Activity-Based Models and ACS Data: What are the implications for use?

D.A. Niemeier

Introduction

The Census is transitioning to the American Community Survey (ACS) for replacement of the decennial census sample. The ACS is a new continuous survey method that is expected to take the place of the Decennial Census survey by 2010. As a continuous survey, the ACS will send out about 250,000 surveys a month to households across the country (U.S. CB, 2005). The U.S. Census Bureau hopes that the ACS will provide accurate data about the U.S. population, and in a manner that will provide planners and decision-makers with information more timely information than can be achieved by the Decennial Census. The continuously collected data from the ACS will be averaged over three year periods so that the total sample size will provide a statistically accurate representation of the population. The highest geographic resolution available for the ACS data is expected to be the traffic analysis zone (*TAZ*) level. However, it is not entirely clear whether or not the sampling rate at the *TAZ* level is large enough to provide statistically accurate information. Therefore, the highest reliable geographic resolution available for most areas will most likely be at the *tract* level.

The purpose of this paper is to evaluate the ACS in terms of newer tour- and activity-based models. The paper begins with a review of the types of data currently used in these models. The paper then reviews the state of the practice with respect current tour and activity-based models. In particular, the ways in which current data are used is highlighted. From this review, a discussion of the points raised in the literature related to using census data in activity-based are reviewed with suggestions for additional research.

PUMS Data Sources

Public Use Microdata Samples (PUMS) is a database of microdata collected by United States Census Bureau (Source: Online Access: http://www.census.gov/acs/www/Products/PUMS/). They are demographic microdata available from Decennial Census (i.e., Census of Population and Housing) and American Community Survey. The U.S. Census Bureau also collects and provides microdata from Current Population Survey, American Housing Survey, Survey of Income and Program Participation and Survey of Program Dynamics.

PUMS files are online accessible files (or files provided in CD-ROMs or tapes) which show all responses made on individual questionnaires. Thus, different from summary data, they provide information about each housing unit and each individual in those housing units. Each PUMS file provides record for 1 in 1000, 1 % and 5% samples of the housing units in the United States and persons in those housing unit (Source: Online Access: http://www.ciesin.org/datasets/pums/pums-home.html) . Each PUMs file provides records and at the geographic levels described below.

- The 5% sample includes data for states and various subdivisions of states called Public Use Microdata Areas (PUMAs) which have population of at least 100,000. These PUMAs are usually based on counties. They can be whole counties, groups of counties. If a county has a population more than 200,000 persons, PUMAs can represent parts of these counties. None of these PUMAs on the sample crosses state lines.

- The 1% sample includes data for metropolitan/nonmetropolitan areas, and contains PUMAs which were made from whole central cities, complete Metropolitan Statistical Areas (MSAs) or Primary Metropolitan Statistical Areas (PMSAs) if these spatial units have populations less than 200,000. If these spatial unit have populations higher than 200,000 persons, 1% PUMAs can represent parts of central cities and MSAs/PMSAs.

The smallest spatial unit that can be represented in PUMS data files are parts of central cities or parts of Metropolitan Statistical Areas. However, it should be noted that the data set which will represent this spatial unit will be a 1 in 1000, 1% sample of the total households. The types of information collected from the PUMS data include a range of population characteristics a few of which are income, social status, education, ethnicity, household composition, etc.

ACS PUMS

The Public Use Microdata Samples (PUMS) provided for the ACS are publicly available for the years 1996 through 2003. The data provide samples from the ACS at the individual level, meaning that data is provided for individual respondents with all identifying information removed to protect the respondent's confidentiality. The PUMS data is provided without the same spatial resolution that can be obtained from the standard ACS tabulations, but may allow users not requiring high spatial resolution to observe relationships among variables that are not provided by other Census Bureau products.

