CEMDAP: Modeling and Microsimulation Frameworks, Software Development, and Verification

Abdul Pinjari
The University of Texas at Austin, Department of Civil, Architectural & Environmental Engineering
1 University Station, C1761, Austin, TX 78712
Phone: (512) 471-4535; Fax: (512) 475-8744; Email: abdul.pinjari@mail.utexas.edu

Naveen Eluru
The University of Texas at Austin, Department of Civil, Architectural & Environmental Engineering
1 University Station, C1761, Austin, TX 78712
Phone: (512) 471-4535; Fax: (512) 475-8744; Email: naveeneluru@mail.utexas.edu

Sivaramakrishnan Srinivasan
University of Florida, Department of Civil and Coastal Engineering
365 Weil Hall, PO Box 116580, Gainesville, FL 32608
Phone (352) 392-9537 Extn. 1456; Fax: (352) 392-3394; Email: siva@ce.ufl.edu

Jessica Y. Guo
University of Wisconsin-Madison, Department of Civil and Environmental Engineering
1206 Engineering Hall, 1415 Engineering Dr, Madison, WI 53706
Phone: (608) 890-1064; Fax: (608) 262-5199; Email: jyguo@wisc.edu

Rachel Copperman
The University of Texas at Austin, Department of Civil, Architectural & Environmental Engineering
1 University Station, C1761, Austin, TX 78712
Phone: (512) 471-4535; Fax: (512) 475-8744; Email: RCopperman@mail.utexas.edu

Ipek N. Sener
The University of Texas at Austin, Department of Civil, Architectural & Environmental Engineering
1 University Station, C1761, Austin, TX 78712
Phone: (512) 471-4535; Fax: (512) 475-8744; Email: ipek@mail.utexas.edu

Chandra R. Bhat*
The University of Texas at Austin, Department of Civil, Architectural & Environmental Engineering
1 University Station, C1761, Austin, TX 78712
Phone: (512) 471-4535; Fax: (512) 475-8744; Email: bhat@mail.utexas.edu

*corresponding author. This paper was written when the corresponding author was a Visiting Professor at the Institute of Transport and Logistics Studies, Faculty of Economics and Business, University of Sydney.

ABSTRACT

The Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns (CEMDAP) is a micro-simulation implementation of a continuous-time activity-travel modeling system. Given as input various socio-demographic, land-use, and transportation level-of-service attributes, the system provides as output the complete daily activity-travel patterns for all individuals of a population. This paper describes the current state of CEMDAP and highlights the salient features of the software. CEMDAP models not only the activity-travel pattern of adults, but also that of children, while incorporating the inter-dependencies between the activity-travel patterns of children and their parents. The software implementation of CEMDAP has been developed using the Object-Oriented (OO) paradigm to support software extensibility and rapid implementation of system variants. Further, the implementation supports multithreading and data caching capabilities to enhance computational performance. The paper discusses these features, and also presents the results from an application of CEMDAP to the Dallas-Fort Worth area. Verification exercises establish the reasonableness of CEMDAP outputs.

1. INTRODUCTION

CEMDAP, or the Comprehensive Econometric Microsimulator for Daily Activity-travel Patterns, is a disaggregate (individual-level), continuous-time, activity-travel forecasting system developed at The University of Texas at Austin. In a paper in 2004, Bhat *et al.* (*1*) described the methodological structure and the software implementation details of the first version of this microsimulation system. Since then, CEMDAP has undergone substantial enhancements in the choice dimensions modeled and the forecasting sequence, as well as the software design. This paper describes the new econometric modeling system and the microsimulation framework embedded within CEMDAP, and also presents an application of the software to the Dallas-Fort Worth (DFW) area.

The reader will note here that the design and architecture of CEMDAP is generic. In particular, the modeling platform can be applied to any metropolitan area, as long as local area models are estimated to produce the appropriate sensitivity parameters. Currently, we have estimated all the CEMDAP models using the DFW data, and the resulting specifications/parameters are embedded as default specifications/parameters. Moreover, the user can use the graphical interface in the program to modify the specifications and/or parameter values if local area specifications/parameters are available (see the CEMDAP user manual by Bhat *et al.* [*2*] for details on modifying the specifications). The system has also been designed to provide a friendly diagrammatic interface to help the user understand the logic of the system.

The rest of the paper is organized as follows. Section 2 describes the econometric modeling system and the microsimulation framework embedded within CEMDAP, highlighting its many salient features. Section 3 is focused on the software design issues. Specifically, the software architecture and the strategies adopted for enhancing the computational performance are discussed. Section 4 provides an overview of the procedures used to generate inputs for applying CEMDAP to the DFW area. The validation of model application is discussed in Section 5. Finally, Section 6 summarizes the paper.

We should point out here that paper length considerations do not permit a comprehensive discussion of all structural, estimation, application, and validation details of this complex microsimulation system. The reader is referred to Pinjari *et al.* (*3*) for complete documentation.

2. CEMDAP FRAMEWORK

2.1 Modeling Framework

CEMDAP comprises a suite of econometric models, each model corresponding to the determination of one or more activity/travel choices of an individual or household. These models may be broadly grouped into two systems: (1) The generation-allocation model system and (2) The scheduling model system. The first system of models is focused on modeling the decision of individuals/households to undertake different types of activities (such as work, school, shopping, and discretionary) during the day and the allocation of responsibilities among individuals (for example, determination of which parent would escort the child to and from school). Table 1 lists the precise econometric structure and the choice alternatives for each of the model components in this system. Further, there is a unique identifier associated with each model. (For example, “GA1” identifies the first model within the “generation-allocation” category, which is the decision of a child to go to school.) To facilitate easy cross-referencing, these identifiers have also been included in subsequent figures that we will reference (and that provide an overview of the microsimulation procedure implemented within CEMDAP for predicting the complete

activity-travel patterns of all individuals in a household). The second system (*i.e.*, the scheduling model system) determines how the generated activities are scheduled to form the complete activity-travel pattern for each individual in the household, accommodating the space-time constraints imposed by work, school, and escort of children activities. That is, these models determine the choices such as number of tours, mode and number of stops for each tour, and the activity-type, location, and duration for each stop in each tour. Table 2 lists the econometric structures and the set of choice alternatives for each model in this second system. The first ten models in Table 2 (WS1-WS10) correspond to worker scheduling components, the next eleven models (NWS1-NWS11) are associated with non-worker scheduling components, the subsequent four models (JS1-JS4) relate to joint discretionary tour scheduling components, and the final seven models (CS1-CS7) focus on children scheduling components.

The reader will observe from Tables 1 and 2 that the econometric structure for each choice dimension being modeled in CEMDAP falls under one of the six econometric model categories: binary logit, multinomial logit, hazard-duration, regression, ordered probit, and spatial location choice. The mathematical model structures of these model categories are provided in Bhat *et al.* (*4*).

The model system described above has several salient features, which include the (1) use of a continuous-time approach that enables the evaluation of such time-of-day varying transportation control measures as dynamic congestion pricing strategies and parking policies at a fine resolution of time (up to a minute), (2) accommodation of within-individual space-time constraints and interactions in daily activity-travel pattern choices, (3) modeling of the activity-travel patterns of children, (4) explicit consideration of the interdependencies between the activity-travel patterns of children and their parents (such as escort to and from school and joint participation in discretionary activities), (5) adoption of a sequencing structure of the models that accommodates intra-personal temporal constraints[1], (6) use of a fine level of disaggregation in the out-of-home activity types considered (the current system uses 11 activity types for adults and 3 for children), (7) explicit distinction between the driver and the passenger in the mode choice alternatives instead of using an aggregate "shared ride" alternative, and (8) ability to be applied at any spatial and temporal resolution (currently, CEMDAP has been applied to a 4874 zone system for the Dallas/Fort-Worth area in Texas, and accommodates varying level-of-service variables for five time periods of the day). The third through eighth features are new features added in the latest version of CEMDAP.

