts), all workers work within their home tract, likely corresponding to less than 3 percent of U.S. workers.

Commuting distances greater than 120 km are assumed in HAPEM to be very atypical for a daily commuter, and thus HAPEM ignores these longer commutes in constructing the commute distance distributions for each tract (see the HAPEM User's Guide[5]). Most home tracts have at least one work tract that is more than 120 km away; that is, in approximately 70 percent of home tracts there is at least one person residing there who commutes farther than 120 km. However, this affects only 3 percent of home-tract/work-tract pairs, and likely affects less than 3 percent of U.S. workers. Ignoring these records with commuting distances greater than 120 km, the average tract-to-tract distance is 22 km (weighting all tract pairs equally, not by numbers of people performing those commutes).

2.3. Commuting-time File – "commute_time_v7.txt"

Whereas the HAPEM commuting file ("commute_flow_v7.txt" see Section 2.2) contains information on the frequency distribution of commuting distances for workers in a given home tract, the HAPEM commuting-time file ("commtime_new.txt" in HAPEM6; "commute_time_v7.txt" in HAPEM7) contains information on the method of commuting (public versus private transit) and the average commuting time. These commuting-time data are provided for all the tracts contained in the HAPEM population file, though no commuting data were available for the U.S. Virgin Islands, as discussed below.

The HAPEM6 commuting-time data were derived from the 2000 Census—specifically, from their Summary File 3, Table P32: "Travel Time to Work by Means of Transportation for Workers 16+ Years who Did Not Work at Home".[11] For the HAPEM7 commuting-time file, we identified equivalent data for the year 2010 from the 2006–2010 five-year summary data from the ACS, as detailed in the following paragraphs.[12]

ACS Table B08301: "Means of Transportation to Work for Workers 16+ Years" contains the numbers of people commuting to work by various means of transit (i.e., "Car, truck, or van: Drove alone", "Car, truck, or van: Carpooled", "Public Transportation: Bus or trolley bus", and so on). We used this table to derive the proportion of commuters traveling by public transit (i.e., bus, trolley bus, streetcar, trolley car, subway, elevated train, railroad, and ferryboat) and the proportion of commuters traveling by private transit (i.e., car, truck, van, taxicab, motorcycle, bicycle, any other non-public means except walking). We excluded from these calculations people working from home (i.e., workers not commuting). We also excluded people walking to work, which are cases where we assume people work within their home tract and thus are not considered commuters for the purposes of HAPEM exposure modeling. As such, the fractions of workers commuting by public and private transit sum to 1, except for a relatively small number of tracts (approximately 858, or 1 percent of all tracts) where nobody reported relevant commuting activity.

[11] We were unable to confirm how the Census Table P32 was used to produce the HAPEM6 commuting-time file. The fractions of workers commuting by public versus private transit can be easily derived from Table P32, though the average commuting time requires processing because Table P32 only specifies the number of people whose commutes are 1-30 minutes, 30-44 minutes, 45-59 minutes, and 60+ minutes.

[12] The ACS has replaced the 2000 long-form Census survey (i.e., Summary File 3).

ACS Table C08134: “Means of Transportation to Work by Travel Time to Work for Workers 16+ Years who Did Not Work at Home” contains the numbers of people commuting to work, irrespective of commuting time as well as for various bins of commuting time (i.e., 1–10 minutes, 10–14 minutes,…,60+ minutes), and irrespective of means of transit as well for specific means of transit in broader groups than in Table B08301 (i.e., “Car, truck, or van: Drove alone”, “Car, truck, or van: Carpooled”, “Public transportation (excluding taxicab)”, and “Taxicab, motorcycle, bicycle, walked, or other means”). We used the population counts by means of transit (irrespective of commuting time) in combination with ACS Table C08136 to derive commuting times, as discussed in the following paragraph.

ACS Table C08136: “Aggregate Travel Time to Work (in Minutes) by Means of Transportation to Work for Workers 16+ Years who Did Not Work at Home” contains travel times to work by the same transit means as in Table C08134, summed across all people who use those means. We divided these aggregate travel times by the corresponding population counts from Table C08134, resulting in average per-person travel times to work, by public transit and by private transit; multiplying by two equals the round-trip times used in the HAPEM7 commuting-time file. Commuting times related to public transit include time spent waiting at a bus or train stop, and commuting times (and population counts from Table C08134) related to private transit include walking commuters; these times are included in our calculations because they cannot be disaggregated from the total commuting time. If the data derived from Table B08301 (used for the proportions of workers commuting by public and private means) indicated that a tract had no commuters using public means, then we set commuting times to 0 for public means; similarly, we set private commuting times to 0 if there were no private commuters, and we set both public and private commuting times to 0 if there were no commuters at all.

Commuting data for the U.S. Virgin Islands were not available from the ACS, so we defaulted all data in the commuting-time file related to the Islands such that all workers work at home (i.e., commute neither by public nor private transit, with commuting times equal to 0). This is consistent with how we approached Island data in the commuting file (see Section 2.2).

Aggregate commuting-time data were also unavailable from Table C08136 (either missing entirely from the table, or present in the table but with flags indicating a lack of reliable data) for 37 percent of tracts in the other U.S. areas. We used county-average aggregate times for 86 percent of these missing tracts (i.e., for 32 percent of all tracts outside of the U.S. Virgin Islands) and state averages for the remaining 14 percent of missing tracts (i.e., for 5 percent of all tracts outside of the U.S. Virgin Islands). We divided those county and state aggregate times by the county and state commuter population counts to produce average, per-person, one-way commuting times, and multiplied by two to obtain round-trip times. These county and state averages were stratified by public and private means. We made these data substitutes for all cases of data missing from Table C08136, including cases where other ACS tables indicated that the tract had no commuters.

2.3.1. Quality Assurance

We thoroughly checked a SAS® data set that contained all the ACS variables used to produce the commuting-time file, the calculated county- and state-average data, and all intermediate calculated variables. We checked for missing values as well as minimum and maximum values for each variable using the "PROC MEANS" SAS® function. We spot-checked these calculations for two tracts.

We discovered and remedied two errors. In one error, we were first calculating the per-person commuting time of each individual means of transit (dividing the commuting times from ACS Table C08136 by the population counts in ACS Table C08136) and then averaging together all those per-person public-transit times (into a final per-person public-transit commuting time) and likewise for the private-transit times. That calculation erroneously gave equal weight to the commuting time of each specific means of transit. The corrected calculation first sums the aggregate commuting times (from ACS Table C08136) into an aggregate public-transit time and an aggregate private-transit time, and likewise sums the population counts taking public and private transit (from ACS Table C08134), and then divides the aggregate times by the aggregate population counts. In the second error, the proportion of commuters did not exclude walking commuters.

We also checked the commuting-time file produced by the SAS® script, ensuring that the public and private commuting fractions summed to 1 for every record. As a bounding check, we checked the distributions of public and private commuting times for reasonableness and we compared these distributions between the HAPEM7 and HAPEM6 files. We manually calculated the commuting times for the same two tracts that we spot-checked within the SAS® code, ensuring that the times matched.

2.3.2. Content and Format

The HAPEM7 commuting-time data are contained in a tab-delimited text file with characteristics shown in Table 3. The file contains five columns and a total of 74,034 rows of data with no header rows. Each row corresponds to a tract, where the first field contains the tract identifier using Census FIPS coding,[8] the second and third fields respectively contain the proportion of commuters who travel by public transit (excluding taxicabs) and by private transit (including taxicabs), and the fourth and fifth fields respectively contain the average round-trip times (in minutes) commuting to work by public transit and by private transit. All values in fields 2–5 are displayed to four decimal places.

Table 3. Characteristics of the HAPEM7 Commuting-time File

Field Number	Description
1	Full Census FIPS Code for Home Tract
2	Proportion of workers commuting outside of the home by public transit
3	Proportion of workers commuting outside of the home by private transit
4	Average round-trip commuting time for workers commuting outside of the home by public transit
5	Average round-trip commuting time for workers commuting outside of the home by private transit

Note: The position where table values begin and the number of characters per value are not relevant in a tab-delimited format.