The ACS PUMS data for "place of work" is provided at the *state* level within the U.S. For place of work outside of the U.S., the data is coded by the *country* where the respondent works. The

ACS PUMS data is not particularly useful to researchers requiring data provided at a geographic area smaller than the *state* level (U.S. CB, 2005).

Census STF3A

The STF3A contains data for the Decennial Census respondents who received the Long-Form Questionnaire, which is about 1 in 6 housing units nationwide. These data are publicly available for the 1980, 1990 and 2000 Decennial Census surveys. The STF3A files contain summary data in the form of counts or aggregate data in the form of means. The STF3A data files also contain nested data, meaning that data from the lower levels of geographic distinction (e.g. block groups) are aggregated and presented again at a higher level of distinction (e.g. tracts) (ATS, 1996).

The highest geographic resolution available for the STF3A data is at the *block group* level. *Block groups* are unique within *tracts*, but may be segmented by other geographic boundaries, such as *place* (ATS, 1996).

STF3A PUMS

The STF3A PUMS provides data at the individual level for about 5% of the respondents for a given area (ATS, 1996). The PUMS data is useful for researchers who do not require a high level of geographic resolution or large sample sizes, and are interested in information that is not explicitly provided by the STF3A data. The highest level of resolution provided by the STF3A PUMS is *place*. However, the data is often only provided in lower spatial resolution such as by *county* or *groups of counties* (ATS, 1996).

Census Transportation Planning Package (CTPP)

The Census Transportation Planning Package (CTPP) provides comprehensive tabulations of households, persons and workers and is a cooperative effort of State DOTs and AASHTO (Source: Online Access: http://www.fhwa.dot.gov/ctpp/about.htm). It is summarized data by place of residence, place of work, and trips made between home and work. It is also known as the "Journey to Work" Package (Source: Online Access: http://www.trbcensus.com/). The data can be used to evaluate existing conditions, develop travel demand models and analyze demographic data. The CTPP data is published for each state and county. In most cases, data is limited to The smallest spatial unit that can be represented in CTPP data files are counties. The data will represent all the population in the county, however, these will be aggregated values.

Activity-Based Modeling and Data Needs

The traditional four-step travel demand modeling processes have long dominated urban transportation planning approach, with reasonably effective applications for regional travel forecasting at the aggregate level. However, they have two main disadvantages that are well recognized and discussed by both researchers and practitioners. First, the conventional travel demand modeling method is trip-based; that is, an individual trip is the analysis unit, with the result being that the underlying behavioral causes are not well modeled. This theoretical deficiency results in a lack of interaction between trips and activities (see McNally, 2000 for details) and provides a somewhat weak foundation in modeling practice for evaluating transportation control measures (Algers et al., 2001). Second, the traditional travel demand modeling method is static and trips are typically aggregated (e.g., peak and non-peak periods).

Highly aggregated modeling results can be insensitive to finer-scale changes of traffic conditions, such as network congestion during a particular hour (Vovsha et al., 2004).

One of the ways in which the disadvantages of traditional four-step processes are being addressed is through activity-based modeling, in which a person's travel is treated as a derived demand from activity participation over time and space (Bhat and Lawton, 2002). This approach represents an advance over traditional travel demand models because it not only recognizes the interactions among a series of individual trips, but also captures the connections between trips made by different household members (RDC Inc., 1995). Since the 1970's, a number of activity-based modeling packages have been developed with varying success. Some of the more popular examples are shown in Table 1. The MPOs using these activity-based travel demand modeling systems include Portland, San Francisco County, New York City, Columbus and Atlanta (Vovsha et al., 2004).

