Transportation Research Board

National Household Travel Survey Conference: Understanding Our Nation's Travel

Resource Paper

Household Travel Survey Data Fusion Issues

Mohan Venigalla
George Mason University

Introduction

Transportation policy and planning studies require data on travel behavior are often obtained from travel activity surveys. In the 40-year span since the early 1960's, when systematic surveying of travelers has begun, response rates for travel surveys have dropped from over 90% to about 30% (MWCOG, 2003.). At the same time, the cost of conducting such surveys has sky-rocketed. With declining response rates for almost all survey approaches, there is a need to seek alternatives to travel surveys. One of the avenues of considerable promise includes supplementing the survey data by integrating with readily available data from other compatible sources.

In this context, the term *data fusion* refers to the process in which two or more databases are integrated so as to obtain necessary parameters or a single database. These parameters may then used for such purposes as travel demand modeling. In this paper issues related to data fusion concerning travel survey and other related databases are discussed. Specific focus of this resource paper is on integrating the Nationwide Household Travel Survey (NHTS) data with other data sources. The following list contains the broad categories of data that may be integrated for analytical needs.

- Household (local) Travel Survey (HTS) data
- Specialized travel survey data (e.g. airport travel and pedestrian surveys)
- Nationwide travel survey data (e.g. NHTS)
- Census demographics
- Public Use Micro-data Sample (PUMS)
- Census Transportation Planning Package (CTPP)
- Transportation network data
- Geographic Information System (GIS) databases

Each of these datasets are designed to serve the needs of a select list of analytical and policy needs. For example, the transportation network data are maintained by MPOs and local jurisdictions vested with the responsibility of developing and applying travel demand models for short-and long range planning needs of the region. These networks consist of a set of zones, nodes and links. Usually zones represent geographic regions with homogeneous demographic characteristics, nodes represent street intersections and zone centroids, and links represent road segments between a pair of nodes. While the network data provides a skeletal structure for performing travel demand modeling, travel survey and demographic data play a pivotal role in meeting the needs of populating the network link with vehicular traffic. The data needs travel

demand models are not only extensive but also complex in terms of level of detail, coverage and accuracy. Often, the TDM tasks require information from a variety of sources.

Other analytical needs requiring integration of data from two or more sources mentioned above include the following:

- Analyzing the state of the nature of the transportation system
- Measuring the performance of the transportation infrastructure, and
- Studying the effectiveness of the transportation investments

Some situations requiring data integration include the following:

- Travel survey data and GIS database are integrated to thematically map travel characteristics of the population. The same can be said about integration of census demographics data and GIS.
- Local and national travel surveys are integrated to bridge data gaps in one or both databases. For example, when HTS data lacked information on long-distance trips, appropriate data elements from the NHTS data may be borrowed.
- Vehicle registration data from state departments of motor vehicles (DMV) may be used in conjunction with NHTS vehicle usage data for developing inputs to emission factor modeling.

When multiple data sources are integrated, the utility of the data may be enhanced in the form of expanded coverage or availability of additional data elements or both. However, data integration cannot fix sampling, measurement and processing errors, as well as non-response and coverage biases that are inherent with many survey data.

Data Sources and Their Use

Household travel surveys deal with collecting information on travel patterns at household level. These data are usually collected by local jurisdictions mainly for supporting the travel demand modeling needs. Special purpose travel survey data are collected to supplement household travel surveys. These include such surveys as urban freight movement survey and special generator (e.g. airport) surveys. Both HTS and special purpose travel surveys are conducted at regional level. The intended users are mainly local planners and policy makers.

The National Household Travel Survey (NHTS) provides data on the full continuum of personal travel in the United States. For the 2001 NHTS, information on daily trips, as well as long distance travel of 50 miles or more was collected. It is the only national source of information on the typical travel of people in the country, and includes valuable information such as:

- amount of travel by purpose and mode
- time and miles spent traveling
- automobile ownership and use of the vehicle fleet
- affects on travel due to household composition and other characteristics

These data allow policy makers, analysts and researchers to accurately describe travel behavior, assess how it is changing over time, and make informed choices regarding transportation planning and policy. The 2001 NHTS is a joint effort funded by two agencies within the

Department of Transportation: the Bureau of Transportation Statistics (BTS) and the Federal Highway Administration (FHWA).