Figure 5 and Figure 6 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 commuting-time files.

```
01001020100    0.0000    1.0000      0.0000    55.3124
01001020200    0.0000    1.0000      0.0000    52.9643
01001020300    0.0000    1.0000      0.0000    51.3188
01001020400    0.0052    0.9948    180.0000    47.4427
01001020500    0.0000    1.0000      0.0000    51.7882
01001020600    0.0156    0.9844     30.0000    45.1287
01001020700    0.0000    1.0000      0.0000    48.5694
01001020800    0.0000    1.0000      0.0000    61.0869
01001020900    0.0000    1.0000      0.0000    75.0188
01001021000    0.0000    1.0000      0.0000    87.1034
```

Figure 5. Excerpt from the HAPEM6 Commuting-time File ("commtime_new.txt")

```
01001020100     0.0472    0.9528    14.4946    35.9467
01001020200     0.0000    1.0000    0.0000     51.6934
01001020300     0.0092    0.9908    14.4946    35.9467
01001020400     0.0000    1.0000    0.0000     40.9754
01001020500     0.0000    1.0000    0.0000     38.9659
01001020600     0.0000    1.0000    0.0000     47.1820
01001020700     0.0000    1.0000    0.0000     50.3055
01001020801     0.0000    1.0000    0.0000     35.9467
01001020802     0.0000    1.0000    0.0000     54.3099
01001020900     0.0000    1.0000    0.0000     67.2094
```

Figure 6. Excerpt from the HAPEM7 Commuting-time File ("commute_time_v7.txt")

Across all tracts (except the U.S. Virgin Islands), the average proportion of commuters using private transit was 93 percent and the conditional-average round-trip private-transit commute was 36 minutes (72 minutes for public transit) (conditional averaging considers only non-zero values). This statistic treats every tract equally, rather than weighting by commuting population; it includes county and state averages where we used them. The longest round-trip commuting times in the data set are 143 minutes for private transit and 367 minutes for public transit.

2.4. Commuting-fraction File – "commute_fraction_v7.txt"

The HAPEM commuting-fraction file ("commfrac.txt" in HAPEM6; "commute_fraction_v7.txt" in HAPEM7) contains the fraction of workers (currently in the labor force and employed) in each tract who do commute and the fraction who do not commute, stratified by each age group. Workers who walk to work are not included as commuters for HAPEM7.

The HAPEM6 commuting-fraction data were derived from the 2000 Census—specifically, Table P31: "Travel Time to Work for Workers 16+ Years" and Table PCT35: "Sex by Age by Employment Status for the Population 16 Years and Over". HAPEM6 data did not consider those in the Armed Forces as part of the employed work force and did include walking as a means of commuting. For the HAPEM7 commuting-fraction file, we identified equivalent data for the year 2010 from the 2006–2010 five-year summary data from the ACS,[12] including Armed Forces members but not including those walking to work, as detailed in the following paragraphs.

ACS Table B23001: "Sex by Age by Employment Status for the Population 16 Years and Over" contains the numbers of people in the labor force and not in the labor force, including those in the labor force but unemployed, by gender and age group. We used this table to derive the number of people per HAPEM age group who are in the labor force as Armed Forces workers or employed in civilian jobs, summed across the two genders. We fit the ACS age groups to the HAPEM age groups using the same apportionment methods discussed for the population file in Section 2.1.

ACS Table B08101: "Means of Transportation to Work by Age for Workers 16+ Years" contains the numbers of people per age group commuting to work by various means of transit (i.e., "Car, truck, or van: Drove alone", "Car, truck, or van: Carpooled", "Public Transportation (excluding taxicab)", and so on). We used this table to derive the numbers of workers who commuted by means other than walking, by age group. As we did in calculating the proportion of workers commuting by public and private transit (see Section 2.3), here we excluded people walking to work because they likely work within their home tract and for simplicity we consider them not to be commuters in HAPEM. We fit the ACS age groups to the HAPEM age groups using the same apportionment methods discussed for the population file in Section 2.1.

Using the two ACS tables discussed above, for each tract and HAPEM age group we calculated the fraction of workers commuting as (number of people aged 16+ years who commute to work other than by walking) ÷ (number of people aged 16+ employed in the labor force). The fraction of workers not commuting is 1 minus the above fraction.

Commuting data for the U.S. Virgin Islands were not available from the ACS, so we set data in the commuting-fraction file such that all workers in the Islands work at home (i.e., did not commute). This is consistent with how we treated Island data in the commuting and commuting-time files (see Sections 2.2 and 2.3, respectively).

2.4.1. Quality Assurance

We thoroughly checked a SAS® data set that contained all the ACS variables used to produce the commuting-fraction file and all intermediate calculated variables. We checked for missing values as well as minimum and maximum values for each variable using the PROC MEANS SAS® function. We spot-checked these calculations for two tracts.

We also checked the commuting-fraction file produced by the SAS® script, ensuring that the fractions of workers in each age group commuting and not commuting summed to 1 for every record. We discovered and remedied one type of error, where some values intended to be "1.0000" were displayed as ".". We compared the HAPEM7 and HAPEM6 files to ensure proper layout. We manually calculated the commuting fractions for three tracts.

2.4.2. Content and Format

The HAPEM7 commuting-fraction data are contained in a tab-delimited text file with characteristics shown in Table 4. The file contains five columns and a total of 74,034 rows of data with no header rows. Each row corresponds to a tract, where the first field contains the tract identifier using Census FIPS coding,[8] the second and third fields respectively contain the fraction of workers (currently in the labor force and employed) aged 0–1 years who do not commute and who do commute, and the remaining fields show the same data for each of the other five HAPEM age groups. All values in fields 2–13 are displayed to four decimal places. Nobody younger than 16 years is considered employed and a commuter, so all values for “does not commute to work” are 1 and all values for “commutes to work” are 0 for the first three HAPEM age groups.

Table 4. Characteristics of the HAPEM7 Commuting-fraction File

Field Number	Description
1	Full Census FIPS Code for Home Tract
2	Proportion of age group 1 (ages 0–1) that does not commute to work
3	Proportion of age group 1 (ages 0–1) that commutes to work
4	Proportion of age group 2 (ages 2–4) that does not commute to work
5	Proportion of age group 2 (ages 2–4) that commutes to work
6	Proportion of age group 3 (ages 3–15) that does not commute to work
7	Proportion of age group 3 (ages 3–15) that commutes to work
8	Proportion of age group 4 (ages 16–17) that does not commute to work
9	Proportion of age group 4 (ages 16–17) that commutes to work
10	Proportion of age group 5 (ages 18–64) that does not commute to work
11	Proportion of age group 5 (ages 18–64) that commutes to work
12	Proportion of age group 6 (ages 65 and older) that does not commute to work
13	Proportion of age group 6 (ages 65 and older) that commutes to work

Note: The position where table values begin and the number of characters per value are not relevant in a tab-delimited format.

Figure 7 and Figure 8 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 commuting-fraction files.

```
01001020100 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5269 0.4731 0.3939 0.6061 0.7357 0.2643
01001020200 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5066 0.4934 0.4604 0.5396 0.8221 0.1779
01001020300 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.3787 0.6213 0.2992 0.7008 0.9644 0.0356
01001020400 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5981 0.4019 0.2902 0.7098 0.8606 0.1394
01001020500 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5785 0.4215 0.2906 0.7094 0.9011 0.0989
01001020600 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.4548 0.5452 0.2764 0.7236 0.9001 0.0999
01001020700 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.7413 0.2587 0.2811 0.7189 0.6625 0.3375
01001020800 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6627 0.3373 0.2903 0.7097 0.8778 0.1222
01001020900 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.7037 0.2963 0.3835 0.6165 0.9243 0.0757
01001021000 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.7633 0.2367 0.3976 0.6024 0.9028 0.0972
```