Table 1. Activity-Based Travel Demand Modeling Packages

Model	Developer	Date Since	Model Description	Practical Application
ALBATROSS	Arentze and Timmermans	1998	A rule-based approach using decision tree formalism to model heuristic choice.	Case study for the Rotterdam region in Netherlands
AMOS	Kitamura, et al., RDC, Inc.	1995	Specific search rules are used for finding feasible schedule adjustment.	AMOS survey and TDM policy analysis applied to Washington, D.C. area (MWCOG)
AURORA	Timmermans, et al.	2001	A utility-based model of activity-travel rescheduling behavior.	N/A
CEMDAP	Bhat, et al.	1999	Modeling a series of activity-travel generations for workers (CATGW) and nonworkers (CATGNW).	N/A
COBRA	Wang and Timmermans	2000	A utility maximizing method based on nested logit model.	N/A
Day Activity Schedule Model	Bowman and Ben-Akiva	1994	Nested logit model used for estimating individual's multidimensional choice of daily activity patterns.	Portland Day Activity Schedule Model System
GISICAS	Mei-Po Kwan	1997	Scheduling an activity agenda in a GIS-based system.	N/A
HAPP	Recker	1995	Modeling household's activities based on a mixed integer mathematical program.	N/A
MASTIC	Dijst, et al.	1995	Identify action space based on a space-time prism.	N/A
MORPC Model	MORPC and PB Consult	2001	Tour-based approach for travel demand modeling using a series of choice models.	Mid-Ohio Region (MPO of Columbus) Travel Demand Modeling
PCATS	Kitamura and Fujii	1997	Maximizing utilities with considering different types of constraints.	Osaka and Kyoto, Japan
PETRA	Fosgerau	1998	Nested logit approach to model activity chains.	N/A
Portland Model	Portland Metro and PB Consult	1999	Based on Bowman and Ben-Akiva's Day Activity Schedule Model.	Portland Metropolitan Travel Demand Modeling
SCHEDULER	Garling, et al.	1989	Choosing a set of activities in the schedule based on a heuristic search.	N/A
SMASH	Ettema, et al.	1995	Sequential scheduling process based on nested logit model and Monte Carlo simulation.	N/A
STARCHILD	McNally and Recker	1986	Predict choice between alternative activity patterns based on multinomial logit models.	Case studies based on activity data from Orange County and Portland

Source: adapted and extended from Timmermans, 2001

In general, activity-based travel forecasting begins with the generation of synthetic households that are then used to populate disaggregate models. The typical process is to first create a socio-economic distribution of households by analysis zones, which are usually either traffic analysis zones or census block groups. The distribution of households by socio-economic characteristics is constructed using representative sample data such as the Census STF-3A or the 5% PUMS.

For example, one of the first tour-based microsimulation models created was for the San Francisco County Transportation Authority (SFCTA 2001). In this modeling framework, a synthesized population for San Francisco is used as direct input for models related to vehicle availability, tour generation, time of day choice, and destination and mode choice. The synthesized population relies on three primary data sources: the PUMs, socioeconomic and employment data collected by the county, and socioeconomic data compiled by the Association of Bay Area Governments (ABAG).

When populations are synthesized using census and other data, typically the way in which the distribution is established is by using what are known as control variables. For example, in the San Francisco County model, control variables are defined as household size and number of workers (9 categories), household income (4 categories), and age of head of household (3 categories), giving a total of 108 possible dimensions over which the population distribution could be synthesized. It is important to note that all of the categorizations used for the control variables are consistent with the census tables available in the CTPP.

The use of control variables is common. Other activity-based models using control variables to establish the baseline population distribution include Transims (which reflects the Portland metropolitan region), the Atlanta Regional Commission model, the Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns (Bhat et al 2003) and the Mid-Ohio Regional Planning Commission model (PB Consult 2005).

The control variables establish the marginal distributions and an interative procedure for fitting the proportions is used to derive the number of households in each of the cells. Most activity-based models using synthesized households start with a distribution supported by aggregate census units (e.g., the Public Use Microdata Areas). Once a joint distribution is established, an allocation to TAZ or census blocks is performed, and for each cell a sample of households is drawn randomly from census data, usually the 5% PUMS. The types of variables used for controls include income, household size, number of workers, children, gender, and race at the household level and age, race (e.g., Hispanic versus not), and employment at the person level.