The congressionally mandated decennial census data are useful not only for distributing congressional seats, but also for a vast array of policy studies. In collaboration with a number of federal and regional governments, the Census Bureau packages the census long-form data in the form of Census Transportation Planning Package (CTPP). The CTPP data are used by various transportation agencies for planning and policy studies. Agencies use CTPP as the primary source for journey-to-work data.

There are many policy, planning, and modeling issues for which NHTS data explains only part of the story, and data from other sources are necessary to complete the picture. The primary objective of this paper is to identify opportunities for data fusion using NHTS data as well as examine the challenges and solutions for integrating one or more databases with NHTS data.

An Overview of Applications of NPTS/NHTS

Presented in this section is a list of studies that employed NPTS/NHTS data with specific emphasis on integration with other data sources.

Bricka [2] used 1995 NPTS data and 1995 ATS data to identify the regional variations in long distance travel in the United States. The data related to long distance travel categorized on the basis of trip purpose, travel mode, household size, household income, household race, and worker status were compared for New York, Massachusetts, and Oklahoma. The results of the study showed that the NPTS and ATS data sets differ in terms of income distribution, overall, and household race specifically for families in Oklahoma. Also, the travel pattern of the respondents from Oklahoma clearly differs from those in New York or Massachusetts.

Zumd and Arce [3] used NPTS, ATS, Consumer Expenditure Survey (CES), and U.S. Bureau of Census to examine the intersection of consumer culture and travel behavior. The data is categorized on the basis of race/ethnicity. The data related to person trip by travel purpose, race/ethnicity distribution, volume of person trips by travel purpose, travel mode for shopping, and trip start time are taken from 1995 NPTS data set. The data related to long distance travel by race/ethnicity are taken from ATS data set. The data related to race/ethnicity distribution and median income by race/ethnicity are taken from U. S. Census bureau data set and the source of data for average annual expenditures by race/ethnicity is CES.

Hu and Young [4] examined the issues of combining NPTS and ATS data. The two data sources were compared for number of data categories and differences were identified. The study found that after removing as many differences between the two surveys as possible, the NPTS and ATS basically drew their samples from a similar population with respect to age, gender, geography, education, and household size. However, the results from simulation study confirmed that limiting data collection period of two weeks in case of NPTS data definitely underestimates the overall extent of long distance travel in the US. Nevertheless, it was found that after appropriately weighting the sample data, both ATS and NPTS produce overall travel statistics that are not identical but that are at least similar. The study recommends that daily trips data from NPTS and the long distance trip data from ATS data set can be combined to portray the overall trends in personal travel.

Niles and Nelson [5] describe the usefulness of NPTS data set along with other national surveys for developing a realistic and cost-effective land use and transportation strategies at national and regional level. The usefulness of other national surveys discussed in the study are; American Housing Survey, Fannie Mae Housing Survey, Survey of Construction, Housing Starts by Design, Consumer Expenditure Survey, and Economic Census. The study suggests that if the NPTS with other national surveys described earlier are coordinated and cross-fertilized properly could provide a better understanding of consumer preferences, industry location decision, and household activities that determine travel pattern.

Pratt [6] used NPTS data set along with other surveys to gain insight into the impact of technology and telecommunications on telework and trip reductions. The study uses Current Population Survey (CPS), NPTS, American Housing Survey (AHS), and CPS Computer Supplement and Cyber Dialogue Internet Survey. The CPS relates job classification and occupation to work at home and frequency. The NPTS contributes trip purpose and distinguishes work related trips from all others. The AHS relates frequency of work at home during normal business hours to job classification, and finally CPS Computer Supplement and Cyber Dialogue Internet Survey suggest the impact that use of the internet by teleworkers may have on travel.

McGukin et al. [7] used the 1990 and 1995 NPTS data sets along with Public-Use Micro Data Set (PUMS) to determine the characteristics of households whose telephone service had been interrupted with characteristics of households with no telephone service. In the analysis the two groups; PUMS nontelephone households and NPTS interrupted telephone households were categorized into the general grouping of no workers in the households, one worker, two workers etc. The goal was to test the hypothesis that the percentages from the two groups were from same population.