The data used in the estimation of all the model components in Tables 1 and 2 were obtained from three main sources: (1) the 1996 DFW household activity survey, (2) the DFW zonal land-use database, and (3) the DFW inter-zonal transportation level of service data. All three data sets were acquired from the North Central Texas Council of Governments (NCTCOG). Details of data preparation and the estimation results of each model component are available in Pinjari *et al.* (3).

[1] Specifically, the current version of CEMDAP models a tour entirely in terms of both the tour-level (mode, number of stops, departure time, and duration) and stop-level (activity-type, duration, travel time, and location) attributes prior to modeling a subsequent tour. This is different from the approach adopted in the previous version in which tour-level characteristics for *all* tours were modeled prior to determining the characteristics of stops within any tour. Our current approach provides better timing of the "return-home" trips of each tour and hence helps achieve better intra-personal temporal consistency.

2.2 Microsimulation Framework

This section provides an overview of the microsimulation procedure implemented within CEMDAP for predicting the complete activity-travel patterns of all individuals in a household. This procedure is repeatedly applied to each household in the input synthetic population to completely determine the activity-travel patterns of all individuals in the study area. The overall prediction procedure (for a household) can be subdivided into two major sequential steps, corresponding to the two broad modeling systems identified in the modeling framework of the previous section. The mathematical procedures to predict the choice outcomes from various econometric models such as the multinomial logit, ordered probit, hazard duration model, and linear regression are available in Bhat *et al.* (*5*).

The microsimulation prediction procedure (for a household) is represented schematically in Figure 1.[2] Each step in the figure involves the application of several models in a systematic fashion. Figure 1 includes the identification numbers (from Tables 1 and 2) of models associated with each of the major steps. As can be observed from Figure 1, the generation-allocation model system is first applied and this comprises the following three sequential steps:

(1) Work and school activity participation and timing decisions,
(2) Children's travel needs (such as mode to school and discretionary activity participation), and allocation of escort responsibilities to parents, and
(3) Independent activities (such as shopping, recreation, and personal business) for personal and household needs.

At the end of the prediction of activity generation and allocation decisions, the following information is available for the simulation day: (1) each child's decision to go to school, the school start time and end time, the modes used to travel to and from school, the decision to undertake a joint discretionary activity with a parent, and the decision to undertake an independent discretionary activity; (2) which (if either) parent undertakes the drop-off activity, the pick-up activity, and the joint discretionary activity with each child in the household; (3) each employed adult's decision to go to work, the work start time and end time, and the decision to undertake work-related activities; (4) each adult student's decision to go to school, and the school start time and end time; (5) each adult's decisions to undertake grocery shopping, personal or household business, social or recreational activities, eating out, and other serve-passenger activities.

Next, the scheduling model system is applied to predict the sequencing of the activities generated in the generation-allocation system, while accommodating the space-time constraints imposed by work, school, and escort-of-children activities. The complete scheduling is accomplished in the following sequence:

(1) Work-to-home and home-to-work and commutes for each worker (determines the commute mode, number of stops each way, and the activity type, episode duration, travel time, and location for each commute stop.)

[2] Due to space constraints, we are unable to discuss the complete details of the microsimulation prediction procedure or the procedures applied to assure intra-individual and inter-individual spatial and temporal consistency of the predicted activity-travel patterns. Further, the exact, detailed sequence of steps applied to determine the complete activity-travel patterns varies from one household to another depending on the household structure and the types of activities generated for the different members. We would like to invite readers to learn more details of the microsimulation procedure from Pinjari *et al.* (*3*), pages 17-56. This report is available at http://www.ce.utexas.edu/prof/bhat/REPORTS/4080_8_draft_Dec11_2006.doc.

(2) Drop-off tour of the non-worker escorting children to school (determines the tour mode, the number of stops following the drop-off stop, and the activity type, episode duration, travel time, and location for each of these stops.)
(3) Pick-up tour of the non-worker escorting children from school (determines the tour mode, the number of stops following the pick-up stop, and the activity type, episode duration, travel time, and location for each of these stops.)
(4) School-to-home and home-to-school commutes for each school-going child. For children who are not escorted by their parents, it is assumed that there are no commute stops and the only attribute determined at this step is the commute duration. Note that the mode for school commute is already known from step 2 of the generation-allocation system. For children escorted by their parents, the attributes are simply copied from the corresponding pick-up or drop-off segments of the corresponding parent.
(5) Joint tour of the adult pursuing discretionary activity jointly with children (determines the departure time for the tour, and the episode duration, travel time, and destination for the joint discretionary activity stop)
(6) Independent home-based tours and work-based tours for each worker (determines the number of before-work, work-based, and after-work tours, and for each tour, home/work-stay duration, mode, and the number of stops, and for each stop in each of the tours, the activity type, episode duration, travel time, and location)
(7) Independent home-based tours for each non-worker (determines the number of home-based tours, and for each tour, home-stay duration before the tour, mode, and the number of stops, and for each stop in each of the tours, the activity type, episode duration, travel time, and location)
(8) Independent discretionary activity tour for each child (determines the tour mode, and departure time, and the activity duration, travel time, and location of the discretionary activity stop)

In addition to these stochastic models, several deterministic rules are also employed within each step based on a descriptive analysis of the DFW survey data. Examples include the following: (a) If a worker picks-up (drops off) his/her child from (at) school, this is taken as the only stop in his/her work-to-home (home-to-work) commute, (b) The mode of travel for a pick-up/drop-off activity is taken as drive with passenger and the mode of travel for the remaining part of a pick-up/drop-off tour is taken as drive alone, (c) The departure time and the travel time to the pick-up/drop-off stop is determined based on the school end/start time and the prevailing travel-times between work/home and school locations at the school end/start time, (d) The duration of a pick-up/drop-off episode is taken as 5 minutes, (d) The travel time to home/work in the final segment of a tour is determined based on the prevailing travel times between origin and destination locations in that time period, (e) If a worker undertakes a joint discretionary activity, the number of after-work tours for him/her is fixed as one joint discretionary tour, and (f) The mode of travel for the adult in a joint discretionary tour is taken as drive with passenger and the number of stops is fixed to one in that tour.

The forecasting sequence described in Figure 1 highlights CEMDAP's *interleaved* approach to determining the activity-travel patterns of all individuals in a household. This idea is illustrated with the following example. In households with school-going children and employed parents, the child's decision/need to go to school and the school timings are first determined. Next, the employed parents' decisions to go to work and the work timings are determined

conditional on the child's school-related choices (since, for example, a parent's decision to go to work may be impacted by a child not being able to go to school due to sickness). The children's travel needs (mode of travel to school) are determined subsequently conditional on both the child's school timings, and parents' work timings. Depending on the children's school mode choice, (*i.e.,* if the mode chosen is "*driven by parents"*), one of the parents is allocated the task of dropping off/picking up the children, and that parent's work timings are adjusted to allow him/her to undertake the drop-off/pick-up activity. Thus, the activity-travel patterns of household members are not generated either purely sequentially (*i.e.*, one person followed by another) or purely simultaneously (*i.e.*, all persons together). Rather, while the individual decisions are modeled sequentially, the overall activity-travel patterns of all household members are generated in an interleaved, parallel, fashion. This approach enables incorporation of intra-household constraints and spatial/temporal consistency across the activity-travel patterns of household members while limiting the computational complexity.