To understand how census data are used in the basic activity-based modeling, it is useful to look at a couple of fairly recently developed population synthesizers. By way of a first example, consider the MORPC travel demand model (PB Consult 2005). In this model, the population is synthesized by TAZ using socioeconomic variables relevant to the travel model and average zonal characteristics. The input parameters include total population, total number of households, total labor force, and average household income. The output variables associated with each household include household income group, status of adult household members (e.g., full time workers, part-time workers, etc), and household children by age category.

The population is synthesized in two steps (Figure 1). First, a distribution of households is created using a set of controlled variables represented as either zonal totals or averages provided as input from the land use, socio-economic modeling. This distribution is then converted to a "list" of household with a set of additional variables added to the characteristics; the additional variables are uncontrolled in the sense that each synthetic household is matched to an identical household in PUMS.

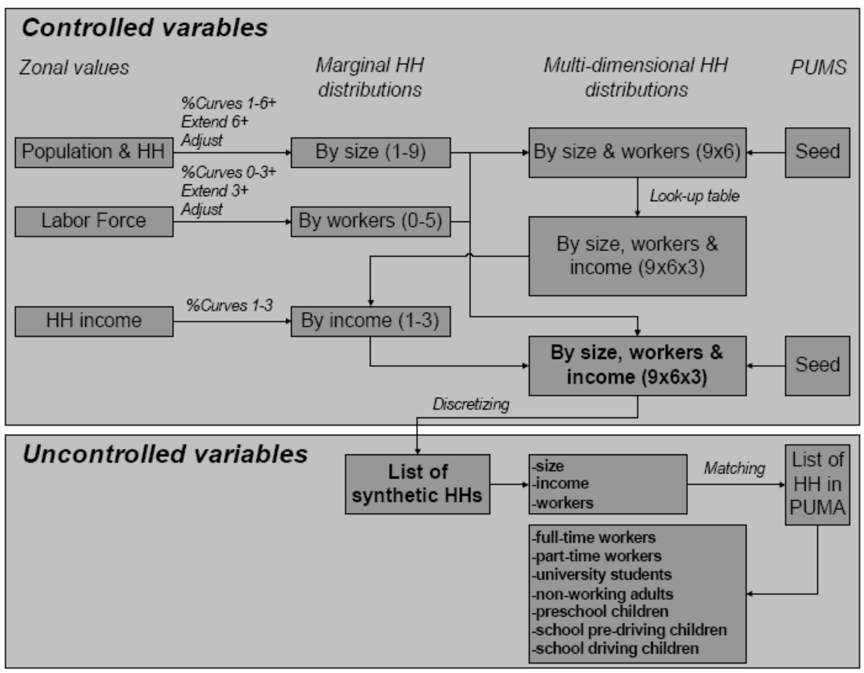

Figure 1. MORPC Population Synthesis Steps (*Source: PB Consult 2005)*

In the first step, the PUMS data is used to provide a multi-dimensional seed distribution of households in each PUMA. In the second step, the uncontrolled variables are added by identifying similar households from PUMS corresponding to the dimensions of controlled variables and adding the relevant uncontrolled variables. This same procedure is undertaken for all households in the PUMA.

The CEMDAP model uses somewhat the same approach for generating the household population. To see how census data are currently used, consider the derivation of “medium-term” activity travel patterns (Bhat et al 2002). For this portion of the activity-based modeling, the PUMA was the Metropolitan Statistical Area, groups of MSAs, parts of MSAs for metropolitan areas larger than 100,000 and groups of non-metropolitan areas.

The estimation of middle-term activity patterns relies on several data sources: an activity survey, Census files, and the PUMS and none of the data provide the full range of information ideally needed for the modeling. The census data (SF1 and SF3) provide only marginal population distributions. This requires that an additional algorithm is available to estimate the eventual multi-dimensional population distribution serving as the basis for the synthesis. The Iterative Proportional Fitting (IPF) process is used to create the multi-dimensional population distribution.