Figure 7. Excerpt from the HAPEM6 Commuting-fraction File (“commfrac.txt”)

```
01001020100   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5500 0.4500 0.2742 0.7258 0.7883 0.2117
01001020200   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.3671 0.6329 0.8627 0.1373
01001020300   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5126 0.4874 0.3107 0.6893 0.8809 0.1191
01001020400   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6906 0.3094 0.2840 0.7160 0.8335 0.1665
01001020500   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6108 0.3892 0.2375 0.7625 0.8590 0.1410
01001020600   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.6727 0.3273 0.3108 0.6892 0.8679 0.1321
01001020700   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.8545 0.1455 0.2739 0.7261 0.9306 0.0694
01001020801   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.7731 0.2269 0.2645 0.7355 0.8790 0.1210
01001020802   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.5788 0.4212 0.3168 0.6832 0.7830 0.2170
01001020900   1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.9576 0.0424 0.4008 0.5992 0.9027 0.0973
```

Figure 8. Excerpt from the HAPEM7 Commuting-fraction File (“commute_fraction_v7.txt”)

2.5. Distance-to-road File – “proximity_road_v7.txt”

The HAPEM distance-to-road file (“distprob.txt” in HAPEM6; “proximity_road_v7.txt” in HAPEM7) contains information on the fraction of a tract’s residents that live within each of three categories of distance from a major roadway, by age group. These distances are 0–75 m, greater than 75 m up to 200 m, and greater than 200 m. The file contains these data for all tracts in the HAPEM7 population file, and the HAPEM6 and HAPEM7 versions were produced using very similar methods. We conducted the proximity assessment at the level of Census blocks and stratified by age and gender, and then we aggregated the block-level results up to the tract level and stratified only by age group.

We used block-level geographies from the 2010 Census Gazetteer and downloaded for a previous work assignment.[6] We used block-level population data from GeoLytics for the 50 states and District of Columbia[13] and from the 2010 Census for Puerto Rico and the U.S. Virgin Islands.

We used roadway location data from the 2013 Census TIGER/Line transportation U.S. roadway geographic file.[14] We assumed first that every person in every tract resides within 150 m of a roadway that can be classified as at least a local road (if not a highway or other significant roadway). We assumed that some transportation features such as logging roads and ferries were unlikely to be located within 150 m of a residence. We also assumed that certain features such as traffic circles, cul de sacs, local or neighborhood roads, rural roads, and city streets did not meet the definition of a “major roadway” for the purposes of evaluating enhanced pollutant exposure to people living near heavy-use roads. Table 5 contains information on the features we assessed for roadway proximity.

Table 5. Types of “Major” Roads Included in the Roadway-proximity Assessment

Roadway Type	Definition
Primary Road	Generally divided, limited-access highways within the interstate highway system or under state management, and distinguished by the presence of interchanges. Accessible by ramps and may include some toll highways.
Ramp	Allows controlled access from adjacent roads onto a limited-access highway, often in the form of a cloverleaf interchange.
Secondary Road	Main arteries, usually in the U.S., state, or county highway systems. Have one or more lanes of traffic in each direction, may or may not be divided, and usually have at-grade intersections with many other roads and driveways.
Service Drive (usually along a limited-access highway)	Usually parallels a limited access highway, provides access to structures along the highway.

We used ESRI® ArcInfo™ software utilizing ArcGIS™ version 10.2 to perform the roadway-proximity geospatial analyses. Due to the size of the roadway and Census-block geography files, we conducted most of the processing at a state level and in some cases at the county level. We describe below each general step of the processing.

[13] As of January 29, 2015, bundles of Census data and proprietary demographic projections are available from GeoLytics at http://www.geolytics.com/.

[14] As of January 29, 2015, the U.S. Census TIGER/Line data available at https://www.census.gov/geo/maps-data/data/tiger-line.html.

1. Populations are not usually evenly distributed within Census blocks, so we assumed that all people live within 150 m of a roadway designated as “local” or greater. We created a 150-m buffer around all such roadways. These buffers defined where everyone lives, and we hereafter refer to them as “populated areas”. We assumed that populations within each block are evenly distributed within these populated areas. We clipped the populated areas at the boundaries of blocks, such that no populated area crossed a block boundary.
2. We created 75- and 200-m buffers around all “major” roadways. We clipped these buffers at the boundaries of blocks, such that no major-roadway buffer crossed a block boundary.
3. For each Census block, we calculated the fraction of the populated area that was within the 75-m major-roadway buffer, the fraction that was within the 200-m major-roadway buffer (subtracting the 75-m portion results in a 75-to-200-m buffer), and the fraction that was outside the 200-m major-roadway buffer. For recordkeeping purposes only, we also calculated the fractions of total block area that fell within each of the major-roadway buffers.
4. We exported the above data to Microsoft® Access™ and appended block population data.
5. For each block and major-roadway buffer, we multiplied the fraction from Step 3 above by the block population count per gender and age group. These are the numbers of people residing within 75 m, beyond 75 m up to 200 m, and beyond 200 m of a major roadway, at the block level and stratified by gender and age.
6. We aggregated the data from Step 5 above to the tract level and summed together the male and female data. We then divided the population counts within the major-roadway buffers by the total tract population, stratified by each of the six HAPEM age groups.

The fractions of people living within the major-roadway buffers are all 0 when tract population is 0.

2.5.1. Quality Assurance

We implemented several layers of QA with several staff members at several stages of the processing. We used the majority of QA time for calculations performed in Steps 4–6 above (i.e., the final steps of processing population data and aggregating to the tract level). We spot-checked that our block-level population data were complete, and discovered that we had some corrupted data from Geolytics (we received new data as a result).

We spot-checked that our processed block-level results represented all blocks nationwide. We discovered a very small percentage of blocks (e.g., 3 percent in Alabama) had no roadways considered “local” or greater, and these areas were usually close to water or mountainous terrain—this was an observation and not a problem. At the tract level, approximately 0.01 percent of all assessed tracts had no people living within 150 m of a roadway “local” or greater, and an additional 0.4 percent of tracts had some people living outside the 150-m buffer (up to 2 percent of tracts in Alaska); again, these are observations and not errors.

We checked that the major-roadway buffer fractions from Step 5 summed to 1 for every block (and in Step 6 summed to 1 for every tract). In this process, we implemented post-processing algorithms to remove rounding errors so that fractions summed to 1 where appropriate (when processed at 4 decimal places). We corrected an error where the populated-area value was not always consistent across different data tables used during processing. We also discovered very small negative values in the areas beyond 200 m of a major roadway—this was an artifact of some of our processing techniques, and we forced the small negative values to be 0.

We spot-checked that the processed tract-population data summed to the correct state-total populations and summed correctly across age groups. We also noted that we should not always expect the fraction of tract area within the individual major-roadway buffers to equal the fraction of tract population within the buffers—this is due to the fact that we performed the assessment at the block level and then aggregated to the tract level, where each block has a unique population density that makes aggregated populations unequal to aggregated areas.

2.5.2. Content and Format

The HAPEM7 distance-to-road data are contained in a tab-delimited text file with characteristics shown in Table 6. The file contains 22 columns and a total of 74,034 rows of data with no header rows. Each row corresponds to a tract, where the first field contains the tract identifier using Census FIPS coding,[8] fields 2–4 respectively contain the fractions of tract area 0–75 m from a major roadway, 75–200 m, and beyond 200 m, and the remaining fields show similar data for the fractions of people in each HAPEM age group who reside within those three major-roadway buffers. All values in fields 2–22 are displayed to four decimal places. The population fractions in tracts with 0 residents are shown as 0 values.