A report [8] prepared by a TRB committee on school transportation safety related to relative risk of school travel used NPTS data along with Fatality Analysis Reporting System (FARS), and National Automotive Sampling System (ASS) to estimate the relative risk to children traveling to and from school and to and from school related activities. The data derived from NPTS include number of trips taken and miles traveled by school age children for all modes, the data from FARS include police related fatal crashes, and data derived from GES contains data on a nationally representative stratified samples of police reported traffic crashes that occur on public roadways in various geographic sites in the US and the result in property damage, injury or death.

Rosenbloom and Waldorf [9] used 1995 NPTS data to determine whether the mode choice of White, non-Hispanic elders differ from otherwise comparable ethnic and racial elders. Two models were used to understand the effects of race and space on mode choice of the elderly. The dependent variable in the first model was privately owned vehicles and in the second model was public transportation, while the independent variables in both model were ethnicity, race, location, HH income, gender, public transit availability, trip purpose, and driver status. The study found that racial minorities are less likely to go by car and more likely to choose public transportation. Also, location is found to be the most dominant covariate of mode choice. The residents of urban areas are likely to travel by public transportation than by car. Moreover, Hispanics are significantly less likely to choose public transportation than non-Hispanics.

Evans [10] used 1995 NPTS data to identify those personal and community characteristics that are associated with trip making among the non-driving 75+ population. The

study suggests that beyond the constraints of physical and economic well being, the housing density and community context mostly influence mobility among the non-driving 75+ population. When the housing density factor is controlled in the analysis living in a central city area appears to be negatively associated with mobility in the age group of 75+ non-driving population, which suggests that perceived safety may influence mobility among this age group.

Gardenhire and Sermons [11] have used 1995 NPTS data samples to study automobile ownership models of residential location choice for poor and non-poor households (HH). The study is done to determine whether the automobile ownership choice behavior of low-income HH is significantly different from that of middle and upper class. The results of the study showed that poor HH convert income to automobiles at twice the rate of non- poor HH. Also, poor HH ownership is more sensitive to residential density than non-poor HH behavior.

Dill [12] used 1995 NPTS data set to derive information regarding the older vehicles in use. The study attempts to find out household that own older vehicles and how they used them. The vehicles built before 1981 are of particular interest. The NPTS data about older vehicles was used to get insight into the impacts of voluntary accelerated vehicle retirement (VAVR) program on older vehicles. The older vehicle data derived from NPTS is based on regional differences, income, race, urban Vs rural, household composition. Also data about VMT based on self Vs odometer reading, and household characteristics is used. The study showed that data derived from NPTS were useful to answer questions like, potential participants of VAVR program, population who could be impacted, older vehicle contribution to pollutant emissions, and effectiveness of VAVR program.

Erlbaum [13] discusses the usefulness of the 1995 NPTS undertaken by New York State DOT (NYSDOT). In this study various analysis of NPTS are done to determine the patterns and characteristics of current and future travel in the state. The 1995 NPTS conducted by NYSDOT collected such data that were not only useful in addressing state transportation issues, but were also valuable in addressing issues related to values used in EPA's MOBILE emission model. The 1995 NPTS data collected by NYDOT also included questions related to telecommuting Vs work at home, hourly vehicle distribution, area wide speeds, vehicle use, and engine mode of operation. The study suggest that the 1995 NPTS-NY provides a valuable county data set which addresses a variety of transportation related questions.

Polzin et al. [14] used 1983, 1990, and 1995 NPTS databases to analyze mobility and mode choice of people of color for non-work travel. The non-travel work includes travel for personal and family business, school activities, religious activities, health care, and social recreational activities. In this study mode choice differences across groups are analyzed by examining how patterns of difference in mode choice vary with personal, household, geographic and trip characteristics as reported in 1995 NPTS. The analysis suggested that the difference in non-work travel behavior for the various racial/ethnic groups has changed dramatically over time with minority travel behavior matching more closely with majority travel behavior.