IPF uses a seed matrix and marginal totals (constraints) to derive cell estimates. If the constraints are internally consistent (i.e., all sources sum to the same totals), the IPF procedure will converges to a unique solution. As Bhat et al note, when data is culled from a number of different sources with varying levels of aggregation (or category collapsing), the assumption that row-column totals will equal becomes a limiting factor in the method’s application. Nonetheless, the method is widely used to develop the multi-dimensional population distributions used in the population synthesis process.

In the CEMDAP application, the IPF is used to produce output file for the 1032 census tracts in the 11-county Dallas Fort-Worth region (Figure 2). For each census tract, a synthesized household distribution by household type is produced, where household types are those defined in the PUMS. For example, the output file for county 113, tract 1 will comprise 7844 household types (rows) corresponding to the households in the PUMS data from county 113, tract 1. Once these files have been established, a second step is performed, which expands the data so that the representative number of households are produced for each household-type distribution and assigning each household to a traffic analysis zone. This step is required because most travel model zone systems will differ from those used by the census. In the case of Dallas-Ft. Worth (the primary application of CEMDAP), there were 919 traffic analysis zones and 1032 census tracts (87,086 blocks). For this second step, the distribution files produced in the first step were needed, and the PUMS person files by county, the Census 2000 tract and block level data and a block-traffic zone equivalence map were required.