3. SOFTWARE DESIGN AND DEVELOPMENT

The development of the CEMDAP software goes beyond a once-off implementation of a specific modeling system calibrated for a specific region. Rather, the goal is to create a generic library of routines that form the building blocks of an activity-based travel-demand modeling system. Correspondingly, CEMDAP has been developed using the Object-Oriented (OO) paradigm, which offers the advantages of code reuse, software extensibility, and rapid implementation of system variants. The software is written in Visual C^{++} using the Microsoft Visual Studio .NET development tool.

CEMDAP uses PostgreSQL to store input databases, which allows the ability to work with a fine resolution of spatial units and/or large study areas. For computational efficiency considerations, CEMDAP supports multithreading and includes data caching techniques to store frequently accessed input data elements in the RAM. Also, the (pseudo)random numbers used to simulate the activity-travel patterns of each individual in CEMDAP are held to be the same across different policy scenario runs. This helps in minimizing the random simulation bias in policy analyses, and allows a disaggregate level (*i.e.,* the individual level) assessment of policies.

The rest of this section is organized as follows. Section 3.1 describes the software architecture. Section 3.2 discusses computational performance issues and methods adopted (multithreading and caching) to enhance the speed. Comprehensive details of the software architecture are available in Chapter 3 of Pinjari *et al.* (*3*).

3.1 Software Architecture

Figure 2 presents a schematic representation of the CEMDAP software architecture. The major components of this software are: the Input Database, the Data Coordinator, the Run-time Data Objects, the Modeling Modules, the Simulation Coordinator, the Application Driver, and the Output Files. Each of these components is further discussed below.

The *input data* are stored in a relational database management system (DBMS). CEMDAP is designed to interact with this *Input Database* through an Open Database Connectivity (ODBC) interface. The *ODBC* provides a product-independent interface between client applications (CEMDAP, in this case) and database servers, allowing applications to be portable across database servers from different manufacturers. Another advantage of interfacing through an ODBC interface is that the database servers and the CEMDAP application can be run on different machines with no additional complexity in interacting with the database over the

network. Further, the ODBC interfacing with CEMDAP is enabled to accept inputs of any given spatial and temporal resolution, within the limits of the processing power at hand.

The *Data Coordinator* is the component responsible for establishing the ODBC connection and interacting with the *Input Database*. It extracts the content and the structural information of the data tables, and converts data into their corresponding data structures that are used within CEMDAP. It is also responsible for all data queries to the database during the process of simulation. By limiting the database interaction to this one system entity, any changes pertaining to the database are easier to make.

The *Run-Time Data Objects* are the main data structures that CEMDAP operates upon internally. Instances of household, person, zone, zone to zone, and LOS entities are created by the *Data Coordinator* from the *Input Database*. The remaining entities (*i.e.* pattern, tour, and stop) are created by the *Simulation Coordinator* as required during the simulation process. The *Run-Time Data Objects* also act as a cache for the data items accessed frequently by the *Simulation Coordinator*.

Each *Modeling Module* in the system corresponds to a behavioral model in the framework described in Section 2. Once a *Modeling Module* is configured via the user interface, it possesses knowledge about the econometric structure and all the relevant parameters required to predict a particular activity-travel choice. Although the *Modeling Modules* are many, they are derived from a limited number of econometric structures. Currently, six types of econometric models are implemented in CEMDAP as model templates: regression, hazard duration, binary logit, multinomial logit, spatial location choice, and ordered probit models. Additional econometric structures may be added to this library of model templates.

The *Simulation Coordinator* is responsible for controlling the flow of the simulation. It coordinates the logic and sequence in which the *Modeling Modules* are called, performs consistency checks, and keeps track of the progress of the overall simulation. The *Simulation Coordinator* holds a reference to the *Data Coordinator* and operates on the *Run-Time Data Objects* which are created and manipulated as choice outcomes are predicted with each modeling component.

The *Application Driver* starts and runs the application. On startup, it triggers the user interface and obtains handles to the *Simulation Coordinator* as well as the *Data Coordinator*. It references the ODBC driver for opening and closing the database connection. It also co-ordinates the functionality offered to the users, such as selecting the input data source, choosing the output path, loading/saving the CEMDAP model specification files, and running the simulation.

The *Output* of CEMDAP is stored in flat-files (plain tabbed formatted files). As the activity-travel patterns are generated sequentially (one household at a time) the CEMDAP outputs can be streamed to flat files. Further, data in flat-file formats can be easily read by spreadsheet, statistical, and DBMS programs thereby providing the user with the flexibility of analyzing the results with any type of software.

3.2 Computational Performance Enhancement

There are two critical aspects which impact the run-time performance (speed) of the CEMDAP software. First, the simulation procedure generates the activity-travel patterns for one household at a time until all the households in the population have been processed. Typically, the synthetic population for a study area might comprise several million households, thereby requiring substantial run time for the simulation of the activity-travel patterns of the entire population. Second, the input data are stored in an external relational database and interfaced with the

program via the ODBC. With increasing number of queries, data access through the ODBC interface can significantly increase the processing time and degrade the system performance. CEMDAP employs the multithreading technique to address the first issue and data caching to address the second. These strategies are described below.

Multithreading functions by loading the data and information pertaining to multiple tasks (instead of a single task) into the memory of a processor and hence improves the overall utilization of the computational resources. The processor rapidly switches between the various tasks at a fixed time interval called the "time slice". In CEMDAP, multithreading is enabled by loading the input data related to several households into the processor. It should be noted here that the time slice has to be small enough to allow a large number of tasks (households in this case) to be handled. At the same time, each time slice has to be large enough so that each task is allocated a sufficient amount of processor time to get useful work done. The number of threads that can be run at a time (or the number of households that can be loaded into the memory of the processor at a time) depends on the processor speed and the Random Access Memory (RAM) of the machine. CEMDAP allows customization of the extent of multithreading via direct changes to the code.

Data Caching involves loading selected sections of the input data into the computer's RAM to reduce the number of data access calls through the ODBC interface. In the case of CEMDAP, caching is done especially for the inter-zonal level-of-service (LOS) data. This is because the LOS data tables are typically very large (the LOS file for the DFW application has 4874*4874 zonal pairs and five time-of-day periods) and accessed frequently (for example, inter-zonal travel times are required for location choice predictions and, hence, the number of times the LOS database has to be accessed for a single individual is at least equal to the number of activity stops made by him/her). It may be possible to cache the entire LOS data for achieving greater simulation speeds. However, any move toward finer spatial and/or temporal resolutions and larger study areas would cause a significant increase in the LOS data size, and limit the extent to which the LOS data can be cached. Hence, cleverly designed partial-data caching routines are built into CEMDAP so that frequently used data are temporarily cached. For example, the LOS data corresponding to an origin zone is cached until all the households belonging to that particular zone have been processed. Similarly, the commute LOS data (the LOS data between residential and employment zones during the commute start and end times) of a worker is cached when (s)he is being processed. The optimal extent of data-caching depends on the machine configuration (RAM and the processor speed), and the size and organization of the input data (*i.e.*, the spatial and temporal resolution at which the LOS files are loaded). The extent of data caching in CEMDAP can be customized via direct changes to the code.

4. GENERATING INPUTS FOR CEMDAP

The application of CEMDAP for a study area requires two major categories of inputs: (1) the estimated model parameters and (2) data inputs for the forecast year (disaggregate characteristics of the population, zonal-level land use descriptors, and inter-zonal transportation level of service (LOS) variables by time of day). In the rest of this section, we briefly discuss how the data inputs were generated for the Dallas-Fort Worth (DFW) region for the base year of 2000. The specific focus here is on the generation of the detailed socioeconomic characteristics of the population, since the land-use and LOS files were directly available from NCTCOG. The other category of input, *i.e.*, the model parameters, were estimated using the 1996 household DFW travel survey, as discussed earlier.