Ziliaskopoulos and Waller [15] introduces the development of an internet based geographic information system (GIS) for bringing together spatio-temporal data, models and users in a single efficient framework to be used for a wide range of transportation applications such as, planning, engineering, and operational. The functional requirements for such a system are identified in terms of interface and user connectivity needs, database integration needs, and modeling needs. To meet the functional requirements of the system and bring together data,

models, users and applications into an efficient system a framework called visual interactive system for transportation algorithms (VISTA) is used. The framework is based on COBRA specification that allows the modules to be written in a separate programming languages, and to be run on different machines over a network.

In an outline of a methodology for the creation of a synthetic baseline population of individuals and households for use with activity-based travel models, Beckman et al. (1996) have discussed techniques for locating households from census data.

Data Integration Challenges

Data fusion poses various challenges, which mainly depend on the application as well as the resources available at the disposal of the agency seeking data integration. First and the foremost challenge pertains to the availability of data itself. That is, pertinent data may or may not be readily be available for fusion considerations. Datasets related to national level survey data may easily be obtained from such centralized repositories of data as the US Census Bureau and the Bureau of Transportation Statistics (BTS). When an agency attempts to integrate local data with well publicized nationwide datasets, the task of obtaining data is relatively simple. It is conceivable that, for many applications integrating available local data at different geographic regions may be an appropriate measure. Such opportunities may not be availed if the data are not readily available. If the data sources are not well publicized, these opportunities may not even be recognized by the potential users.

Significant differences may exist among datasets in terms of survey scope, content and coverage. In order to reconcile these differences in the data integration process, various statistical tests may have to be performed on individual and combined datasets. In many cases resolving statistical issues related integrating data sets from disparate cross-sectional characteristics is not trivial. For example, when integrating a 5% sample data with 1% sample data the analysts should carefully consider maintaining the integrity of weights (such as household weights) in the integrated data.

One of the most important challenges occurs when identical data elements are present in the data sets that are being integrated. It is to be expected that results of analysis of these identical data elements may be considerably different from each other. In such cases, integrating these data sets may require normalization of the common data elements along with necessary weight adjustments.

A Generic Approach to Integrating NHTS Data with Other Datasets

Data fusion methods vary with analytical needs as well as with the content and extent of data sources. In order to combine data from multiple sources, at a minimum, the details of two data sources must be comprehended. While it is impossible to develop a *one-size-fits-all* methodology for data fusion, in this section a generic approach to integrating NHTS data with other data sources is presented. This approach is flexible enough to apply to a majority of situations that require integration of NHTS data with one or more other datasets from other sources.

Step 1. Identify appropriate databases and the data elements.

If all the data elements can be obtained from a single data source such as NHTS or CTPP, there is no need to integrate the data. Data fusion may be considered as an option if one of the data sets contains required elements but lack the necessary sample size or spatial resolution. The term data

elements refers to variables or attributes that are within the database. For example, in order to develop trip production rates cross-classified by income level at TAZ level, at a minimum the number of trips by household and household income for each of the TAZ's are needed. Appropriate data tables elements in NHTS data for this purpose will be household income and trip purpose from the person trip table. It is possible the same data elements may be available form a variety of data sources. For simplicity, number of data sources may be kept to a minimum.

Step 2. Examine the data characteristics of each of these data sources.

Important considerations in integrating the data sources include the following:

- What are the sample sizes of source and target data sets
- Do these data sets include compatible data elements for the geographic region in question?
- Do the data set contain adequate demographic references that facilitate integration?
- How should the records be weighted?

Step 3. Identify common (or similar) data elements that facilitate data fusion.

Each of the databases to be integrated are normally organized as several tables that are related to each other by way of one or more common attributes (or fields). This would mean that individual records (or rows) in each table may contain a corresponding record in another table. When two or more databases are to be integrated, there relational databases may be combined by data elements that are statistically compatible. Usually individual data elements may have to be aggregated to achieve this statistical compatibility. For example, several tables within the NHTS data are related by the household identification number. A statistically compatible aggregate data group between the data sets could be the income group.

Step 4. Analyze and integrate datasets.

An important consideration for integrating data is the choice of database and other analytical tools. Most survey data are available in such formats such as SAS, database and ASCII. Available information processing resources dictate the choice of analysis platform such as statistical software, relational databases and/or custom applications. Using the same or different tools, the analyzed and integrated data may be transformed into a form where the data can be presented for the intended end use. Examples of end uses of the integrated data include such tasks as presentation of the results in thematic map using GIS or environment and input files to travel demand models.