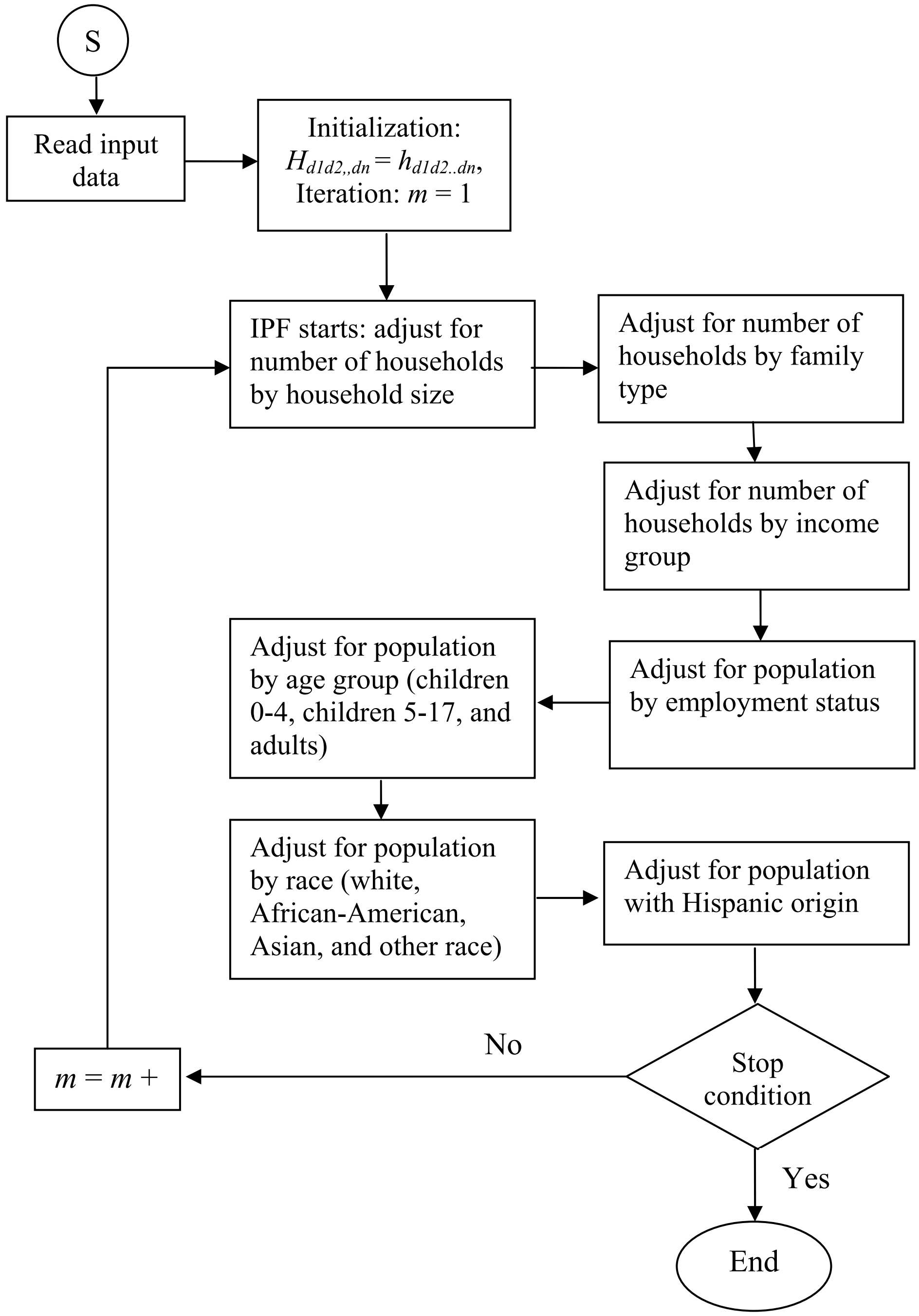

Figure 1. CEMDAP Population Synthesis Steps (*Source: Bhat et al 2003b)*

Spatial Resolution of ACS Estimates

Issues associated with small area samples in the ACS have been discussed in a number of papers. In this section, the focus is on synthesizing the major findings of previous work and linking this work to potential issues associated with small area estimation associated with the ACS. Perhaps the most relevant of these more recent studies related to small area estimation issues with the ACS, is a comparison between the ACS estimates and the Census 2000 estimates.

Estimates were compared across 36 counties and associated census tracts, with the exception of two counties in Texas: Harris and Fort Bend. The ACS estimates represent a three-year average (1999-2001) while the Census data were collected in 2000. It is important to note that the compared estimates do not include basic count data since total population and total number of housing units is used to control the ACS population estimates (at the county level only) and likewise, 100% weighted Census counts (at roughly equivalent to a census tract) are used to control sample estimates.

To compare between the ACS and Census 2000 sample data estimates, standard errors were used to statistically evaluate differences. At the county level, a z-score with a Bonferroni adjustment was computed and evaluated at the 90% confidence level. At the tract level, calculations were made without the Bonferroni adjustment. Note also that larger counties will be more likely to have greater statistically significant differences than smaller counties simply as a result of sample sizes.

In Table 2, a summary of the major findings in Report 8 are presented. The results of the comparison indicate that some of the Census variables used to date for activity-based modeling may be significantly different from those produced through the ACS. Looking at sex, the results indicated that small differences existed between the Census 2000 estimates and the ACS estimates, and those differences that did exist were for all practical purposes unimportant (i.e., <1%). It is important to remember that many of these variables are also used in the ACS weighting process, and age is one such control variable. This means that the ACS estimates are forced to agree with census information unless categories are collapsed. So for control variables, we would not expect to see big differences.

Significant differences in county level estimates were found for race. Although race is used as a population control variable, because of category collapsing the totals are not likely to match. Most of the differences were attributable to responses for Hispanic. No significant differences were found for average household size, but both total household population and total housing units are control variables so it is unlikely there would be major differences. In terms of the employment categories, ACS and Census estimates were significantly different. ACS estimates were higher for the in-labor force estimates and Census estimates were higher for the not in labor force estimates.

In terms of household income, ACS estimates tended to be higher for income categories below $50,000 and Census estimates tended to be higher for income categories greater than $50,000. However, for practical purposes the differences in the estimates were rarely largely than 2%. In evaluating household size, ACS estimates were generally higher than Census estimates (e.g., averaged over all counties, Census household size was 3.14, ACS was 3.16).