CEMDAP requires detailed, individual- and household-level population characteristics as input. The individual-level attributes include age, gender, availability of driver's license, ethnicity, education level, income, employment-related characteristics (such as work location, weekly duration, flexibility, and industry type), and school-related characteristics (such as school location and grade). Household-level attributes include household size, composition, residential location, tenure, housing unit type, and automobile ownership. The age, gender, and ethnicity attributes at the individual level, and the household size, composition, and residential location attributes at the household level, are generated for the base year using the Synthetic Population Generation (SPG) module which implements an iterative proportional fitting (IPF) algorithm. Other base year socioeconomic attributes related to driver's license, schooling, and employment at the individual level, and residential tenure, housing unit type, and vehicle ownership at the household level, that are difficult to synthesize (or cannot be) synthesized directly from the aggregate socioeconomic data for the base year are simulated by the Comprehensive Econometric Microsimulator for SocioEconomics, Land-use, and Transportation System (CEMSELTS).[3]

The details of the procedures used in SPG are provided in Guo and Bhat (*6*), while the details of the procedures used in CEMSELTS are available in Eluru *et al.* (*7*). For the current application, three individual-level variables and four household-level variables were used as *control variables in the SPG module*. The individual-level variables include: (a) gender (2 categories), (b) race (7 categories), and (c) age (10 categories), while the household-level variables include: (a) family/non-family indicator (2 categories), (b) household type (5 categories), household size (7 categories), (c) presence of children (2 categories), and (d) age of household head (2 categories).The Census 2000 summary file SF1 is used to create the *aggregate target dataset* for the above mentioned *control variables,* and data from the US Census' Public Use Microdata Samples (PUMS) is used as the *disaggregate* "seed" data. Together, these two data sets are used to synthesize the base year population by gender, race, and age at the individual level and by family/household type, household size, presence of children, and age of household head at the household level.[4] The remaining data on schooling grade and school location for students, and employment characteristics (whether or not employed, employment industry, employment location, work duration, work flexibility, and personal income) at the individual level are generated in CEMSELTS. Also, housing tenure (own or rent home), housing unit type (Single-family detached, Single-family attached, Apartment, and Mobile home or trailer), and household vehicle ownership at the household level are generated in CEMSELTS.

[3] The base year synthetic disaggregate-level sociodemographic data generated by SPG and the base-year activity-travel environment attributes are used by CEMSELTS to generate additional disaggregate-level base-year socioeconomic data. The reader will note that an advantage of using stochastic models in CEMSELTS to generate some of the base year socioeconomic characteristics is that the synthetic population has more variation than would be obtained by simply expanding the disaggregate-level sample (usually the Public-Use Microdata Samples or PUMS data) employed in the SPG module. Also, SPG is used only to generate the disaggregate-level synthetic population for the base-year and is not used beyond the base year. CEMSELTS generates all the socioeconomic attributes of the population for any future year (see Eluru et al. [*7*]).

[4] The population synthesized by SPG locates households in block groups, since this is the spatial level used by the Census 2000 summary file SF1. The corresponding Traffic Analysis Zone (TAZ) locations, as required by CEMDAP, were generated by mapping the block groups to TAZs using a GIS software with the assumption that the households within a block group are uniformly distributed in space.

The disaggregate input population generated for the year 2000 DFW application using the methodology discussed above comprises 4,815,916 individuals from 1,785,653 households. The characteristics of this population have been validated against aggregate marginal distributions available from the 2000 PUMS and the 2000 US Census. As an example, Figure 4 illustrates the verification exercise carried out at an aggregate level for selected socio-demographic characteristics of the population from Tarrant County. The predicted distributions of the population closely track the Census distributions.

5. MODEL VERIFICATION

The verification of the DFW application of CEMDAP involved two efforts. First, the survey data used in model estimation were input to CEMDAP and the predicted activity-travel patterns were compared to the observed patterns (Section 5.1). Second, the activity-travel patterns were generated for the entire DFW population for the year 2000 (using inputs generated as described in Section 4). The generated patterns were then aggregated and compared with the travel-demand measures generated by the current DFW trip-based model and observed link counts (Section 5.2).

5.1 Validation Against the Estimation Data

The validation against the estimation data was undertaken at the aggregate level by comparing the predicted percentage shares of discrete choices and distributions of continuous choices with the observed percentage shares and distributions in the estimation survey sample.

Table 3 compares selected pattern-, tour-, and trip-level characteristics predicted by CEMDAP with those observed in the estimation survey data [see (*3*) for additional validation results]. Overall, the CEMDAP outputs match reasonably with the observed patterns in the DFW survey. Among the pattern-level characteristics (first part of Table 3), the predicted and observed number of non-school tours for children show some difference. This may be attributed to the small sample from which the models for children's non-school travel were estimated. An examination of the tour-level characteristics (second part of Table 3) shows that CEMDAP is under-predicting the average number of stops in the home-work and work-home commutes. In the context of trip-level characteristics (last part of Table 3), CEMDAP performs well in predicting the average number of daily trips per person for all trip types (*i.e.*, home-based work, home-based non-work, and non home-based trip types). However, we find a slight under-prediction in the average travel times for all trip types, possibly because CEMDAP directly uses the inter-zonal travel time values from the LOS files for certain trip segments (such as the return-home trips) as opposed to the door-to-door travel times reported in surveys. We also find an over-prediction of PMT and VMT for home-based other trips and an under-prediction of PMT and VMT for non home based trips. However, overall, the statistics are similar in range between the CEMDAP-predicted values and the actual survey observations.

Figure 5 presents the distribution of the work start and end times in the DFW survey data and as predicted by CEMDAP. CEMDAP replicates the overall shape of the profile; however, the sharp peaks observed in the survey are not captured.

5.2 Comparison with the DFW Trip-based Model and Observed Link Counts

The comparison of CEMDAP with the DFW's current trip-based model involved the following steps. First, the travel-demand patterns predicted by the DFW's current trip-based model for the year 1999 (4,848,237 persons from 1,808,402 households) were obtained. Second, the activity-

travel patterns for the entire DFW synthetic population for the year 2000 (4,815,916 individuals from 1,785,653 households generated as described in Section 4) were generated using CEMDAP. Third, the CEMDAP-generated activity-travel patterns were aggregated into origin-destination (O-D) trip tables by time-of-day for each auto mode (single occupancy and multiple occupancy). Fourth, estimates of external trips and truck trips were borrowed from the trip-based model and suitably added to the OD matrices from CEMDAP. Fifth, static traffic assignments were conducted, with the OD matrices as inputs into the traffic assignment procedures in DFW's current modeling software (CEMDAP does not perform traffic assignment). The results from this step are deemed as CEMDAP's travel predictions.

Table 4 presents a summary of the overall travel indicators from the CEMDAP and DFW trip-based model results. While the travel indicators such as total number of person trips, total number of vehicle trips, and average trip speed are quite close, other measures such as average trip length and total vehicle miles traveled (VMT) show some differences (with CEMDAP predicting higher values). Further, on examining the travel volumes by trip purpose, we find that CEMDAP predicts fewer home-based work trips and greater numbers of home-based-other and non-home-based trips than the DFW model. These differences can be attributed to the difference between the number of employed individuals in the CEMDAP input as predicted from CEMSELTS (the percentage of employed individuals was 48.1% of the overall population from CEMSELTS, which matches well with the 49.4% employment rate for the DFW population from the 2000 Census data statistics), and the number of employed individuals used in the DFW trip-based model input (which is 62.3% of the DFW population).