Case Study: Integrating NHTS with CTPP for Trip Generation Models by TAZ

The above mentioned five-step is illustrated with a case study involving the development of trip generation models for a medium-sized urban area. Trip generation models are usually developed as cross-classification models for such market segments as income group or auto ownership. Usual practice in developing travel demand models for medium to large urban areas is to conduct a household travel survey and develop model parameters based on this survey. Primary assumptions for this case study include the following:

- The incorporated area has a population of 135,000.
- No local travel survey data are available.
- Trip generation rates are to be cross-classified by income group

- The regional planning agency is aiming at integrating data from readily available sources to meet this modeling need and
- The 2000 NHTS survey sampled about 100 households from this urban area

Identification of appropriate databases and data elements

The NHTS data contains extensive information on trip purpose and other demographic variables used for developing trip generation models. Appropriate demographic variables in this case include household income and auto ownership. Since the city has been sampled by NHTS, potential exists for using the NHTS data as a data source for developing trip generation models for this city. The CTPP database for this city contains demographic information on much larger number of households. The CTPP data are available to the traffic analysis zone (TAZ) level spatial resolution. Therefore, the potential exists to integrate the NHTS data with CTPP data to obtain trip generation rates at zonal level.

Characteristics of NHTS and CTPP data

Since NHTS sampled only 100 households from this city, which represents less than 0.1 percent of the households, the sample may not be adequate enough to generate trip generation rates solely from NHTS data. On the other hand, CTPP which contains data on 10 percent of the households, has information only on household work trips. For this reason trip generation rates for non-work and non-home based trips cannot be obtained from CTPP. The following table summarizes demographic references that are appropriate for developing trip generation models using integrated NHTS and CTPP.

Table 1
Availability of Appropriate Demographic References in NHTS and CTPP Data

Demographic	NHTS	CTPP
Household income	Yes	Yes
Auto ownership	Yes	Yes
Trip purpose	Yes – complete	Yes – work trips only
Sample size	Varies	10%

Data elements and procedures for trip generation models

Day trips file of NHTS, like most local household travel surveys, focuses on trips taken by household member on a given day. For deriving trip generation rates, these day trips are first counted at household level and then aggregated by income category to which each of the households belong. Trip generation rates for each income group are computed by dividing total number of trips divided by the number of households in that category.

As the term cross-classification implies, the data will be segmented into classes – which in this case are income groups. When small samples are further segmented, certain categories stand the risk of not being represented. For example, suppose the 100 NHTS households for the case study city are to be grouped into five income categories: less than $15,000; $15,000-34,999; $35,000-$74,999; and greater than $75,000. In this case there is not only a danger of non-representation for one or more income groups, but also a risk of over representation of one or more groups. On

the other hand, the 10 percent CTPP sample for the same city may be expected to more realistic representation of the households in the specified income groups (which is the same as CTPP income category #5).

A legitimate question to ask is that 'is it statistically appropriate to apply the trip generation rates developed using a small sample to a large sample where income groups are relatively better represents?' The answer to this question is that it is not appropriate to do so. One way to resolve this issue is to compare the trip generation characteristics of this city with similar cities and establish a larger sample for developing trip generation rates. For instance a statistical tests may be performed to compare the city with 135,000 population to the other cities in NHTS database where the metropolitan statistical area (MSA) populations are less than 250,000. For this groups, cities with similar trip making characteristics may be grouped to increase the representative sample size in NHTS data.

Whatever may be the size of NHTS sample from which trip generation rates are derived, before applying these rates to the CTPP income groups, the work trip generation rates (a category common to both NHTS and CTPP) for the two data sources should be compared. Any significant differences between the two must be reconciled.

Analyzing and integrating data elements

The data manipulation process in the above mentioned steps is fairly involved. Agencies must choose proper tools and allocate appropriate to properly analyze data and combine data elements. Listed among the most commonly used tools for these tasks are database management systems, statistical analysis system and electronic spreadsheets. tools

Upon developing trip generation rates of transferable quality are developed using the NHTS data, the rates may be applied to each of the households in the same market segments (i.e. income group) in the CTPP data. Since most households in the CTPP data are mapped to the TAZ level, total trips generated by all household in the TAZ may then be obtained. At this point, the trip generation characteristics of the NHTS data are *transferred* to the CTPP data for the subject city in the case. Once this step is completed, developing trip generation models using the CTPP is not restricted just the income group market segment. Rather, the trip generation rates for CTPP data may be developed for any demographic category.