Table 6. Characteristics of the HAPEM7 Distance-to-road File

Field Number	Description
1	Full Census FIPS Code for Home Tract
2	Proportion of tract area located 0–75 m from major roadway
3	Proportion of tract area located beyond 75 m of major roadway, up to 200 m
4	Proportion of tract area located beyond 200 m of major roadway
5	Proportion of age group 1 (ages 0–1) residing 0–75 m from major roadway
6	Proportion of age group 1 (ages 0–1) residing beyond 75 m of major roadway, up to 200 m
7	Proportion of age group 1 (ages 0–1) residing beyond 200 m of major roadway
8	Proportion of age group 2 (ages 2–4) residing 0–75 m from major roadway
9	Proportion of age group 2 (ages 2–4) residing beyond 75 m of major roadway, up to 200 m
10	Proportion of age group 2 (ages 2–4) residing beyond 200 m of major roadway
11	Proportion of age group 3 (ages 3–15) residing 0–75 m from major roadway
12	Proportion of age group 3 (ages 3–15) residing beyond 75 m of major roadway, up to 200 m
13	Proportion of age group 3 (ages 3–15) residing beyond 200 m of major roadway
14	Proportion of age group 4 (ages 16–17) residing 0–75 m from major roadway
15	Proportion of age group 4 (ages 16–17) residing beyond 75 m of major roadway, up to 200 m
16	Proportion of age group 4 (ages 16–17) residing beyond 200 m of major roadway
17	Proportion of age group 5 (ages 18–64) residing 0–75 m from major roadway
18	Proportion of age group 5 (ages 18–64) residing beyond 75 m of major roadway, up to 200 m
19	Proportion of age group 5 (ages 18–64) residing beyond 200 m of major roadway
20	Proportion of age group 6 (ages 65 and older) residing 0–75 m from major roadway
21	Proportion of age group 6 (ages 65 and older) residing beyond 75 m of major roadway, up to 200 m
22	Proportion of age group 6 (ages 65 and older) residing beyond 200 m of major roadway

Note: The position where table values begin and the number of characters per value are not relevant in a tab-delimited format.

We have retained the block-level part of the processing for any potential future use, though it is not used in HAPEM. These block-level results identify each block using the full Census identifier, then provide the block’s total population, total area, fractions of areas within the three major-roadway distance buffers, and numbers of people residing within those buffers, the latter stratified by age and gender.

Figure 9 and Figure 10 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 distance-to-road files.

```
 01001020100  0.3268 0.2578 0.4154 0.6570 0.3318 0.0112 0.7354 0.2496 0.0150 0.6920 0.2961 0.0119 0.7101 0.2791 0.0108 0.6966
        0.2893 0.0141 0.6536 0.3305 0.0159

01001020200   0.6165 0.2142 0.1693 0.7997 0.2003 0.0000 0.7731 0.2269 0.0000 0.8016 0.1984 0.0000 0.8392 0.1608 0.0000 0.8114
        0.1886 0.0000 0.8255 0.1745 0.0000

01001020300   0.5894 0.2725 0.1381 0.7724 0.2226 0.0050 0.7500 0.2437 0.0063 0.7327 0.2612 0.0061 0.7387 0.2547 0.0066 0.7540
        0.2389 0.0071 0.7725 0.2225 0.0050

01001020400   0.5822 0.2046 0.2132 0.8608 0.1334 0.0058 0.8584 0.1360 0.0056 0.8529 0.1412 0.0059 0.8116 0.1824 0.0060 0.8472
        0.1479 0.0049 0.8716 0.1242 0.0042

01001020500   0.4175 0.2381 0.3444 0.6969 0.2515 0.0516 0.7122 0.2375 0.0503 0.7311 0.2251 0.0438 0.7395 0.2188 0.0417 0.7227
        0.2326 0.0447 0.7036 0.2415 0.0549

01001020600   0.5169 0.2952 0.1879 0.7351 0.2496 0.0153 0.7545 0.2289 0.0166 0.7337 0.2486 0.0177 0.7091 0.2736 0.0173 0.7239
        0.2551 0.0210 0.7283 0.2518 0.0199

01001020700   0.2132 0.1680 0.6188 0.8004 0.1851 0.0145 0.7538 0.2241 0.0221 0.7653 0.2290 0.0057 0.7510 0.2340 0.0150 0.7703
        0.2201 0.0096 0.7529 0.2401 0.0070

01001020800   0.1574 0.1451 0.6975 0.5187 0.3968 0.0845 0.5197 0.3904 0.0899 0.5385 0.3860 0.0755 0.5344 0.3919 0.0737 0.5279
        0.3934 0.0787 0.5011 0.4131 0.0858

01001020900   0.1195 0.1210 0.7595 0.4542 0.4101 0.1357 0.4430 0.4336 0.1234 0.4169 0.4123 0.1708 0.4469 0.4053 0.1478 0.4218
        0.4060 0.1722 0.4330 0.4073 0.1597

01001021000   0.1354 0.1252 0.7394 0.4855 0.4396 0.0749 0.4864 0.4420 0.0716 0.4833 0.4390 0.0777 0.4907 0.4493 0.0600 0.4894
        0.4367 0.0739 0.5102 0.4086 0.0812
```

Note: Contents wrap around due to space constrictions in this figure.

Figure 9. Excerpt from the HAPEM6 Distance-to-road File ("distprob.txt")

```
01001020100   0.0643 0.0970 0.8387 0.0582 0.0882 0.8535 0.0797 0.1179 0.8023 0.0598 0.0913 0.8489 0.0488 0.0794 0.8717 0.0634
        0.0977 0.8389 0.0767 0.1022 0.8211

01001020200   0.0326 0.0622 0.9052 0.0031 0.0158 0.9811 0.0284 0.0627 0.9089 0.0273 0.0512 0.9215 0.0202 0.0435 0.9363 0.0287
        0.0655 0.9058 0.0803 0.0789 0.8408

01001020300   0.0586 0.1039 0.8375 0.0484 0.0824 0.8692 0.0528 0.0818 0.8654 0.0551 0.0910 0.8539 0.0567 0.0925 0.8508 0.0586
        0.1029 0.8384 0.0671 0.1395 0.7934

01001020400   0.1293 0.2003 0.6704 0.1742 0.2039 0.6219 0.1382 0.2498 0.6121 0.1270 0.2046 0.6684 0.1247 0.1873 0.6880 0.1331
        0.1992 0.6678 0.1156 0.1942 0.6902

01001020500   0.0214 0.0426 0.9361 0.0189 0.0380 0.9431 0.0215 0.0407 0.9378 0.0196 0.0399 0.9405 0.0185 0.0386 0.9429 0.0213
        0.0430 0.9357 0.0265 0.0484 0.9252

01001020600   0.1399 0.2283 0.6319 0.1447 0.2417 0.6136 0.1346 0.2135 0.6518 0.1386 0.2257 0.6358 0.1244 0.2010 0.6746 0.1409
        0.2319 0.6272 0.1419 0.2231 0.6350

01001020700   0.0898 0.1699 0.7404 0.0833 0.1573 0.7594 0.0796 0.1509 0.7694 0.0825 0.1540 0.7634 0.0877 0.1669 0.7454 0.0945
        0.1810 0.7245 0.0788 0.1400 0.7812

01001020801   0.0609 0.0826 0.8565 0.0586 0.0823 0.8591 0.0661 0.0774 0.8565 0.0499 0.0700 0.8801 0.0559 0.0762 0.8679 0.0610
        0.0824 0.8566 0.0775 0.1068 0.8156

01001020802   0.0597 0.0836 0.8567 0.0586 0.0909 0.8505 0.0681 0.0910 0.8409 0.0605 0.0867 0.8528 0.0568 0.0788 0.8644 0.0598
        0.0831 0.8571 0.0554 0.0782 0.8665

01001020900   0.1019 0.1052 0.7929 0.1195 0.1237 0.7567 0.0985 0.1091 0.7924 0.1032 0.1067 0.7901 0.0957 0.1121 0.7922 0.1019
        0.1041 0.7941 0.0991 0.1014 0.7995
```

Note: Contents wrap around due to space constrictions in this figure.

Figure 10. Excerpt from the HAPEM7 Distance-to-road File (“proximity_road_v7.txt”)

3. Updating Microenvironments and Exposure-factors Files ("factors_*_v7.txt")

HAPEM6 utilized 14 microenvironments (MEs), each corresponding to one of five broader MEs used for the activity cluster analysis (see Section 3), and each also designated as being related to commuting or not related to commuting. The HAPEM6 MEs are shown in Table 7.