It is important to note here that the results above cannot be directly interpreted as over-predictions or under-predictions by any one modeling approach, as neither predictions represent the "ground truth". The intent of the above comparison is to just ensure that CEMDAP does not produce results that are completely unreasonable. Another way to check the CEMDAP results is to examine the link flows predicted by CEMDAP with link vehicle counts. The results (%RMSE values) are presented in the second part of Table 4 by roadway functional class. Overall, the validation results indicate that the performances of both CEMDAP and the trip-based model (without K factors) against the ground "truth" are close to each other.

The aggregate-level comparisons of the CEMDAP results with the trip-based model results and observed ground counts (as discussed above) are intended to establish the preliminary reasonableness of CEMDAP outputs. It is also important to note here that, unlike the DFW trip-based model, the CEMDAP results have not been "calibrated/adjusted" in any way. Rather the CEMDAP results are direct predictions based on the estimated models from the DFW survey data. Besides, CEMDAP provides several other details of the activity-travel characteristics (such as activity episode durations, extent of trip chaining, and inter-personal constraints/consistency) which are simply not provided by trip-based models Further, the use of the static assignment process does, to an extent, "undo" the benefits of a continuous-time modeling system. This is because the activity-travel patterns are grouped into aggregate time periods in the static assignment stage and the static assignment process does not consider the dynamics of vehicle delays.

The activity-based predictions may be validated in a more rigorous manner by using a dynamic traffic assignment procedure to predict the traffic volumes. In any case, the real validity of any model should be measured in terms of its ability to forecast well into the future and respond appropriately to transport policies. In this context, the focus should be on the level of behavioral fidelity captured in the model. The better the behavioral fidelity of a model, the better

it will be in terms of transferability in time and space (especially if the demographics and travel environment change substantially over time and space). The behavioral fidelity of CEMDAP and trip-based models can be qualitatively examined by comparing the outputs of the two models for several policy scenarios. While we have undertaken such an extensive exercise (see Pinjari *et al*. [*3*], Chapter 6), one problem is that one still does not know which output predictions are the right ones in the absence of "ground truth". One fruitful way forward to assess activity-based models and trip-based models would be to compare before-after results in response to such policy actions as implementation of auto-use disincentives (congestion pricing, toll roads), car pooling incentives (HOV lanes), transit improvements, and land-use changes, *etc.*. Such an exercise is planned as part of our future work in the Dallas-Fort Worth area.

6. SUMMARY AND FUTURE WORK

This paper describes the current state of CEMDAP and highlights the salient features of the software. CEMDAP models not only the activity-travel pattern of adults, but also that of children, while incorporating the inter-dependencies between the activity-travel patterns of children and their parents. The software implementation of CEMDAP has been developed using the Object-Oriented (OO) paradigm to support software extensibility and rapid implementation of system variants. Further, the implementation supports multithreading and data caching capabilities to enhance computational performance. The paper discusses these features, and also presents the results of an application of CEMDAP to the Dallas Fort Worth area. The results indicate the reasonableness of the activity-travel predictions from CEMDAP, and the readiness of the system for more rigorous before-after sensitivity testing.

ACKNOWLEDGEMENTS

The research in this paper was funded by a Texas Department of Transportation (TxDOT) project entitled "Second Generation Activity-Based Travel Modeling System for Metropolitan Areas in Texas Accommodating Demographic, Land Use, and Traffic Microsimulation Components". The authors would like to thank Janie Temple and William Knowles of the Transportation Planning and Programming Division of TxDOT for their input and suggestions during the course of the TxDOT project. The authors are also very grateful to Ken Cervenka of the Federal Transit Administration (FTA) and Arash Mirzaei of the North Central Texas Council of Governments (NCTCOG) for their help in the comparison of CEMDAP predictions against the DFW trip-based model and observed link counts. Thanks to Sanketh Indarapu for coding the SPG software, and to Sahil Thakar for providing technical assistance to code some of the CEMDAP software modules. Finally, the corresponding author also acknowledges the support of an International Visiting Research Fellowship and Faculty grant from the University of Sydney.

REFERENCES

1. Bhat, C.R., J.Y. Guo, S. Srinivasan, and A. Sivakumar. A Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1894*, TRB, National Research Council, Washington, D.C., 2004, pp. 57-66.

2. Bhat C.R., A.R. Pinjari, N. Eluru, R. Copperman, I.N. Sener, J. Guo, and S. Srinivasan. CEMDAP Software User Manual. 4080-P6, prepared for the Texas Department of Transportation, Center for Transportation Research, The University of Texas at Austin, October 2006.

3. Pinjari, A.R., N. Eluru, R. Copperman, I.N. Sener, J.Y. Guo, S. Srinivasan, and C.R. Bhat. Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP Models, Framework, Software Architecture and Application Results. Research Report 4080-8, Center for Transportation Research, The University of Texas at Austin, October 2006.

4. Bhat, C.R., S. Srinivasan, and J. Guo. Activity Based Travel Demand Analysis for Metropolitan Areas in Texas: Model Components and Mathematical Formulations. Research Report 4080-2, Center for Transportation Research, The University of Texas at Austin, September 2001.

5. Bhat, C.R., S. Srinivasan, J. Guo, and A. Sivakumar. Activity Based Travel Demand Analysis for Metropolitan Areas in Texas: A Micro-simulation Framework for Forecasting. Report 4080-4, Center for Transportation Research, The University of Texas at Austin, February 2003.

6. Guo, J.Y., and C.R. Bhat. Population Synthesis for Microsimulating Travel Behavior. Forthcoming, *Transportation Research Record: Journal of the Transportation Research Board*, 2007.

7. Eluru, N., A.R. Pinjari, J.Y. Guo, I.N. Sener, S. Srinivasan, R. Copperman, and C.R. Bhat. Population Updating System Structures and Models Embedded with the Comprehensive Econometric Microsimulator for Urban Systems (CEMUS). Technical paper, Department of Civil, Environmental and Architectural Engineering, The University of Texas at Austin, July 2007.