More half of the counties had statistically significant differences in mean travel time to work estimates. In nearly all counties, the Census 2000 estimates were higher than the ACS estimates, with the largest differences appearing in the range of 20 to 30 minute commute times. In terms of commute mode, all modes showed significant differences between the census and ACS estimates. Census estimates were generally higher for the carpool to work mode while ACS estimates were higher for all other modes. In terms of vehicle availability, ACS estimates were higher for the no vehicle category and Census estimates were higher for the 3+ vehicle categories. There was no statistical difference found for the 2-vehicle category.

The report also evaluated a limited number of transportation relevant variables at the tract level. Before reviewing these findings, however, it is important to understand how the ACS data at the tract level are weighted. In producing tract estimates, the Census data is weighted at the tract level (or sometimes groups of tracts depending on the sample size). In contract, the ACS data are weighted at the county level. Obviously, this difference in weighting may be more likely to produce larger differences when comparing tract estimates between Census and ACS, but these differences may be less likely to produce a statistically significant difference because sample sizes are much smaller (and variability much greater).

To compare ACS estimates and Census estimates at the tract level, tracts were divided into five groups by population (for greater detail, see Report 8), and some of the observations are interesting. For race, the differences in tracts level estimates generally reflect those found at the county level.

For mean travel time to work, a general trend of higher Census estimates was observed, although very few of the differences were actually statistically significant. The Census was significantly higher for the mid-range household income levels ($25k to $75k); for the remainder of the categories, the two estimates were similar.

Table 2. Travel Model Related Findings U.S. Census Bureau: Report 8 (2004)

Variable	**Major Findings**	
	County Level (Practical Difference)	**Tract Level**
Age°	ACS ≈ Census 2000	
Race°	ACS ≠ Census 2000	ACS ≠ Census 2000
Household Size	ACS ≈ Census 2000	
Employment Status	ACS ≠ Census 2000	
Household Income	ACS ≠ Census 2000 (<2%)	ACS ≠ Census 2000
Household Size	ACS ≠ Census 2000	
Mean Work Travel Time	ACS ≠ Census 2000	ACS ≠ Census 2000
Commute Mode	ACS ≠ Census 2000	
Vehicle Availability	ACS ≠ Census 2000	

° ACS Control Variable

Overall Comment on ACS PUMS for Travel Modeling

There is little doubt that PUMS and Census data plays a key role in activity-based modeling, most prominently through the household synthesis process. In the using these data, there are hierarchies of zonal systems – mostly constructed as a function of the available data - that demand modelers have accommodated in development of models. For example, consider the San Francisco model again. The households were synthesized using a data hierarchy that went from six PUMAS to 127 Metropolitan Transportation Commission traffic analysis zones to 766 San

Francisco Traffic Analysis Zones. And in the Dallas-Ft. Worth models, there are a total of 1032 census tracts enclosed by 919 transportation analysis zones.

Bhat et al (2003b) identified three major practical issues associated with using the census or PUMS data for household synthesis: the geographic resolution associated with the zone system, the need for demographic information at the household and person-level, and finally the issue associated with a starting point for the IPF process.

In terms of the geographic resolution, recall that census data are available at the county, tract and block levels. When traffic analysis zones and census tracts are roughly comparable, then socioeconomic data is readily available. However, households must then be reallocated to TAZs. One approach is to estimate a location model that is constrained to the census tract where the household resides. However, very few data sets contain both the TAZ and census designation.

The second general problem identified by Bhat relates to whether the household synthesis can be performed at the household level or the population level. Ideally, the analysis is conducted at the household level retaining information about individuals within the household. In the CEMDAP application, the IPF is conducted at the household level using the PUMS sample. Once households have been created, individuals are sampled within households again from the PUMS data.