Table 7. The HAPEM6 Microenvironments (MEs)

ME Number	ME Description	Broader ME (for clustering)	Commuting?
1	Residential	1 Indoors Residence	No
2	Residential Garage	3 Outdoors Near-roadway	No
3	School	2 Indoors Other	No
4	Hospital	2 Indoors Other	No
5	Office	2 Indoors Other	No
6	Public Access	2 Indoors Other	No
7	Bar/Restaurant	2 Indoors Other	No
8	Car/Truck	5 In-vehicle	Yes - Private
9	Public Transit	5 In-vehicle	Yes - Public
10	Air Travel	2 Indoors Other	No
11	Outdoors, Near Roadway	3 Outdoors Near-roadway	No
12	Outdoors, Service Station	3 Outdoors Near-roadway	No
13	Outdoors, Parking Garage	3 Outdoors Near-roadway	No
14	Outdoors, Other	4 Outdoors Other	No

For HAPEM7, we expanded the number of commuting-related MEs from two to six by

- disaggregating "Motorcycle/Bicycle" from the "Outdoors, Near Roadway" ME and assigning it to private commuting (though still being part of the broader "Outdoors Near-roadway" ME for the purposes of activity clustering);
- disaggregating "Ferryboat" from "Outdoors, Other" and assigning it to public commuting (still part of the broader "Outdoors Other" ME);
- disaggregating "Waiting Outdoors for Public Transit" from "Public Transit" and assigning it to public commuting (still part of the broader "Outdoors Near-roadway" ME); and,
- disaggregating "Waiting Indoors for Public Transit" from "Public Transit" and assigning it to public commuting (still part of the broader "Indoors Other" ME).

We created the new "Motorcycle/Bicycle" and "Ferryboat" MEs so that they could be assigned as commuting activities (they were part of non-commuting outdoor MEs in HAPEM6). This assumes that the majority of time spent in these MEs is due to commuting to or from work, at least for respondents in CHAD who commute and for days when the respondents work. Because the HAPEM7 public-transit commuting times include time spent waiting for transit to arrive (aggregated with times spent doing other commuting activities), we created the new "Waiting Indoors for Public Transit" and "Waiting Outdoors for Public Transit" commuting MEs (they were part of non-commuting MEs in HAPEM6).

Finally, we reordered the 18 HAPEM7 MEs so that the superset of in-vehicle and indoor MEs received consecutive numbering (a requirement in the HAPEM program code). The HAPEM7 MEs are shown in Table 8.

Table 8. The HAPEM7 Microenvironments (MEs)

ME Number	ME Description	Broader ME (for clustering)	Commuting?
1	Residential	1 Indoors Residence	No
2	School	2 Indoors Other	No
3	Hospital	2 Indoors Other	No
4	Office	2 Indoors Other	No
5	Public Access	2 Indoors Other	No
6	Bar/Restaurant	2 Indoors Other	No
7	Car/Truck	5 In-vehicle	Yes - Private
8	Public Transit	5 In-vehicle	Yes - Public
9	Air Travel	2 Indoors Other	No
10	Waiting Indoors for Public Transit	2 Indoors Other	Yes - Public
11	Waiting Outdoors for Public Transit	3 Outdoors Near-roadway	Yes - Public
12	Motorcycle/Bicycle	3 Outdoors Near-roadway	Yes - Private
13	Ferryboat	4 Outdoors Other	Yes - Public
14	Residential Garage	3 Outdoors Near-roadway	No
15	Outdoors, Near Roadway	3 Outdoors Near-roadway	No
16	Outdoors, Service Station	3 Outdoors Near-roadway	No
17	Outdoors, Parking Garage	3 Outdoors Near-roadway	No
18	Outdoors, Other	4 Outdoors Other	No

The HAPEM7 mapping of CHAD location codes to HAPEM7 MEs is provided in Table 9.

Table 9. Mapping of CHAD Location Codes to HAPEM7 Microenvironments (MEs)

CHAD	HAPEM7	
Location	ME	Broader ME (for clustering)
U Uncertain of correct code	-1 Unused	0 Unused
X No data	-1 Unused	0 Unused
30000 Residence, general	1 Residential	1 Indoors Residence
30010 Your residence	1 Residential	1 Indoors Residence
30020 Other residence	1 Residential	1 Indoors Residence
30100 Residence, indoor	1 Residential	1 Indoors Residence
30120 Your residence, indoor	1 Residential	1 Indoors Residence
30121 ..., kitchen	1 Residential	1 Indoors Residence
30122 ..., living room or family room	1 Residential	1 Indoors Residence
30123 ..., dining room	1 Residential	1 Indoors Residence
30124 ..., bathroom	1 Residential	1 Indoors Residence
30125 ..., bedroom	1 Residential	1 Indoors Residence
30126 ..., study or office	1 Residential	1 Indoors Residence
30127 ..., basement	1 Residential	1 Indoors Residence
30128 ..., utility or laundry room	1 Residential	1 Indoors Residence

Table 9. Mapping of CHAD Location Codes to HAPEM7 Microenvironments (MEs)

CHAD	HAPEM7	
Location	ME	Broader ME (for clustering)
30129 ..., other indoor	1 Residential	1 Indoors Residence
30130 Other residence, indoor	1 Residential	1 Indoors Residence
30131 ..., kitchen	1 Residential	1 Indoors Residence
30132 ..., living room or family room	1 Residential	1 Indoors Residence
30133 ..., dining room	1 Residential	1 Indoors Residence
30134 ..., bathroom	1 Residential	1 Indoors Residence
30135 ..., bedroom	1 Residential	1 Indoors Residence
30136 ..., study or office	1 Residential	1 Indoors Residence
30137 ..., basement	1 Residential	1 Indoors Residence
30138 ..., utility or laundry room	1 Residential	1 Indoors Residence
30139 ..., other indoor	1 Residential	1 Indoors Residence
30200 Residence, outdoor	18 Outdoors, Other	4 Outdoors Other
30210 Your residence, outdoor	18 Outdoors, Other	4 Outdoors Other
30211 ..., pool or spa	18 Outdoors, Other	4 Outdoors Other
30219 ..., other outdoor	18 Outdoors, Other	4 Outdoors Other
30220 Other residence, outdoor	18 Outdoors, Other	4 Outdoors Other
30221 ..., pool or spa	18 Outdoors, Other	4 Outdoors Other
30229 ..., other outdoor	18 Outdoors, Other	4 Outdoors Other
30300 Residential garage or carport	14 Residential Garage	3 Outdoors Near-roadway
30310 ..., indoor	14 Residential Garage	3 Outdoors Near-roadway
30320 ..., outdoor	17 Outdoors, Parking Garage	3 Outdoors Near-roadway
30330 Your garage or carport	14 Residential Garage	3 Outdoors Near-roadway
30331 ..., indoor	14 Residential Garage	3 Outdoors Near-roadway
30332 ..., outdoor	17 Outdoors, Parking Garage	3 Outdoors Near-roadway
30340 Other residential garage or carport	14 Residential Garage	3 Outdoors Near-roadway
30341 ..., indoor	14 Residential Garage	3 Outdoors Near-roadway
30342 ..., outdoor	17 Outdoors, Parking Garage	3 Outdoors Near-roadway
30400 Residence, none of the above	1 Residential	1 Indoors Residence
31000 Travel, general	7 Car/Truck	5 In-vehicle
31100 Motorized travel	7 Car/Truck	5 In-vehicle
31110 Car	7 Car/Truck	5 In-vehicle
31120 Truck	7 Car/Truck	5 In-vehicle
31121 Truck (pickup or van)	7 Car/Truck	5 In-vehicle
31122 Truck (not pickup or van)	7 Car/Truck	5 In-vehicle
31130 Motorcycle or moped	12 Motorcycle/Bicycle	3 Outdoors Near-roadway
31140 Bus	8 Public Transit	5 In-vehicle
31150 Train or subway	8 Public Transit	5 In-vehicle
31160 Airplane	9 Air Travel	2 Indoors Other
31170 Boat	18 Outdoors, Other	4 Outdoors Other
31171 Boat, motorized	13 Ferryboat	4 Outdoors Other

Table 9. Mapping of CHAD Location Codes to HAPEM7 Microenvironments (MEs)