List of Figures and Tables

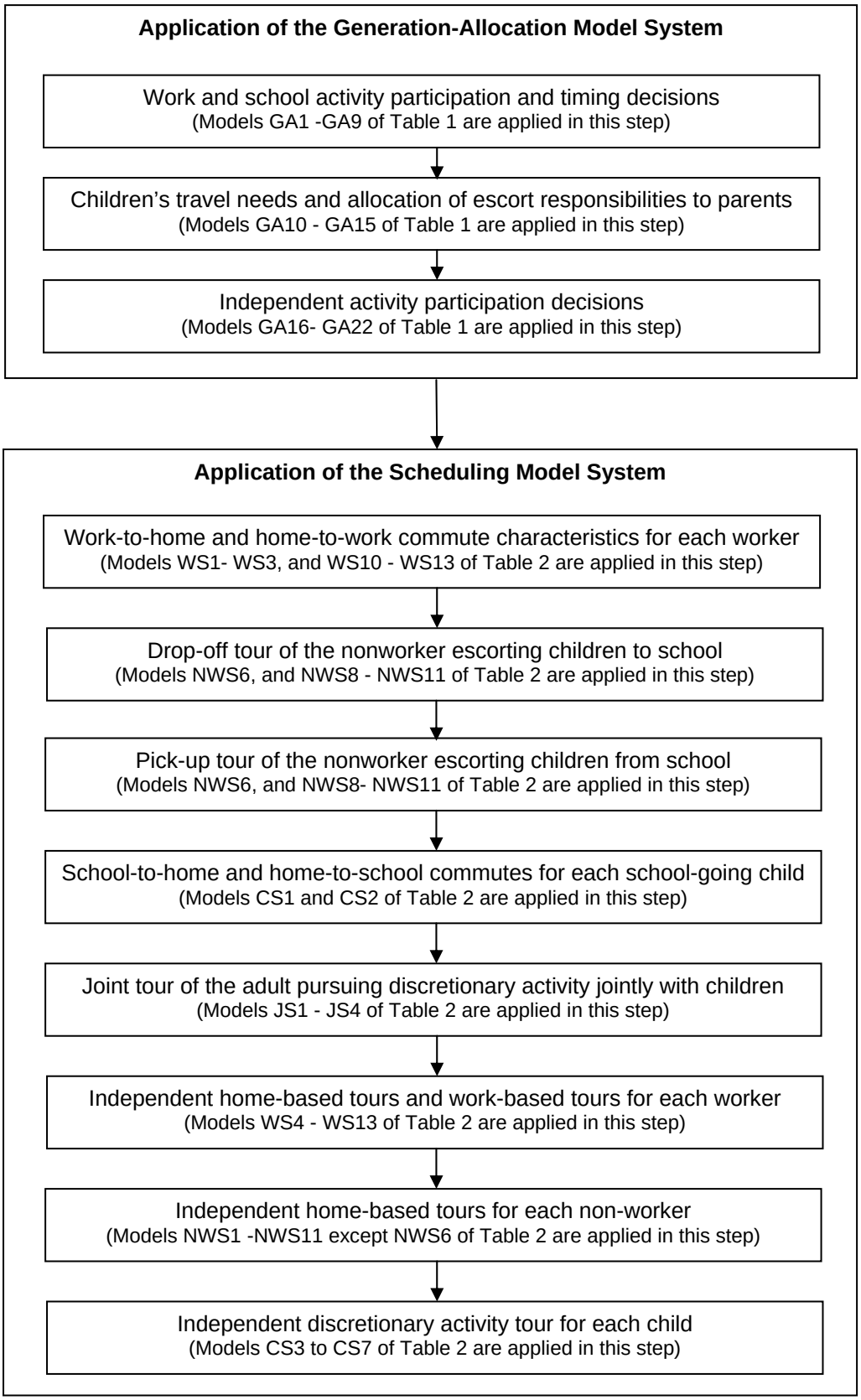

FIGURE 1 Activity-travel forecasting sequence.

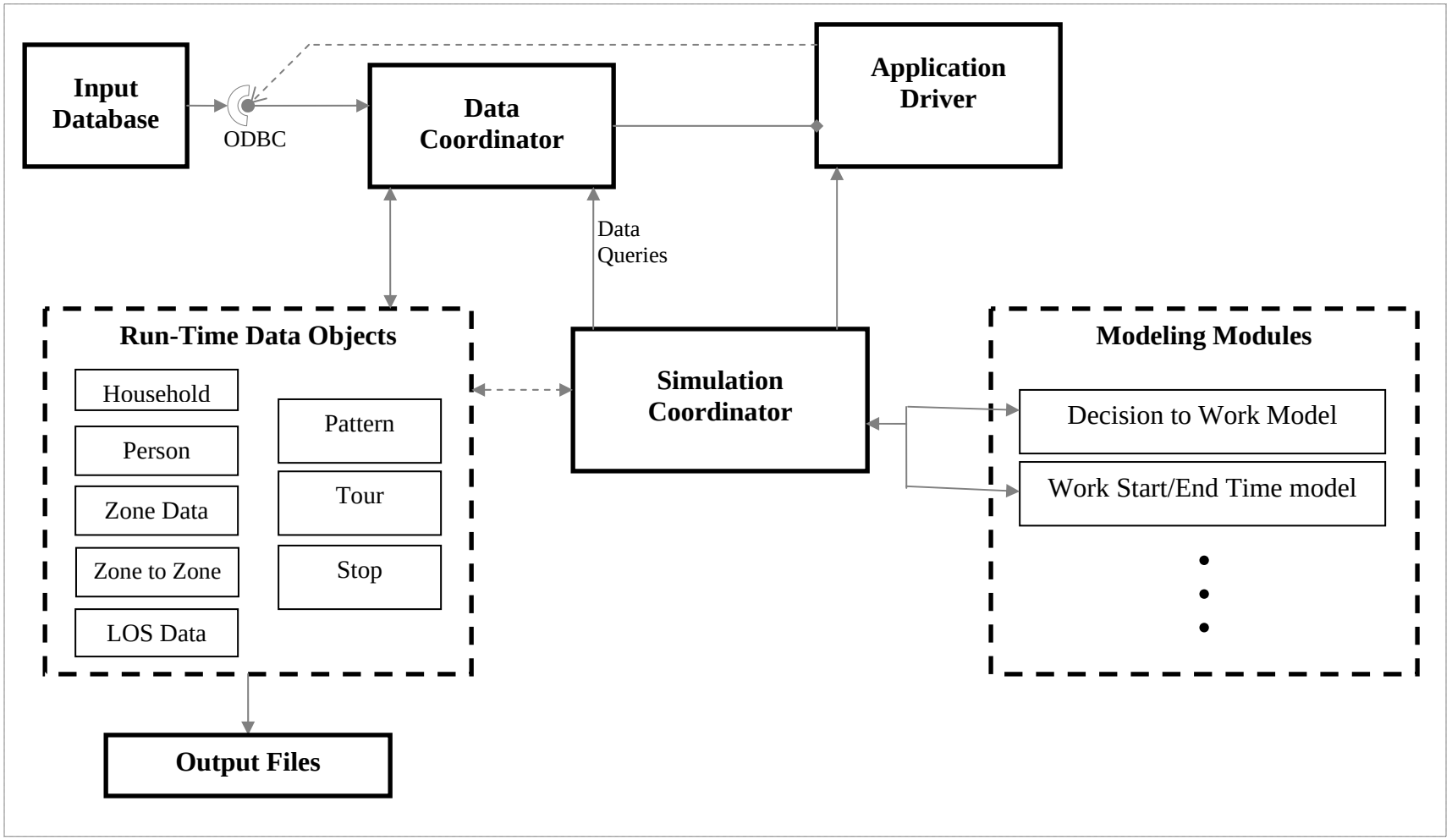

FIGURE 2 CEMDAP software architecture.