Emerging Issues of Importance

During the time period 2000-2009 American Community Survey (ACS) will replace the census long form. While many of the details of ACS data are still being worked out, in the coming years this new survey is expected to play a pivotal role in integrating appropriate data from the NHTS data.

Recent emphasis on transportation needs of the elderly and the disabled and the unavailability of adequate resources has resulted in looking to data fusion as an alternative to meeting these needs.

References

1. Metropolitan Washington Council of Governments (MWCOG). *FY-2003 Models Development Program for COG/TPB Travel Models*. National Capital Region Transportation Planning Board. June 2003.
2. Bricka, Stacy (2001). "Variations in long Distance Travel." *Transportation Research Circular*, No. E-C026. March 2001.
3. Zmud, Johanna, P. and Arce, Carlos, H. (2001). "Influence on Consumer Culture and Race on Travel Behavior." *Transportation Research Circular*, No. E-C026. March 2001.
4. Hu, Patricia, S. and Young, Jenny (2001). "Using the NPTS and ATS Together: A Case Study." *Transportation Research Circular*, No. E-C026. March 2001.
5. Niles, John and Nelson, Dick (2001). "Enhancing Understanding of Non-Work Trip Making: Data Needs for the determination of TOD Benefits." *Transportation Research Circular*, No. E-C026. March 2001.
6. Pratt, Joanne, H. (2002). "Teleworkers, Trips, and Telecommunication: Technology Drives Telework-But Does it Reduce Trips?" *Transportation Research Record*, Issue, 1817, pp. 58-66.
7. McGukin, Nancy, Keyes, Ann, Mary, and Banks, David (2001). "Are Households with Interrupted Phone Service Like Those with No Telephone Service? Comparison Using Public-Use Microdata Set and Nationwide Personal Transportation Survey." *Transportation Research Record*, Issue, 1768, pp. 99-105.
8. *The Relative Risk of School Travel: A National Perspective and Guidance for Local Community Risk Assessment*. TRB Special Report 269. Transportation Research Board, National Research Council, Washington, D.C. 2002.
9. Rosenbloom, Sandi and Waldorf, Brigitte (2001). "Older Travelers: Does Place or Race Make a Difference." *Transportation Research Circular*, No. E-C026. March 2001.
10. Evans, Edward, L. (2001). "Influence on Mobility Among Non-Driving Older Americans." *Transportation Research Circular*, No. E-C026. March 2001.
11. Gardenhire, Alissa, D. and William, Sermons, M. (2001). "Understanding Automobile Ownership Behavior of Low-Income Households: How Behavioral Differences May Influence Transportation Policy." *Transportation Research Circular*, No. E-C026. March 2001.
12. Dill, Jennifer (2001). "Older Vehicles and Air Pollution: Insights from the 1995 NPTS." *Transportation Research Circular*, No. E-C026. March 2001.
13. Erlbaum, Nathan, S. (2001). "Improving Air Quality Models in New York State: Utility of the 1995 Nationwide Personal Transportation Survey." *Transportation Research Circular*, No. E-C026. March 2001.
14. Polzin, Steven, E., Chu, Xuehao, and Rey, Joel, R. (2001). "Mobility and Mode Choice of People of Color for Non-Work Travel." *Transportation Research Circular*, No. E-C026. March 2001.
15. Ziliaskopoulos, Athanasios, K. and Waller, Travis, S. (2000). "An Internet Based Geographic Information System that Integrated Data, Models and Users for Transportation Applications." *Transportation Research C*, Issue 8, pp. 427-444
16. Beckman, RJ; KA Baggerly; and MD McKay. (1996). Creating Synthetic Baseline Populations. Transportation Research. Part A: Policy and Practice. Volume: 30 Issue: 6. pp 415-429

17. Stopher, P., S. Greaves and P. Bullock. Simulating Household Travel Survey Data: Application to Two Urban Areas. Paper Presented at the Transportation Research Board Annual Meeting, January 2003.