CHAD	HAPEM7	
Location	ME	Broader ME (for clustering)
31172 Boat, other	18 Outdoors, Other	4 Outdoors Other
31200 Non-motorized travel	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
31210 Walk	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
31220 Bicycle or inline skates/skateboard	12 Motorcycle/Bicycle	3 Outdoors Near-roadway
31230 In stroller or carried by adult	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
31300 Waiting for travel	11 Waiting Outdoors for Public Transit	3 Outdoors Near-roadway
31310 ..., bus or train stop	11 Waiting Outdoors for Public Transit	3 Outdoors Near-roadway
31320 ..., indoors	10 Waiting Indoors for Public Transit	2 Indoors Other
31900 Travel, other	7 Car/Truck	5 In-vehicle
31910 ..., other vehicle	7 Car/Truck	5 In-vehicle
32000 Non-residence indoor, general	5 Public Access	2 Indoors Other
32100 Office building/ bank/ post office	4 Office	2 Indoors Other
32200 Industrial/ factory/ warehouse	4 Office	2 Indoors Other
32300 Grocery store/ convenience store	5 Public Access	2 Indoors Other
32400 Shopping mall/ non-grocery store	5 Public Access	2 Indoors Other
32500 Bar/ night club/ bowling alley	6 Bar/Restaurant	2 Indoors Other
32510 Bar or night club	6 Bar/Restaurant	2 Indoors Other
32520 Bowling alley	6 Bar/Restaurant	2 Indoors Other
32600 Repair shop	5 Public Access	2 Indoors Other
32610 Auto repair shop/ gas station	16 Outdoors, Service Station	3 Outdoors Near-roadway
32620 Other repair shop	5 Public Access	2 Indoors Other
32700 Indoor gym /health club	5 Public Access	2 Indoors Other
32800 Childcare facility	2 School	2 Indoors Other
32810 ..., house	1 Residential	1 Indoors Residence
32820 ..., commercial	2 School	2 Indoors Other
32900 Large public building	5 Public Access	2 Indoors Other
32910 Auditorium/ arena/ concert hall	5 Public Access	2 Indoors Other
32920 Library/courtroom/museum/theater	5 Public Access	2 Indoors Other
33100 Laundromat	5 Public Access	2 Indoors Other
33200 Hospital/ medical care facility	3 Hospital	2 Indoors Other
33300 Barber/ hair dresser/ beauty parlor	5 Public Access	2 Indoors Other
33400 At Work, no/moving among locations	4 Office	2 Indoors Other
33500 School	2 School	2 Indoors Other
33600 Restaurant	6 Bar/Restaurant	2 Indoors Other
33700 Church	5 Public Access	2 Indoors Other
33800 Hotel/ motel	5 Public Access	2 Indoors Other
33900 Dry cleaners	5 Public Access	2 Indoors Other
34100 Parking garage	17 Outdoors, Parking Garage	3 Outdoors Near-roadway

Table 9. Mapping of CHAD Location Codes to HAPEM7 Microenvironments (MEs)

CHAD	HAPEM7	
Location	ME	Broader ME (for clustering)
34200 Laboratory	3 Hospital	2 Indoors Other
34300 Indoor, none of the above	5 Public Access	2 Indoors Other
35000 Non-residence outdoor, general	18 Outdoors, Other	4 Outdoors Other
35100 Sidewalk, street, neighborhood	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
35110 Within 10 yards of street	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
35200 Public garage/ parking lot	17 Outdoors, Parking Garage	3 Outdoors Near-roadway
35210 ..., public garage	17 Outdoors, Parking Garage	3 Outdoors Near-roadway
35220 ..., parking lot	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
35300 Service station/ gas station	16 Outdoors, Service Station	3 Outdoors Near-roadway
35400 Construction site	15 Outdoors, Near Roadway	3 Outdoors Near-roadway
35500 Amusement park	18 Outdoors, Other	4 Outdoors Other
35600 School Grounds/Playgrounds	18 Outdoors, Other	4 Outdoors Other
35610 ..., school grounds	18 Outdoors, Other	4 Outdoors Other
35620 ..., playground	18 Outdoors, Other	4 Outdoors Other
35700 Stadium or amphitheater	18 Outdoors, Other	4 Outdoors Other
35800 Park/ golf course	18 Outdoors, Other	4 Outdoors Other
35810 Park	18 Outdoors, Other	4 Outdoors Other
35820 Golf course	18 Outdoors, Other	4 Outdoors Other
35900 Pool/ river/ lake	18 Outdoors, Other	4 Outdoors Other
36100 Restaurant/ picnic	18 Outdoors, Other	4 Outdoors Other
36200 Farm	18 Outdoors, Other	4 Outdoors Other
36300 Outdoor, none of the above	18 Outdoors, Other	4 Outdoors Other

Creating a new and reordered set of MEs for HAPEM7 necessitated revising the HAPEM factors files that are used to estimate ME concentrations from outdoor, ambient concentrations. We also corrected the mobile factors related to “Outdoors, Service Station” and “Outdoors, Parking Garage,” which were identical to “Outdoors, Other” in HAPEM6 but which we revised to be identical to “Outdoors, Near Roadway” in HAPEM7. These changes to the factors files are provided in Table 10.

Table 10. New and Revised Factors Assignments in HAPEM7

ME Number	ME Description	Mapping to Factors	Mapping to Mobile Factors
10	Waiting Indoors for Public Transit	Same as ME 5 “Public Access”	
11	Waiting Outdoors for Public Transit	Same as ME 15 “Outdoors, Near Roadway”	
12	Motorcycle/Bicycle	Same as ME 15 “Outdoors, Near Roadway”	
13	Ferryboat	Same as ME 18 “Outdoors, Other”	
16	Outdoors, Service Station	No change	Same as ME 15 “Outdoors, Near Roadway”
17	Outdoors, Parking Garage	No change	Same as ME 15 “Outdoors, Near Roadway”

Separate from these changes, we also corrected the PROX (i.e., roadway proximity) factors in the three factors files that apply to all source categories (i.e., the *factors* files; not the factors files that are specific to on-road mobile sources, which are the *mobiles* files). The *factors* files include PROX factors for four source categories, where source categories other than on-road mobile received PROX factors of 1 (i.e., no roadway proximity effect), and where the third source category is hard-coded to be on-road mobile sources with some PROX factors greater than 1 (indicating a roadway proximity effect). The *mobiles* files were then introduced to apply more-specific PROX factors for pollutants emitted by on-road mobile sources. The model programming was changed so that the modeler specifies which source category corresponded to on-road mobile and then the model would apply PROX factors from the *mobiles* file (overriding those in the *factors* file). However, the on-road mobile PROX factors in the *factors* files were never removed or set to 1 because these changes were not fully tested and had not, until now, been used for NATA. This override for on-road mobile PROX factors occurs correctly when the modeler sets the third source category to be on-road mobile, but if a model run has some other source category as the third one, it will be subjected to the on-road mobile PROX factors in from the *factors* files. To correct for this, we set all on-road mobile PROX factors to 1 in the *factors* files, so that no override is necessary and whichever source category the modeler sets as on-road mobile will receive the proper PROX factors from the *mobiles* files.

4. Updating Activity (“activity_CHAD_v7.txt”), Cluster (“activity_cluster_v7.txt”), and Cluster-transition (“activity_ClusterTransition_v7.txt”) Files

We updated the HAPEM activity file (“durhw.txt” in HAPEM6; “activity_CHAD_v7.txt” in HAPEM7) to reflect the most recent version of CHAD as of June 2014 (i.e., the July 2013 version). Accordingly, we also updated the HAPEM cluster file (“durhw_cluster.txt” in HAPEM6; “activity_cluster_v7.txt” in HAPEM7) and the HAPEM cluster-transition file (“clustertransa.txt” in HAPEM6; “activity_ClusterTransition_v7.txt” in HAPEM7).