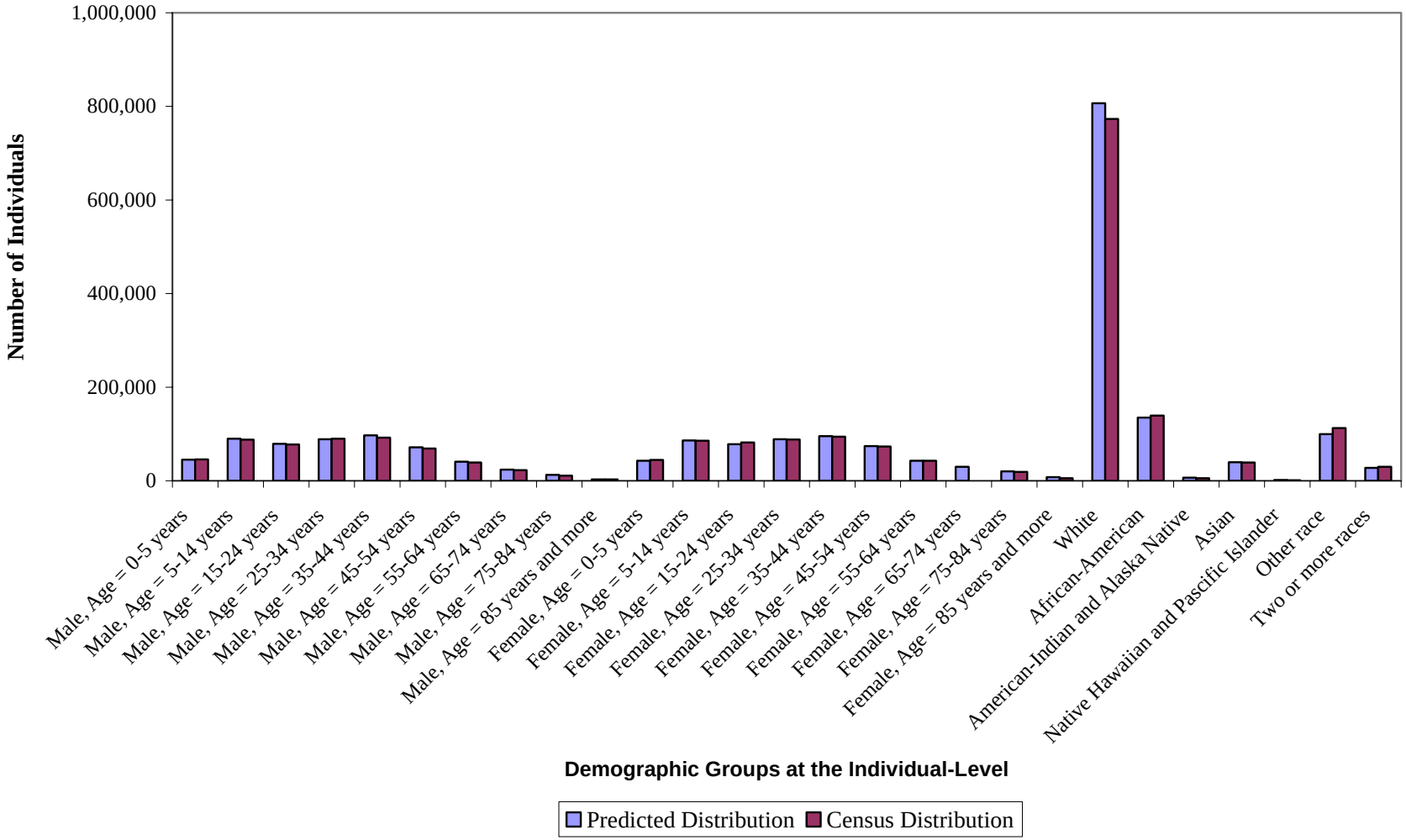

FIGURE 3 A comparison of generated and observed marginal distributions of selected socioeconomic inputs.

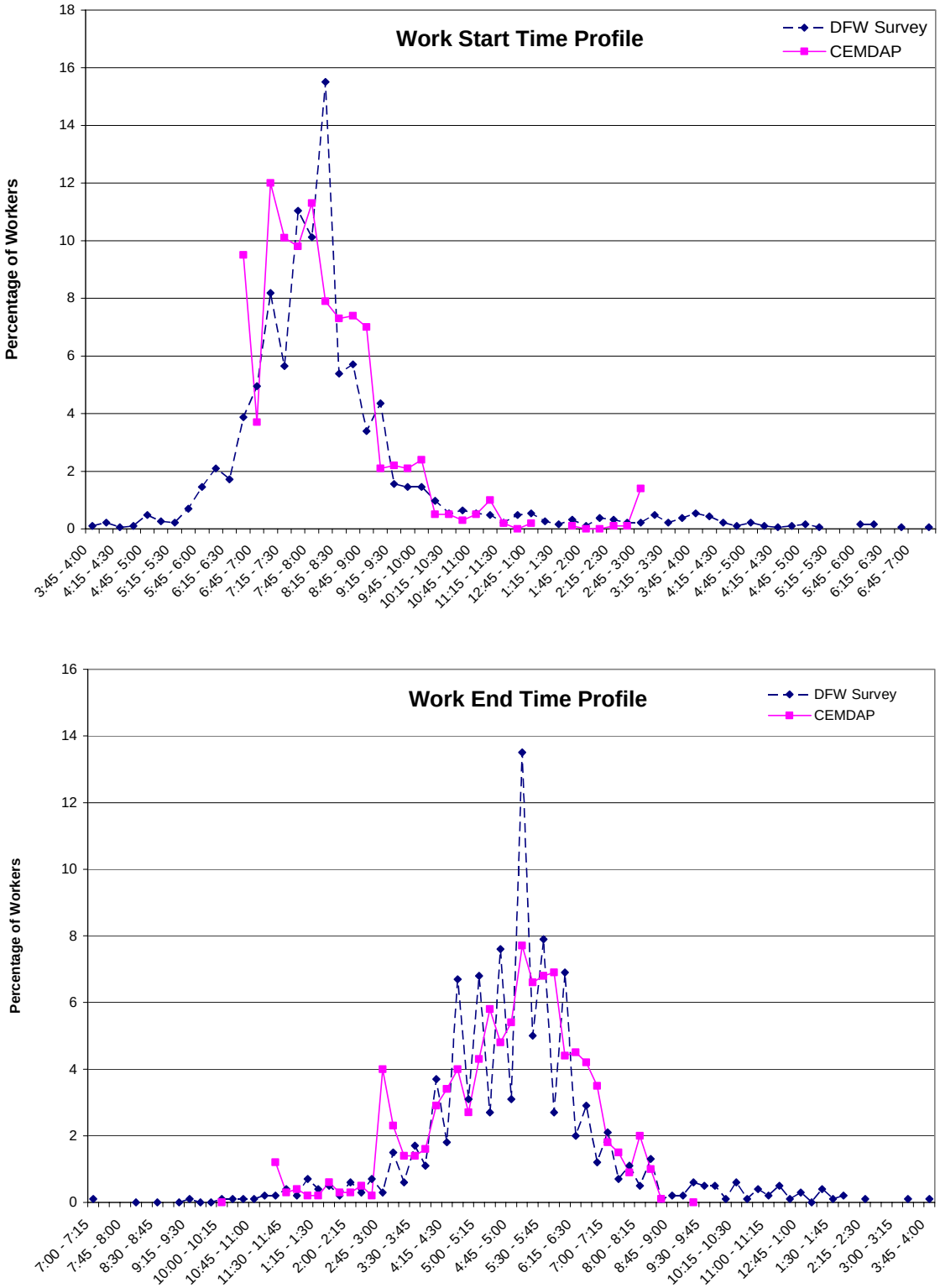

FIGURE 4 Validation against the estimation data: work start and end time profiles.

TABLE 1 The Generation-Allocation Model System

Model ID	Model Name	Econometric Structure	Choice Alternatives	Comments
GA1	Children's decision to go to school	Binary logit	Yes, No	Applicable only to children who are students. The determination of whether or not a child is a student is made in the CEMSELTS module (see Eluru *et al*. [*7*])
GA2	Children's school start time (time from 3 AM)	Hazard-duration	Continuous time	
GA3	Children's school end time (time from school start time)	Hazard-duration	Continuous time	
GA4	Decision to go to work	Binary logit	Yes, No	Applicable only to individuals above the age of 16 and who are workers. The determination of whether or not an individual is a worker is made in the CEMSELTS module
GA5	Work start and end times	Multinomial logit	528 discrete time period combinations	
GA6	Decision to undertake work related activities	Binary logit	Yes, No	
GA7	Adult's decision to go to school	Binary logit	Yes, No	Applicable only to adults who are students, as determined in CEMSELTS
GA8	Adult's school start time (time from 3 AM)	Regression	Continuous time	
GA9	Adult's school end time (time from school start time)	Regression	Continuous time	
GA10	Mode to school for children	Multinomial logit	Driven by parent, Driven by other, School bus, Walk/bike	Applicable only to children who go to school
GA11	Mode from school for children	Multinomial logit	Driven by parent, Driven by other, School bus, Walk/bike	
GA12	Allocation of drop off episode to parent	Binary logit	Father, Mother	Applicable only to non-single parent household with children who go to school
GA13	Allocation of pick up episode to parent	Binary logit	Father, Mother	
GA14	Decision of child to undertake discretionary activity jointly with parent	Binary logit	Yes, No	Second model in this row is applicable only to non-single parent households with children who go to school
GA15	Allocation of the joint discretionary episodes to one of the parents	Binary logit	Father, Mother	
GA16	Decision of child to undertake independent discretionary activity	Binary logit	Yes, No	
GA17	Decision of household to undertake grocery shopping	Binary logit	Yes, No	Second model in this row is applicable only if the household is determined (using the first model in this row) to undertake shopping
GA18	Decision of an adult to undertake grocery shopping	Binary logit	Yes, No	
GA19	Decision of an adult to undertake household/personal business activities	Binary logit	Yes, No	
GA20	Decision of an adult to undertake social/recreational activities	Binary logit	Yes, No	
GA21	Decision of an adult to undertake eat out activities	Binary logit	Yes, No	
GA22	Decision of an adult to undertake other serve passenger activities	Binary logit	Yes, No	