Finally, when a household distribution is created, a starting seed distribution is necessary. If the census tract is the geographic unit, then ideally the seed distribution at the census tract should be the starting point as well. However, PUMS data can only identified at the county level so in essence the seed distribution is taken at the county level.

Most of these basic constraints can be overcome in a practical setting, but almost certain result in errors. Although there have been substantiated issues related to small area estimation using the ACS, these issues do not seem any more likely to hinder transportation analysts than currently used data. Perhaps of equal or greater concern is that key ACS variables defining households characteristics may not match census information, and in fact become less reliable as the spatial resolution increases. This could present a significant issue in terms of synthesizing households that are reasonable approximations of reality.

Additional research on the effect of the control variables and variables that are statistically significantly different from census information on household synthesis would be very useful. It should also be noted that other aspects of the ACS may present difficulties for travel demand models, however, these difficulties (e.g., workplace location) will be present regardless of the type of model used. Travel forecasters have already adapted methods for using more aggregated PUMA information and it is likely that these methods will continue to be applicable when using ACS PUMS. One of the issues that will have to be addressed for unique region is whether to use census blocks/tracts for household synthesis or county estimates. At the block/tract level ACS PUMS may have multiple collapsing of categories due to small sample sizes. This, in turn, can impact the socioeconomic characteristics both for the seed distribution as well as the simulated households. One line of research that might be useful here is the development of quality measures that assess the use of ACS data for multiple category collapsing. For example, Alexander (2001) computed standard errors compared to census standard errors. The ACS computed standard errors were approximately 1.33 times as large as those of the census for domains with average response rates. However, this estimate can differ greatly depending on the domain and non-response. If the quality measure (perhaps based on the coefficient of variation or standard errors) is relatively straightforward to compute, it would allow individual regions to assess the impact of using various spatial resolutions in their activity-based models, most prominently the household synthesis process.

REFERENCES

Academic Technology Services (ATS) (1996). *Understanding the Census STF3A File For California*. Training and Consulting, Statistical Computing; University of California, Los Angeles. October, 1996. Available [online]: http://www.ats.ucla.edu/stat/census/census1980_1990/1B.htm

Alexander, C., *Still rolling: Leslie Kish's rolling samples and the American Community Survey*, Proceedings of the Statistics Canada Symposium, Achieving Data Quality in a Statistical Agency: A Methodological Perspective, 2001.

Bhat, C., Sivaramakrishnan Srinivasan, Jessica Y. Guo, *Activity-Based Travel Demand Modeling for Metropolitan Areas in Texas: A Micro-Simulation Framework for Forecasting*, Center for Transportation Research, University of Texas, FHWA/TX-03/4080-4, 2003.

Bhat, C., Sivaramakrishnan Srinivasan, Jessica Y. Guo, Aruna Sivakumar, *Activity-Based Travel Demand Modeling for Metropolitan Areas in Texas: Data Sources, Sample Formation, and estimation Results*, Center for Transportation Research, University of Texas, FHWA/TX-03/4080-3, 2002.

Bhat, C. R., Zhao, H. and Sivakumar, A.. *Synthetic population generation for micro-simulation activity-based travel demand modeling systems*, Technical Document, University of Texas at Austin, 2003b

PB Consult, *The MORPC Travel Demand Model, Validation and Final Report*, Mid-Ohio Regional Planning Commission, Final Report, Feb 2005.

San Francisco County Transportation Authority (SFCTA), *Transportation Modeling System, Overview and Summary*, San Francisco County Transportation Authority, June 2001.

U.S. Census Bureau, *Meeting 21st Century Demographic Data Needs – Implementing the American Community Survey, Report No. 8, Comparison of the American Community Survey Three-Year Averages and the Census Sample for a Sample of Counties and Tracts,* U.S. Department of Commerce, U.S. Census Bureau, Washington DC, 2004.

U.S. Census Bureau (U.S. CB) (2005). American Community Survey (website). Available [online]: http://www.census.gov/acs/www/