Starting with HAPEM5, we analyzed CHAD data to create longitudinal activity patterns using Markov chains. For HAPEM7, we have refitted the Markov chain model to the most recent CHAD that now includes more activity pattern studies and, thus, more daily activity patterns. The data analysis groups the CHAD daily activity patterns into one, two, or three activity categories (or “clusters”) of similar activity patterns for each of 36 combinations of type of day (i.e., summer weekday, non-summer weekday, and weekend), age (i.e., 0–1, 2–4, 5–15, 16–17, 18–64, and 65 and older), and commuter type (i.e., commutes or does not commute).[15] Whether one, two, or three activity clusters are assigned to a day-age-commuter combination depends on the availability of CHAD data; for HAPEM7, 17 day-age-commuter combinations were assigned three clusters, two were assigned two clusters, and 17 were

[15] Appendix A of the HAPEM6 User’s Guide[5] was not updated to reflect these 36 groups. HAPEM5 used different age groups and used gender instead of commuter type. We included commuter type in HAPEM6 and HAPEM7 to better reflect the different activity patterns and exposures between commuters and non-commuters.

assigned one cluster. We defined clusters based on similar times spent in five broad MEs (i.e., indoors residence, indoors other, outdoors near-roadway, outdoor other, and in-vehicle).[16]

In HAPEM, for each such day-age-commuter combination, one daily activity pattern is randomly selected from all the CHAD data that correspond to that combination. The starting activity category (i.e., for the first day) is selected according to the relative frequencies of each category. The activity category for the second day is selected according to the transition probabilities from the starting category. Transition probabilities are the relative frequencies of each activity category when the same subject was in the starting category on the first day and the given activity category on the next day. The activity 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 activity categories. For a given subject, each day is assigned an activity pattern representative of the day's activity category. Once a particular activity pattern is selected as representative of an activity category, that pattern is always used for that category for that subject. Further details on the cluster and cluster-transition analysis can be found in Appendix A of the HAPEM6 User's Guide[5] (and, more specific to HAPEM7, in a 2015 memorandum from ICF to EPA's Ted Palma and Terri Hollingsworth,[17] which will likely also become Appendix A of the forthcoming HAPEM7 User's Guide[1]).

4.1. Quality Assurance

We ensured that each CHAD record was represented in the activity file and formatted appropriately, including the proper sets of columns for each day-age-commuter combination. We ensured that the same CHAD records were represented in the cluster file and formatted appropriately. We ensured that the cluster-transition file contained the correct combinations of day-age-commuter and was formatted appropriately. We discovered one error where people aged 15 years were placed into the wrong HAPEM age group; we repaired the script and regenerated the outputs.

4.2. Content and Format

The HAPEM7 activity data are contained in a fixed-with, space-delimited text file with characteristics shown in Table 11. The file contains 877 columns and a total of 45,628 rows of data with one header row. Each row corresponds to a person-day of activity in CHAD, where the first field contains an identifier for the record, the next 12 fields can be used together to describe the study subject, and the remaining fields contain duration values for how long the subject spends in each ME, for each hour of a day, and at work versus at home. All values in fields 14–877 are displayed as whole numbers (i.e., whole minutes).

[16] There were no activity patterns for the day-age-commuter = 1-1-2, but we included that case in cluster-transition file in order for HAPEM to execute properly.

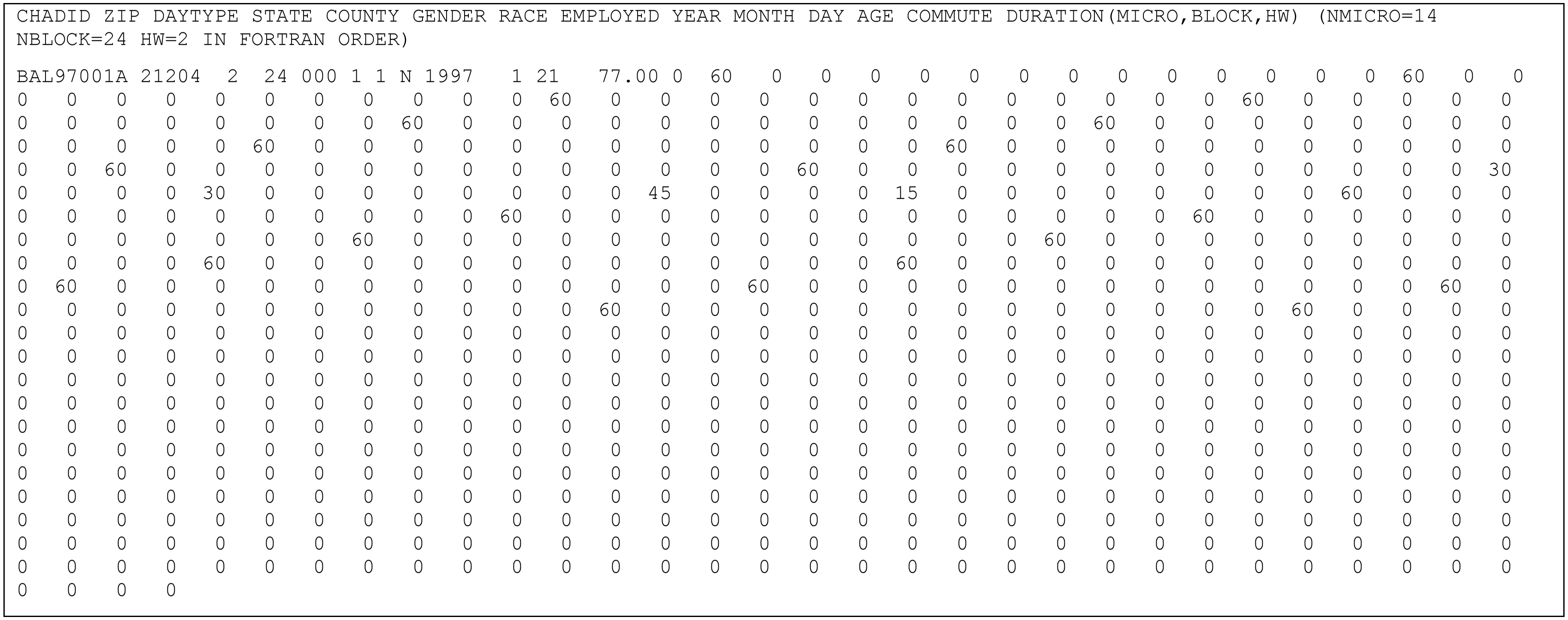
[17] ICF Memorandum "Update to: Proposed modification of HAPEM algorithm for creating longitudinal activity patterns: Results of data analysis." from March 30, 2015, addressed to Ted Palma and Terri Hollingsworth at EPA-OAQPS.

Table 11. Characteristics of the HAPEM7 Activity File

Variable Number	Variable	Description	Character Start Position on Data Row	Character Length on Data Row[a]
1	CHADID	ID of event in CHAD	1	10
2	ZIP	ZIP Code of subject's residence	11	5
3	DAYTYPE	Type of day when the event took place (1=summer weekday, 2=non-summer weekday, 3=weekend)	18	1
4	STATE	2-character FIPS code of state where event took place	21	2
5	COUNTY	3-character FIPS code of county where event took place	24	3
6	GENDER	Gender of subject (1=female, 2=male, 9=unknown)	28	1
7	RACE	Race of subject (1=white non-Hispanic, 2=black non-Hispanic, 3=Hispanic any race, 4=Asian or other non-Hispanic, 9=unknown)	30	1
8	EMPLOYED	Employment status of subject ("Y"=employed, "N"=unemployed, "X"=missing)	32	1
9	YEAR	Year when the event took place	34	4
10	MONTH	Month when the event took place	40	2
11	DAY	Day when event took place	43	2
12	AGE	Age of subject (presented to two decimal places; -999.00=missing)	46	7
13	COMMUTE	Commuter status of subject (0=does not commute, 1=commutes)	54	1
14-877	DURATION(MICRO,BLOCK,HW)	Duration of event (minutes). There are 864 of these fields, cycling through each of the 18 microenvironments; within each microenvironment, cycling through each of the 24 hours of the day; within each microenvironment-hour, cycling through whether the subject is at work or at home.	57	Each entry is 2 characters. Data record is complete after character 3,510.

[a]Any unused character space before a number or character and/or between fields consists of blank spaces.

Figure 11 and Figure 12 respectively contain the header and first data row of the HAPEM6 and HAPEM7 activity files.



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

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

Note: Contents wrap around due to space constrictions in this figure.