General Notes:

(1) A child is an individual whose age is less than 16 years, and an adult is an individual whose age is 16 years or more.
(2) CEMSELTS = Comprehensive Econometric Microsimulator for SocioEconomics, Land-use, and Transportation Systems.
(3) In the CEMDAP architecture, all individuals in the population have to be classified into one of the following three categories: (1) student (2) worker, and (3) non-student, non-worker. CEMDAP, in its current form, does not accept the category of "student and worker".
(4) GA1- GA9 model the work/school participation decisions, GA10-GA15 model the children's travel needs and allocation of escort responsibility, and GA16-GA22 model the individual-level activity participation choices.

TABLE 2 The Scheduling Model System

Model ID	Model Name	Econometric Structure	Choice Alternatives
WS1	Commute mode	Multinomial logit	Solo driver, Driver with passenger, Passenger, Transit, Walk/Bike
WS2	Number of stops in work-home commute	Ordered probit	0,1,2
WS3	Number of stops in home- work commute	Ordered probit	0,1,2
WS4	Number of after-work tours	Ordered probit	0,1,2
WS5	Number of work-based tours	Ordered probit	0,1,2
WS6	Number of before-work tours	Ordered probit	0,1
WS7	Tour mode	Multinomial logit	Solo driver, Driver with passenger, Passenger, Transit, Walk/Bike
WS8	Number of stops in a tour	Ordered probit	1,2,3,4,5
WS9	Home/work stay duration before a tour	Regression	Continuous time
WS10	Activity type at stop	Multinomial logit	Work-related, Shopping, Household/personal business, Eat out, Other serve passenger
WS11	Activity duration at stop	Linear Regression	Continuous time
WS12	Travel time to stop	Linear Regression	Continuous time
WS13	Stop location	Spatial location choice	Choice alternatives based on estimated travel time
NWS1	Number of independent tours	Ordered probit	1,2,3,4
NWS2	Decision to undertake an independent tour before pickup-up/joint discretionary tour	Binary logit	Yes, No
NWS3	Decision to undertake an independent tour after pickup-up/ joint discretionary tour	Binary logit	Yes, No
NWS4	Tour Mode	Multinomial logit	Solo driver, Driver with passenger, Passenger, Transit, Walk/Bike
NWS5	Number of stops in a tour	Ordered probit	1,2,3,4,5
NWS6	Number of stops following a pick-up/drop-off stop in a tour	Ordered probit	0,1
NWS7	Home stay duration before a tour	Regression	Continuous time
NWS8	Activity type at stop	Multinomial logit	Work-related, Shopping, Household/personal business, Eat out, Other serve passenger
NWS9	Activity duration at stop	Linear Regression	Continuous time
NWS10	Travel time to stop	Linear Regression	Continuous time
NWS11	Stop location	Spatial location choice	Choice alternatives based on estimated travel time
JS1	Departure time from home	Regression	Continuous time
JS2	Activity duration at stop	Regression	Continuous time
JS3	Travel time to stop	Regression	Continuous time
JS4	Location of stop	Spatial location choice	Continuous time
CS1	School-home commute time	Regression	Continuous time
CS2	Home-school commute time	Regression	Continuous time
CS3	Mode for independent discretionary tour	Multinomial logit	Drive by other, Walk/Bike
CS4	Departure time from home for independent discretionary tour	Regression	Continuous time
CS5	Activity duration at independent discretionary stop	Regression	Continuous time
CS6	Travel time to independent discretionary stop	Regression	Continuous time
CS7	Location of independent discretionary stop	Spatial location choice	Pre-determined subset of zones

TABLE 3 Validation Against the Estimation Data: Pattern, Tour, and Trip Characteristics

Pattern Characteristics	**DFW Survey**	**CEMDAP**
Avg. number of before-work tours (made by workers)	0.04	0.02
Avg. number of work-based tours (made by workers)	0.30	0.34
Avg. number of after-work tours (made by workers)	0.32	0.39
Avg. number of tours (made by non-workers)	1.14	1.19
Avg. number of non-school tours (made by children)	0.28	0.18
Tour Characteristics	**DFW Survey**	**CEMDAP**
Avg. number of stops in before-work tours	1.33	1.36
Avg. number of stops in work-based tours	1.31	1.27
Avg. number of stops in after-work tours	1.43	1.41
Avg. number of stops in home-work commute	0.22	0.15
Avg. number of stops in work-home commute	0.45	0.39
Avg. number of stops in non-worker tours	1.71	1.78
Trip Characteristics	**DFW Survey**	**CEMDAP**
Home-based work		
Avg. number of daily trips per person	1.79	1.70
Avg. person minutes of travel per person	27.67	26.92
Avg. person miles of travel (PMT) per trip	11.68	11.96
Avg. vehicle miles of travel (VMT) per trip	12.17	12.67
Home-based other		
Avg. number of daily trips per person	2.59	2.65
Avg. person minutes of travel per person	18.06	17.49
Avg. person miles of travel (PMT) per trip	9.38	10.72
Avg. vehicle miles of travel (VMT) per trip	9.27	11.05
Non home-based		
Avg. number of daily trips per person	2.43	2.57
Avg. person minutes of travel per person	17.78	15.15
Avg. person miles of travel (PMT) per trip	9.78	8.29
Avg. vehicle miles of travel (VMT) per trip	9.94	8.86

Note: The trip level characteristics are averaged over all individuals who made at least one out-of-home stop.

TABLE 4 Comparison with the Trip-Based Model and Observed Link Counts

Overall travel indicators in the DFW area		
Travel Indicator	**DFW Trip-based Model (Year: 1999)**	**CEMDAP (Year: 2000)**
Total person trips (Millions)	16.91	17.12
Home-based work trips (Millions)	3.91	2.74
Home-based other trips (Millions)	8.60	9.44
Non home-based trips (Millions)	4.40	4.94
Total vehicle trips (Millions)	13.42	13.35
Total vehicle miles traveled (Million miles)	121.37	140.32
Average trip length (miles/vehicle trip)	9.04	10.51
Average trip speed (vehicle miles/hour)	39.10	38.08
Predicted Traffic Volumes vs. Observed Traffic Counts (% RMSE)		
Roadway functional class	**DFW Trip-based Model (Year: 1999)**	**CEMDAP (Year: 2000)**
Freeways	21.48	25.84
Major Arterials	36.69	42.07
Minor Arterials	43.02	44.61
Collectors	70.11	70.10
Ramps	54.32	66.88
Frontage Roads	75.76	79.88
Overall	42.60	47.23

$$\%RMSE_f = \frac{1}{N_f} \sum_{\forall links} \sqrt{(Actual\ Link\ Count - Predicted\ Link\ Count)^2}\ x\ 100,$$

where N_f represents the number of links of functional class f.