Figure 11. Excerpt from the HAPEM6 Activity File (“durhw.txt”)

```
CHADID ZIP DAYTYPE STATE COUNTY GENDER RACE EMPLOYED YEAR MONTH DAY AGE COMMUTE DURATION(MICRO,BLOCK,HW) (NMICRO=18
NBLOCK=24 HW=2 IN FORTRAN ORDER)

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

Note: Contents wrap around due to space constrictions in this figure.

Figure 12. Excerpt from the HAPEM7 Activity File (“activity_CHAD_v7.txt”)

The HAPEM7 activity-cluster data are contained in a fixed-with, space-delimited text file with characteristics shown in Table 12. The file contains six columns and one data row corresponding to each data row in the activity file, plus one header row. The first field contains an identifier for the record, the next three fields together identify the age-day-commuter combination of the event, and the final two fields respectively identify the cluster number of the event and the number of clusters that exist for all records corresponding to the age-day-commuter combination.

Table 12. Characteristics of the HAPEM7 Cluster File

Variable Number	Variable	Description	Character Start Position on Data Row	Character Length on Data Row[a]
1	CHADID	ID of event in CHAD	1	10
2	Demographic	Number corresponding to the HAPEM age group of the subject (i.e., 1=ages 0–1, 2=ages 2–4, and so on).	16	1
3	DayType	Type of day when the event took place (1=summer weekday, 2=non-summer weekday, 3=weekend)	23	1
4	Comtype (1=non-commute,	Commuting status of subject (1=does not commute, 2=commutes)	30	1
5	CLUSTER	Cluster category of event	37	1
6	Ncluster	Number of clusters for the corresponding combination of demographic, DayType, and Comtype	43	1

[a] Any unused character space before a number or character and/or between fields consists of blank spaces.

Figure 13 and Figure 14 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 cluster files.

```
CHADID Demographic   DayType       "Comtype (1=non-commute,"   CLUSTER       Ncluster
CAC01166A     1       1       1       1       1
CAC01251A     1       1       1       1       1
CAC01489A     1       1       1       1       1
CAC01562A     1       1       1       1       1
CAC01568A     1       1       1       1       1
CAC01809A     1       1       1       1       1
CAC01830A     1       1       1       1       1
CAC01982A     1       1       1       1       1
CAC02036A     1       1       1       1       1
CAC02132A     1       1       1       1       1
```

Figure 13. Excerpt from the HAPEM6 Cluster File (“durhw_cluster.txt”)

```
CHADID Demographic DayType "Comtype(1=non-commute," CLUSTER Ncluster
CAC01166A      1        1        1        1      1
CAC01251A      1        1        1        1      1
CAC01489A      1        1        1        1      1
CAC01562A      1        1        1        1      1
CAC01568A      1        1        1        1      1
CAC01809A      1        1        1        1      1
CAC01830A      1        1        1        1      1
CAC01982A      1        1        1        1      1
CAC02036A      1        1        1        1      1
CAC02132A      1        1        1        1      1
```

Figure 14. Excerpt from the HAPEM7 Cluster File (“activity_cluster_v7.txt”)

The HAPEM7 activity-cluster-transition data are contained in a fixed-with, space-delimited text file with characteristics shown in Table 13. The file contains 16 columns and one data row corresponding to each age-day-commuter combination, plus one header row. The first three fields together identify these combinations, the fourth field identifies the number of clusters that exist for that combination, fields 5–7 contain the cumulative fractions of the corresponding age-day combinations within each cluster, and the remaining fields identify the cumulative transition probabilities of all possible combinations of the subject’s cluster number on day X and the subject’s cluster number on day X+1.

Table 13. Characteristics of the HAPEM7 Cluster-transition File

Variable Number	Variable	Description[a]	Character Start Position on Data Row	Character Length on Data Row[b]
1	Demographic	Number corresponding to the HAPEM age group of the subject (i.e., 1=ages 0–1, 2=ages 2–4, and so on).	1	1
2	DayType	Type of day when the event took place (1=summer weekday, 2=non-summer weekday, 3=weekend)	8	1
3	Comtype (1=non-commute,2=commuting)	Commuting status of subject (1=does not commute, 2=commutes)	15	1
4	Ncluster	Number of clusters for the corresponding combination of demographic, DayType, and Comtype	22	1
5	cluster1	Cumulative fraction of demographic group/day type in cluster #1	28	7
6	cluster2	Cumulative fraction of demographic group/day type in clusters #1–2	36	7
7	cluster3	Cumulative fraction of demographic group/day type in clusters #1–3	44	7
8	prob11	Cumulative transition probability from cluster #1 to #1	52	7
9	prob12	Cumulative transition probability from cluster #1 to clusters #1–2	60	7

Table 13. Characteristics of the HAPEM7 Cluster-transition File

Variable Number	Variable	Description[a]	Character Start Position on Data Row	Character Length on Data Row[b]
10	prob13	Cumulative transition probability from cluster #1 to clusters #1–3	68	7
11	prob21	Cumulative transition probability from cluster #2 to #1	76	7
12	prob22	Cumulative transition probability from cluster #2 to clusters #1–2	84	7
13	prob23	Cumulative transition probability from cluster #2 to clusters #1–3	92	7
14	prob31	Cumulative transition probability from cluster #3 to #1	100	7
15	prob32	Cumulative transition probability from cluster #3 to clusters #1–2	108	7
16	prob33	Cumulative transition probability from cluster #3 to clusters #1–3	116	7

[a] For the cluster* fields, if Ncluster = 1 then cluster2 and cluster3 = 0 in the file; similarly, if Ncluster = 2 then cluster3 = 0. The same is true for the prob* fields (if Ncluster = 1 then prob12, prob13, prob21, prob22, prob23, prob31, prob32, and prob33 = 0, and if Ncluster = 2 then prob13, prob23, prob31, prob32, and prob33 = 0).

[b] Any unused character space before a number or character and/or between fields consists of blank spaces.

Figure 15 and Figure 16 respectively contain the first ten data rows of the HAPEM6 and HAPEM7 cluster-transition files.

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

1      1      1      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

1      1      2      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

1      2      1      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

1      2      2      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

1      3      1      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

1      3      2      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

2      1      1      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

2      1      2      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000

2      2      1      3      0.50435       0.78261       1.00000       0.78947       0.84211       1.00000       0.06250
       0.93750       1.00000       0.42857       0.57143       1.00000

2      2      2      1      1.00000       0.00000       0.00000       1.00000       0.00000       0.00000       0.00000
       0.00000       0.00000       0.00000       0.00000       0.00000
```

Note: Contents wrap around due to space constrictions in this figure.

Figure 15. Excerpt from the HAPEM6 Cluster-transition File (“clustertransa.txt”)

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

1       1       1       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000

1       1       2       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000

1       2       1       2      0.79132 1.00000 0.00000 0.96552 1.00000 0.00000 0.50000 1.00000 0.00000 0.00000 0.00000 0.00000

1       2       2       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000

1       3       1       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000

1       3       2       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000

2       1       1       3      0.56255 0.88783 1.00000 0.60392 0.91373 1.00000 0.37383 0.84579 1.00000 0.32000 0.69333 1.00000

2       1       2       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000

2       2       1       3      0.48133 0.69467 1.00000 0.66667 0.82456 1.00000 0.07692 0.71795 1.00000 0.19608 0.39216 1.00000

2       2       2       1      1.00000 0.00000 0.00000 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
```

Figure 16. Excerpt from the HAPEM7 Cluster-transition File (“activity_ClusterTransition_v7.txt”)

5. Updating Source Code

We made several modifications to the module source code for HAPEM7. All modifications were minor and functioned either to ensure proper execution from the command line or to ensure that array dimensions were large enough to accommodate the revised default model input data. We describe below the specific changes we made to the specific modules.

- DURAV, COMMUTE, AIRQUAL, and HAPEM modules:
  - Added GETARG function allowing for model parameters file (i.e., run control file) to be called from command line.
  - Expanded variables and dimensions related to array sizes, necessary to accommodate revised default model input data that contain more lines and/or columns than in HAPEM6:
    - DURAV and HAPEM: Expanded Max1 to 10000
    - COMMUTE: Expanded wrtractu, Udispro, ngroup, Uareapro, htracid, first2, and last2 to 80000
    - AIRQUAL: Expanded m to 80000
- INDEXPOP module:
  - No changes.

The model executables for HAPEM7 are named: durav7.exe, indexpop7.exe, commute7.exe, airqual7.exe, and hapem7.exe.